IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 July 2025, accepted 30 July 2025, date of publication 5 August 2025, date of current version 22 August 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3595665

SURVEY

LLM-Driven APT Detection for 6G Wireless Networks: A Systematic Review and Taxonomy

MUHAMMED GOLEC[1], YASER KHAMAYSEH[2], SUHIB BANI MELHEM[3], AND ABDULMALIK ALWARAFY[4], (Member, IEEE)

[1]Department of Computer Engineering, Boğaziçi University, 34342 Istanbul, Türkiye
[2]College of Technological Innovation, Zayed University, Abu Dhabi, United Arab Emirates
[3]Department of Cybersecurity, College of Engineering, Al Ain University, Abu Dhabi, United Arab Emirates
[4]Department of Computer and Network Engineering, College of Information Technology, United Arab Emirates University, Al Ain, United Arab Emirates

Corresponding author: Abdulmalik Alwarafy (aalwarafy@uaeu.ac.ae)

The work of Yaser Khamayseh was supported by the Zayed University Research Office under Grant 23153.
The work of Abdulmalik Alwarafy was supported by United Arab Emirates University.

ABSTRACT Sixth Generation (6G) wireless networks, which are expected to be deployed in the 2030s, have already created great excitement in academia and the private sector with their extremely high communication speed and low latency rates. However, despite the ultra-low latency, high throughput, and AI-assisted orchestration capabilities they promise, they are vulnerable to stealthy and long-term Advanced Persistent Threats (APTs). Large Language Models (LLMs) stand out as an ideal candidate to fill this gap with their high success in semantic reasoning and threat intelligence. This paper presents the first systematic review and taxonomy for LLM-assisted APT detection in 6G networks. It also provides insights by reviewing recent research on the intersection of LLMs, APTs, and 6G. Key challenges such as limitations in edge deployment, data scarcities, and explainability gaps are identified and a multidimensional taxonomy is provided in line with the APT lifecycle and 6G contexts. The paper is based on 142 studies from 2018 to 2025, searching leading databases such as IEEE Xplore, ACM Digital Library, SpringerLink, and Elsevier ScienceDirect.

INDEX TERMS 6G wireless networks, advanced persistent threat (APT), large language model (LLM), natural language processing (NLP), security.

I. INTRODUCTION

The rapid development of wireless technologies increases expectations for 6G networks with ultra-low latency and artificial intelligence-based orchestration architecture [1]. To meet these expectations, 6G networks operate with a heterogeneous architecture where many layers, such as physical and network layers, work together, which means a larger attack surface [2]. The wide attack surface in 6G systems requires that measures be taken against Advanced Persistent Threats (APTs), one of the hidden and long-stage attack methods that are difficult to detect with traditional detection mechanisms [3].

Large Language Models (LLMs) with semantic and contextual reasoning features are one of the most promising developments that can be used against APTs [4]. In particular, it can be used for the detection of APT in 6G networks by analyzing fragmented logs and increasing situational awareness [5]. Despite this potential, there is no comprehensive taxonomy or systematic analysis in the literature on LLM-based APT detection for 6G networks.

To the best of our knowledge, this paper is the first comprehensive systematic review and classification study on LLM-based APT detection in 6G networks. Therefore, this paper aims to fill the underexplored research gap regarding LLM-based APT detection in 6G networks. As a result of the current studies examined by the authors using identification, screening, eligibility, and inclusion (snowballing) techniques, 142 articles were analyzed. Our aimed is to synthesize the

The associate editor coordinating the review of this manuscript and approving it for publication was Nurul I. Sarkar.

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

intersection of LLM architectures, APT lifecycle modeling, and 6G-specific security challenges and provide insights for future research.

A. MOTIVATION AND CONTRIBUTIONS

LLMs and 6G technologies are very recent research areas and their intersection in cyber threat detection, such as APT attacks, has not been sufficiently investigated in the literature. Existing studies are scattered across either LLM-based cybersecurity or 6G network security issues. Furthermore, 6G networks are still in their infancy (expected to become widespread after the 2030s) and contain obstacles for AI and rule-based systems due to issues such as a fragmented structure of source data and end device limitations [6]. For all these reasons, a detailed investigation should be conducted to explore the potential of LLMs in providing explainable detection mechanisms throughout the 6G infrastructure. The main contributions of this paper can be summarized as follows:

- We present the first Systematic Literature Review (SLR)-based review focusing on LLM-enabled APT detection in 6G networks. To do this, we searched more than 300 most recent and relevant papers in academic and industrial databases between 2018-2025. As a result of the systematic analysis (Kitchenham's SLR approach and Petersen's Systematic Mapping Study (SMS) [7], [8]), the most relevant 142 papers in the field were obtained.
- We define five-point research questions to conduct the systematic review (Section IV-B). In line with these questions: (i) Semantic correlation of fragmented logs generated in 6G networks and how it can be used for LLMs threat detection (Section V-A), (ii) Limitations of 6G encrypted channels and how it can address LLMs visibility and reasoning challenges (Section V-B), (iii) Challenges of deploying LLM to edge nodes on 6G networks and optimization techniques for these challenges (Section V-C), (iv) Datasets and modeling techniques used in LLM-based APT detection studies (Section V-D), and (v) Exploration of publication trends, platform distribution, and reproducibility for LLM-focused APT research (Section V-E).
- LLM deployment models, threat lifecycle stages, optimization strategies for edge inference, and taxonomy studies for dataset types are presented.
- Research gaps, such as explainability gaps, dataset scarcity, and 6G orchestration risks, are highlighted through critical analysis. And future directions, such as slice-aware XAI pipelines and unified demand tuning techniques, are highlighted.
- A comparison of this paper with 16 previous reviews in the literature. The comparison is made to highlight the novelty and necessity of the paper.

Terms and abbreviations used throughout the paper are given in table 1.

TABLE 1. List of abbreviations used in this paper.

Abbreviation	Definition
6G	Sixth Generation Wireless Networks
APT	Advanced Persistent Threat
LLM	Large Language Model
NLP	Natural Language Processing
IDS	Intrusion Detection System
XAI	Explainable Artificial Intelligence
RIS	Reconfigurable Intelligent Surface
SDN	Software Defined Networking
FL	Federated Learning
THz	Terahertz Communication
CTI	Cyber Threat Intelligence
PEFT	Parameter-Efficient Fine-Tuning
RAG	Retrieval-Augmented Generation
FSM	Finite State Machine

B. ARTICLE ORGANIZATION

Figure 1 shows the organizational chart for this paper. Section II provides a comparison with related surveys to highlight the novelty of the paper. Section III explains the basic background of APTs, 6G networks, and LLMs, and explains their roles in cybersecurity. Section IV explains the methodological structure, such as the article selection methods and research questions for this systematic review. Section V provides an in-depth analysis of five key research questions. Section VI indicates open challenges and future directions for researchers in the relevant research area. Section VII concludes the paper by summarizing the findings and emphasizing the importance of LLM-based APT detection in 6G.

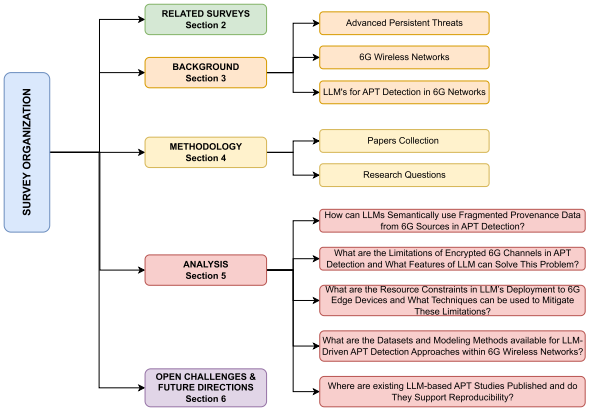

FIGURE 1. The organization of the survey.

II. RELATED SURVEYS

The use of LLMs in the detection and prevention of APTs, which are potential cybersecurity threats in 6G, is still an area that needs to be investigated. The main reason for this is that there are limited datasets in the literature on APTs and 6Gs can only be modeled simulation-based. When the literature is examined, it is seen that although there are surveys focused on 6G, APT, LLMs, and LLM-based security, however, there is no systematic review and taxonomy study that addresses LLM, APT, and 6G in a combined manner.

LLM-Focused Cybersecurity Surveys: Several recent studies in the literature have examined the role of LLMs in the field of cybersecurity. Hassanin and Moustafa [9] provide an overview of the role of LLMs in applications such as threat intelligence and phishing detection in their review. In [10], [11], [12], and [13], they examine the architectures used for LLM-based attack detection and threat analytics in a more systematic way. Zuo et al. [14] presented an analysis study examining LLMs usage for APTs. This study investigated the semantic augmentation of LLMs (such as GPT-4o) origin logs for APT detection. However, this study is superficial, not a survey or taxonomy, and does not include the 6G context. In another study, the authors present a review of language models (including APT), but do not include any information about 6G and do not use a formal methodology such as SLR [15]. Although LLMs sheds light on the applications in cybersecurity, none of these studies cover 6G and its limitations and opportunities.

APT-Focused Surveys: Some of the literature studies investigated APT detection using DL and rule-based learning. In [16] and [17], classifications and threat lifecycle analyses of APTs are examined, while in [18] DL-based cyber attack detection systems (partially addressing APT) are investigated. Although all these survey studies partially or in detail mention APTs, none of them provide detailed information about LLM-based approaches or 6Gs (such as network layer dynamics).

6G Focused Surveys: Another area of survey research in the literature examines the technical foundations of 6G. Shen et al. [19] covers five main aspects of 6G (such as spectrum and positioning) in detail. In [20], [21], [22], and [23], important components in 6G (such as IoT integration and federated learning) are comprehensively examined. In another survey study, Sun et al. [24] investigate the importance and use of explainable AI (XAI) in 6G network slicing and vehicle contexts. However, none of these studies address LLM and security issues.

Among the reviewed literature studies, [10], [11], [12], [13], [18] examine model architectures and use cases in detail in their focused research topic using systematic research methodology such as PRISMA. Furthermore, some of the studies [11], [13] provide binary taxonomies of cybersecurity tasks, while in [19], [20], [21], and [23] they strongly address 6G at the architecture and protocol level. However, none of the reviewed articles address LLM, APT, and 6G in a unified manner.

A. CRITICAL ANALYSIS

Table 2 provides a comparison of 16 recent survey studies with this paper. When the table is examined, it is seen that this paper fills the following three critical gaps:

- **Combining consideration of LLM, APT, and 6G**: None of the reviewed studies simultaneously address the intersection of LLM-based threat detection, APT lifecycle modeling, and 6G network features.
- **Providing a detailed taxonomy for APT detection**: The vast majority of studies are in the form of a general survey, and those that do include a taxonomy lack details such as the lifecycle of APTs.
- **Providing a comparative synthesis across fields**: Few of the reviewed studies include multidimensional comparisons (such as methodology, model types). The lack of such comparative syntheses makes it difficult to assess the overlap and gaps between the topics covered by the survey.

III. BACKGROUND

This section provides the basic background necessary for the reader to better understand the concepts related to LLM-based APT detection in 6G networks.

A. ADVANCED PERSISTENT THREATS

Advanced Persistent Threats (APT) are one of the most effective cyber attacks known due to their characteristics, such as stealth and longevity. This subsection defines APT and explains its key characteristics, lifecycle, and attacker behaviors (TTPs). Then, a comparison of traditional attacks and APTs is provided for APT.

1) KEY CHARACTERISTICS OF APT

APTs use multiple vectors to gain long-term access to an IT environment and are like an attacker with significant expertise and resources [25]. They have three basic characteristics [26]:

- **Advanced**: These attacks specialize in zero-day attacks and tactics to evade detection.
- **Persistent**: They encourage new APT attacks by leaving backdoors in the systems they penetrate.
- **Threat**: They carry out attacks such as espionage, sabotage, or exfiltration of critical data from the systems.

2) THE LIFECYCLE OF APTs

The lifecycle for APT attacks is shown in figure 2 and can be summarized in five basic stages [15], [16], [17]:

- **Reconnaissance**: This is the first stage of the attack, and information about the target is collected (Open Source Scanning (OSINT)), and system vulnerabilities and weaknesses are investigated.
- **Initial Intrusion**: The entry into the target system is achieved through methods such as phishing and malware.
- **Command and Control (C2)**: Preparation of the infrastructure to communicate with the APT inserted into the target system (such as backdoor channels).
- **Lateral Movement**: Infiltration of other devices connected to the same network within the system and detection of high-value targets.
- **Data Exfiltration**: The final stage involves malicious operations such as exfiltration of data in the target system using APT and system sabotage.

TABLE 2. Comparison of our systematic review and taxonomy with existing survey studies.

Paper	Focus Area	6G-Specific	APT-Specific	LLM-Specific	Type	SLR Methodology	Publisher	Year
[9]	General Cyber Defence	✗	✗	✓	Review	✗	Arxiv	2024
[10]	LLMs in Cybersecurity	✗	✓	✓	Systematic Review	✓	Arxiv	2024
[11]	IDS with Transformers & LLMs	✗	✓	✓	Review and Taxonomy	✓	Elsevier	2024
[12]	LLMs in Cybersecurity	✓	✓	✓	Systematic Survey	✓	IEEE	2025
[13]	LLMs in Cyber Threat Detection	✗	✓	✓	Systematic Review	✓	Elsevier	2024
[14]	LLM-Augmented Provenance for APT Detection	✗	✓	✓	Research Paper	✗	Sandia TR	2025
[15]	PLMs/LLMs in Cybersecurity	✗	✓	✓	Review	✗	IEEE (ISDFS)	2024
[16]	APT Analysis & Countermeasures	✗	✓	✗	Review and Taxonomy	✗	Springer	2019
[17]	APT Detection Techniques	✗	✓	✗	Survey	✓	TechScience (CMC)	2024
[18]	DL Techniques for IDS	✗	✓	✗	Systematic Survey	✓	ACM	2025
[19]	6G Architectures & Networking	✓	✗	✗	Survey	✗	ACM CSUR	2023
[20]	Federated Learning in 5G/6G Cybersecurity	✓	✗	✗	Comprehensive Survey	✓	IEEE	2025
[21]	6G and IoT Integration	✓	✗	✗	Comprehensive Survey	✗	IEEE	2022
[22]	Optimization & Performance of LIS in 6G	✓	✗	✗	Survey	✗	IEEE Access	2020
[23]	6G Technologies & Architectures	✓	✗	✗	Survey	✗	IEEE	2022
[24]	Explainable AI for 6G	✓	✗	✗	Systematic Survey	✓	IEEE OJ-COMS	2025
Our Paper	**LLMs for APT Detection in 6G**	✓	✓	✓	**Systematic Review and Taxonomy**	✓	**Computer Science Review (Planned)**	**2025**

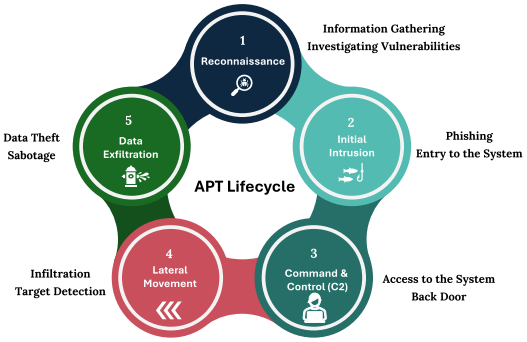

FIGURE 2. The five-stage lifecycle of an APT.

3) TACTICS, TECHNIQUES, AND PROCEDURES (TTPs)

TTPs are shown in Figure 3 and are the framework used to classify the behavior of an APT attack. It can be defined as follows [27]:

- **Tactics**: Used to define the goal of the attack, such as gaining access to a system.
- **Techniques**: Refers to the technique used to achieve this goal. An example would be DLL injection.
- **Procedures**: Refers to the way the attack is implemented, such as sending a special email.

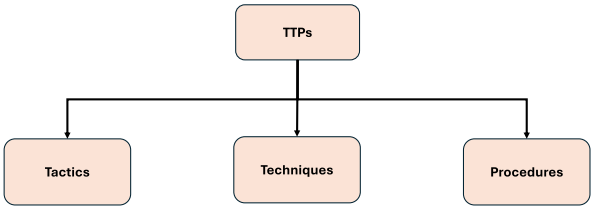

FIGURE 3. The hierarchical structure of TTPs.

4) APT VS. TRADITIONAL ATTACKS

APTs and traditional attacks differ from each other in many ways, such as target, tactics, duration, and these differences are shown in Table 3. Traditional attacks aim to cause general damage and aim for quick gain, while APTs are long-term and professional attacks (usually state-sponsored) [26]. Traditional attacks identify weak systems by simultaneously attacking many targets, while APTs are more target-oriented and the attack process is carried out in a long and sneaky way [28].

TABLE 3. Comparative characteristics: Traditional attacks vs. advanced persistent threats (APT).

Attribute	Traditional Attack	Advanced Persistent Threat (APT)
Target	Broad or Random	Highly Specific
Duration	Short-lived	Prolonged (Months or Years)
Entry Vector	Known Exploits	Custom Zero-Days, Spear Phishing
Goal	Financial Gain, Disruption	Espionage, Strategic Access
Tools Used	Commodity Malware	Tailored, Multi-Stage Toolkits

B. 6G WIRELESS NETWORKS

1) ARCHITECTURAL FOUNDATIONS OF 6G

6G is the new generation of wireless communication paradigm that emerges with the integration of advanced physical technologies and software-defined network solutions [29]. In order to provide uninterrupted communication in the 6G architecture, it is a heterogeneous structure (Ultra-Dense Heterogeneous Networks) that combines three basic layers: terrestrial, aerial, and satellite [30]. In order to reach a data rate of more than 1 Tbps, technologies such as Terahertz (THz) communication and Visible Light Communication (VLC) are used [31]. In addition, power-sensitive technologies such as Reconfigurable Smart Surfaces (RIS) and Software-Defined Metasurfaces (SDM) are used to reduce latency [32].

The 6G networks with heterogeneous architecture shown in Figure 4 try to reduce computational loads with edge devices and AI-based systems [33]. While 5G architectures use a centralized system, in the 6G architecture, thanks to decentralization, network slices can be optimized autonomously via AI-based engines. However, despite these advantages, the heterogeneous and decentralized architecture offers a large attack surface and is exposed to cyberattack threats [33].

2) KEY FEATURES OF 6G

Table 4 shows a feature comparison of 5G and 6G, and as can be seen, 6G (key features) is superior in every aspect. 6G is expected to work in harmony with real-time

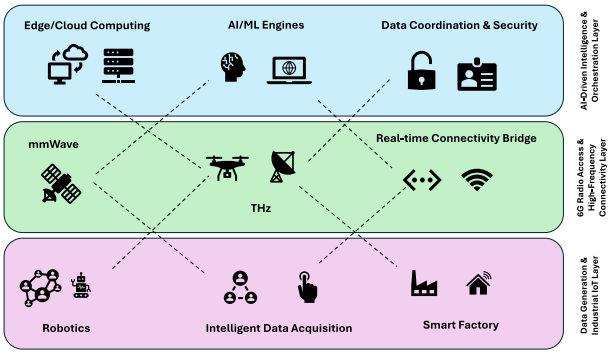

FIGURE 4. 6G architectural pillars and deployment layers.

TABLE 4. Comparison of key features between 5G and 6G wireless networks.

Feature	5G	6G
Latency	Approximately 1 millisecond	Less than or equal to 0.1 milliseconds
Data Rate	Up to 20 Gbps	At least 1 Tbps
Frequency Range	Sub-6 GHz and millimeter wave	Sub-Terahertz (Sub-THz) and Visible Light Communication (VLC)
Architecture	Centralized network control	Distributed and AI-powered network architecture
Security	Add-on security features	Built-in, intent-aware security mechanisms

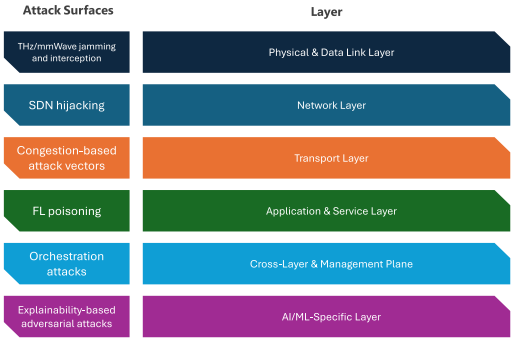

FIGURE 5. Illustration of potential attack surfaces across the hierarchical 6G network.

holography and trusted autonomous systems once it is available for daily use [34]. 6G relies on AI-powered protocols and advanced infrastructure capabilities to meet these demands [35].

3) VULNERABILITIES AND THREAT LANDSCAPE IN 6G

Despite the high speed and wide infrastructure opportunities they offer, 6G networks also carry risks such as misconfiguration and hostile exploitation due to AI-based control logic and network software such as SDN/NFV [36]. If vertical slicing and segmentation operations in networks do not work correctly, they become vulnerable to lateral attacks (sourced by APT's etc.) [37].

Possible potential attacks that may occur in 6G are shown in Figure 5. Attack types can range from physical layer compression to manipulation. Another potential danger is that the AI mechanisms responsible for 6G orchestration are vulnerable to attack and data leakage in cases where RIS and THz communication channels are not properly set [38].

4) 6G-SPECIFIC CHALLENGES FOR APT DETECTION

APTs are expected to threaten the rapid lateral movements that 6G will bring to our lives [39]. In addition, behavioral detection becomes more complex for traditional signature-based intrusion detection systems (IDS) due to architectural features such as encrypted layers and dynamic topologies [40].

Another challenge for 6G networks is the scarcity of APT datasets, which makes it difficult to train security models [41]. Another challenge is the fragmented nature of source logs, as this limits the correlation between layers that can be used in APT detection [42].

5) RESEARCH TRENDS INTEGRATING 6G AND AI FOR SECURITY

Literature studies investigate the use of FL at edges to detect attacks while also concerning privacy [43]. In addition, XAI methods for decision-making mechanisms are another frequently investigated method [44]. Beyond these, mapping TTPs and analyzing logs with LLMs based systems are promising [45]. However, since edge devices are resource-constrained and storage-limited devices, these limitations should be taken into consideration when deploying LLMs at edges, and strategies such as model distillation should be applied [46].

C. LLMs FOR APT DETECTION IN 6G NETWORKS

1) OVERVIEW OF LLM ARCHITECTURES AND SECURITY-ORIENTED SPECIALIZATIONS

Developed on Transformer architecture, LLMs have made a great breakthrough in the field of AI, and these models provide representation learning by making sense of the context [47]. In other words, these LLMs indicate the ability to understand the meanings of words in context beyond their dictionary meanings. Thanks to this ability, they achieve great success in natural language understanding (NLU) and generation (NLG) tasks [48].

This technological development (LLMs) has begun to be used in many areas, especially in cybersecurity, and the evolution of LLMs in cybersecurity use is shown in Figure 6. These areas of use can be examined under three main headings [49], [50]:

- **General-Purpose**: LLM models that can be trained with large text collections and used for various purposes.
- **Domain-Specific**: LLM models trained (fine-tuned) using purpose-oriented cybersecurity data.
- **Emerging Techniques**: These are methods that make LLMs models lightweight and specific to their intended use.

General-purpose models such as BERT (2018), GPT-2 (2019) have shown success in text classification and question-answer tasks, and with the customization of these models, domain-specific models such as SecBERT (2020), CyBERT (2021) have been developed and started to be used in special tasks such as malware detection (such as APT). And studies

on emerging techniques continue to increase the performance of these models.

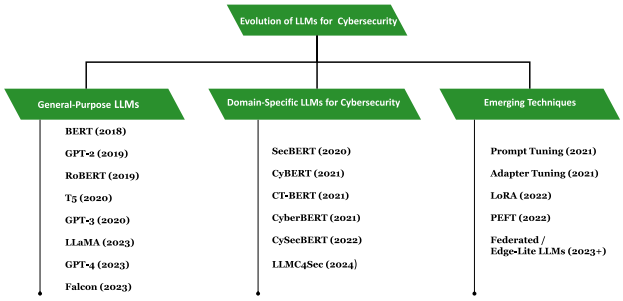

FIGURE 6. Evolution of large language models (LLMs) for cybersecurity.

2) APPLICATIONS IN CYBERSECURITY AND APT DETECTION

LLMs have been used in cybersecurity for multi-APT detection and response, returning based on attack type [13]. LLMs features, such as contextual reasoning and linguistic understanding, make it particularly suitable for APTs with multi-stage attacks [51]. Figure 7 shows a tree structure for LLM application areas in APT detection. As can be seen from the figure, LLMs are versatile in the cybersecurity context [52], [53], [54]:

- **Threat Intelligence**: LLMs models can extract TTPs using open-source data such as threat reports.
- **APT Behavior Modeling**: Logs and lineage data can be used to semantically interpret multi-stage APTs.
- **Anomaly Detection**: LLMs context-aware feature can detect anomalous behavior (network and system logs).
- **Alert Triage and Incident Response**: Natural language summarization translates alerts into insights to extract meaningful information (helpful for analysts).
- **TTP Alignment**: LLMs fine-tuning can map hostile behaviors for low-fire systems to MITRE ATT&CK stages.

In dynamic networks (especially 6G networks), the GPT family provides situational awareness in complex architectures where traditional detection methods can struggle.

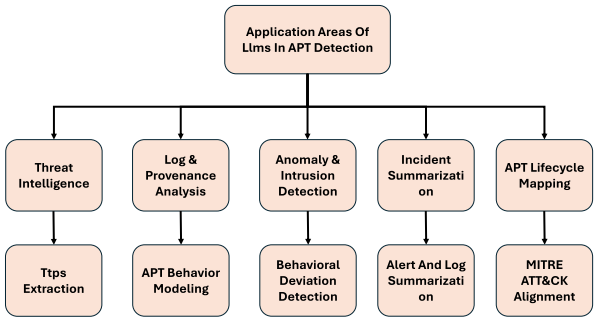

FIGURE 7. Application areas of LLMs in APT detection.

3) LLM INTEGRATION CHALLENGES IN 6G EDGE ENVIRONMENTS

Edge devices positioned close to the data source have advantages such as low latency and low bandwidth usage, but also disadvantages such as heterogeneous structure and limited processing power [55]. For this reason, problems arise due to these limitations when LLMs are deployed on 6G-based edge devices. Figure 8 summarizes these limitations and possible solutions [1], [2]:

- **Resource Constraints**: Since LLMs require high processing power and storage, their implementation on resource-constrained 6G edge devices (RAM & Compute Energy Efficiency) is one of the major problems that can be encountered.
- **Latency Constraints**: Edge devices, which are expected to offer a low latency advantage due to being positioned close to the data source, may lose this advantage due to the high computational time of LLMs.
- **Privacy & Compliance**: Sensitive data such as biometrics must take into account some privacy concerns when processed on edge devices [55], [56].

The methods to solve these challenges can be summarized as follows [57], [58]:

- **Compression**: LLMs can be downgraded to lower versions to reduce memory and processing load.
- **Knowledge Distillation**: Information obtained from large models can be transferred to smaller models to minimize performance loss.
- **Federated/Split Inference**: Both privacy and efficiency can be increased by distributing the components of the LLM model to different edge nodes and processing them.

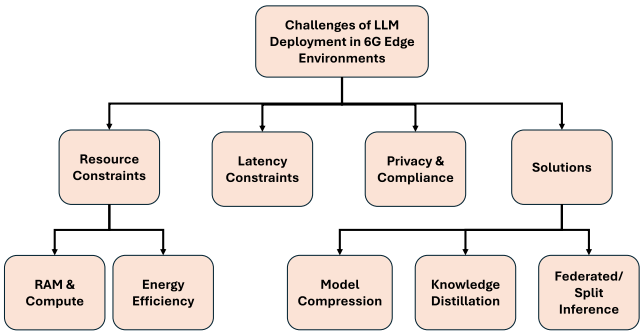

FIGURE 8. Challenges of LLM deployment in 6G edge environments.

4) APT DETECTION-SPECIFIC BENEFITS OF LLMs IN 6G CONTEXT

LLM models have great potential in APT threat reduction studies in 6G networks, which are expected to be used in the near future. This potential stems from the success of LLM models in establishing semantic correlations between data types and their ability to analyze the attack lifecycle as a whole [59]. The contributions of LLM models for APT

detection in 6G networks can be generalized as follows [60], [61]:

- **Cross-Layer Fusion**: It can be used in the detection of multi-vector attacks by combining log records from the control and user planes and the cloud layers.
- **Lifecycle Prediction**: LLM models can use past attack data to predict the next step in the APT kill-chain.
- **Semantic Generalization**: LLM models can capture attacks in a contextual manner in encrypted and hidden attack situations where traditional systems are inadequate.

Figure 9 shows the contributions of LLM models for layers. As can be seen from the figure, LLM models can perform multi-layered threat modeling by performing detection not only at the packet level but also at various levels of the network.

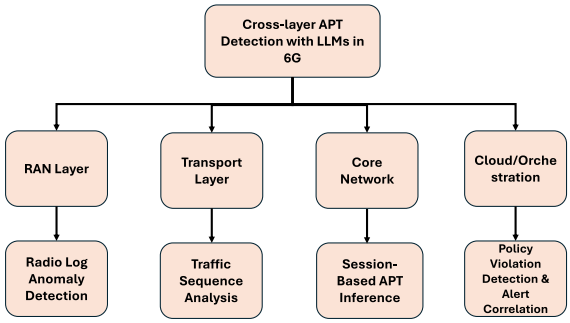

FIGURE 9. Cross-layer APT detection with LLMs in 6G.

In addition, LLM models not only interpret the behavior of APT attacks but also provide insight into the attack lifecycle phases and response mechanisms. Figure 10 shows the tasks that LLMs undertake in the APT kill chain model.

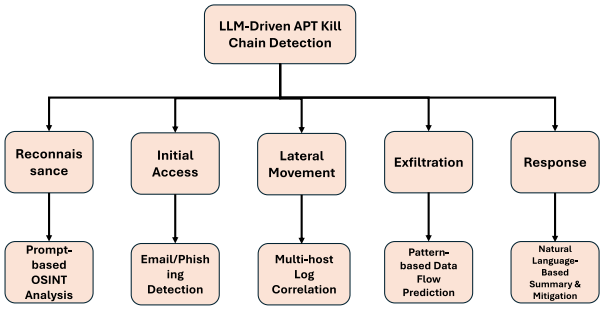

FIGURE 10. LLM-driven apt kill chain detection.

5) PROPOSED TAXONOMY: LLM-DRIVEN APT DETECTION IN 6G

This paper aims to classify LLM-based APT detection approaches in 6G networks and provide a comprehensive taxonomy. Figure 11 shows this taxonomy and can be summarized in five dimensions:

- **Input Modalities**: LLM models can be fed from various data sources such as logs and PCAP.
- **Detection Granularity**: LLMs can perform APT detection at different levels, such as single-packet analysis and session-based modeling.
- **LLM Techniques**: LLM models can be trained in various ways (prompt tuning, etc.) according to different scenarios.
- **Deployment Models**: LLMs can be deployed on different platform environments, such as cloud computing and edge computing.
- **Threat Lifecycle Phase**: LLM models can provide analysis and interventions at various stages of the APT kill chain.

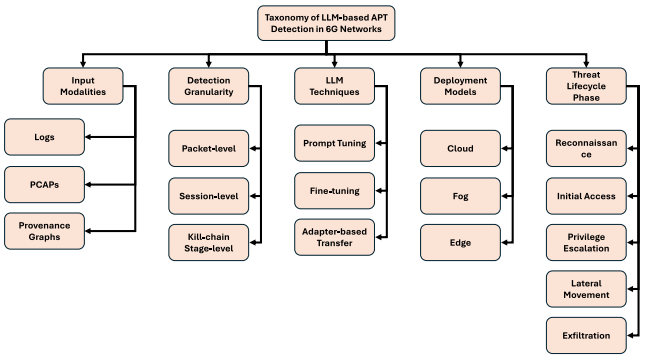

FIGURE 11. LLM-based APT detection taxonomy in 6G networks.

IV. REVIEW METHODOLOGY

This section presents the methodologies employed in the systematic review of LLM-driven APT detection approaches within 6G wireless networks and subsequently outlines the formulated research questions.

A. PAPERS COLLECTION

Since LLM-focused APT detection approaches in 6G wireless networks are a very current topic, we targeted the years 2018-2025 (current) to collect the relevant current literature studies. The following keywords were used to identify the studies related to the research topic:

1) [(LLM) | (LargeLanguageModel)] & [(APT) || (AdvancedPersistentThreat)]|
2) [(6G) | (WirelessNetworks)] & [(LLM) || (APTDetection)] & [(Edge) || (CrossLayerSecurity)]|
3) [(CyberThreatIntelligence) |(ProvenanceLogs)] & [(LLM) || (APT)] & [(6G)]|
4) [(LLM)] |[(APT)] || [(6G)]|

Figure 12 shows the collection and filtering process of the articles examined in this study. The steps carried out for this process are aimed at providing a comprehensive and structured analysis by following Kitchenham's Systematic Literature Review (SLR) and Petersen's Systematic Mapping Study (SMS) approaches [7], [8]. The steps summarizing this process are explained below:

- **Identification**: Known major academic literature sources (IEEE, ACM, Elsevier, Springer), technical reports, book chapters, and reference lists were scanned.
- **Screening**: Duplicate documents were removed, and the number of papers decreased to 126.
- **Eligibility**: Papers collected by our expert authors were analyzed, and only quality and scope-compliant papers were selected (the number of papers decreased to 120).
- **Included**: Additional relevant studies were added using the backward and forward snowball method, and the paper set was determined as 142 [62].

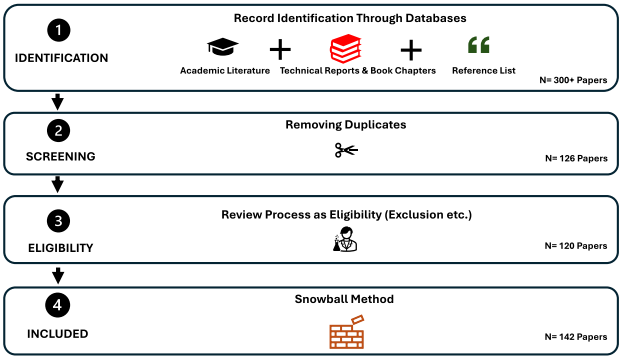

FIGURE 12. The paper collection process.

B. RESEARCH QUESTIONS

The research questions used in the systematic review and the section in which they are examined are shown in Table 5. Using these research questions, the current literature is examined and analyzed.

V. ANALYSIS

In this section, we discuss how we addressed the research questions in this study.

A. SEMANTIC CORRELATION OF FRAGMENTED PROVENANCE LOGS IN 6G (RQ1)

6G networks contain a lot of fragmented lineage data due to their heterogeneous structure (edge, cloud, etc.), and this data is distributed and inconsistent, which causes difficulties for security analysts and attack systems in APT detection [63]. One example of these difficulties is that rule-based and statistical detection methods fail to capture the nuanced context required for attack detection [64], [65]. Recent studies have focused on LLM-based methods to semantically combine fragmented lineage data and provide context-aware correlation. Figure 13 shows how fragmented lineage data can be semantically associated with LLM-enabled systems in a multi-layered manner. LLMs offer a promising solution by generating consistent security narratives by syntactically and temporally handling various records (such as security logs) [66], [67], [68], [69], [70].

Recent findings have shown that LLMs models can effectively utilize many different sources, such as audit

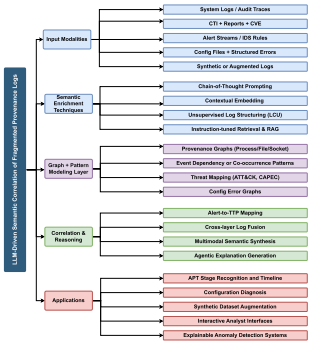

FIGURE 13. RQ1 taxonomy: LLM-based semantic correlation of fragmented provenance data across heterogeneous 6G sources.

logs [66], IDS alerts [4], CTI reports [71], and even static code artifacts [72]. Models that transform low-level source sequences into textual formats, such as APT-LLM [66], GENTTP [71], and LLMeLog [68], have been developed to reflect the system behavior semantics of models such as BERT or RoBERTa. Frameworks based on multitasking instructions and thought chains, such as SEVENLLM [73] and AnomalyGen [74], have been proposed for reasoning in data-scarce environments. For enrichment techniques, the literature includes studies such as retrieval-augmented generation [67], clustering embedding [68], [75], and ATT&CK alignment via request templates [76].

Many frameworks have been proposed that support fragmented logs with deep reasoning by capturing temporal, causal, and entity-level relationships and that resort to graph-based modeling. SHIELD [67], MultiKG [77], and MAD-LLM [78] frameworks use source graphs that encode dependencies of edges and represent system events at nodes. Other works such as AURORA [79] and DroidTTP [70] reconstruct attack sequences by applying classical planning and LLM. Works such as LocalIntel [80] and MCM-LLAMA [81] prefer dynamic association of SOC information and external alerts, while works such as LUNAR [75] and AnomalyGen [74] prefer association with CTI corpora. For high-level reasoning and explanation generation, these works resort to semantically annotated graph-based modeling.

Despite all these developments, there are still limitations that remain to be addressed. SEVENLLM [73] and SHIELD [67] frameworks use organized and synthetic logs, but this does not fully reflect the dynamic, heterogeneous nature of 6G. Another point to note is that mitigation strategies such as hybrid verification [66] and instruction fine-tuning [73] are rarely applicable to edge contexts. In addition, LLMs' high processing power and storage requirements make their application in 6G edge nodes a serious challenge, and therefore, the need for lightweight alternative methods such as TinyLM agents [82] or MoE-based distributed inference [78] is increasing. For future research, areas such as cross-layer lineage fusion, real-time semantic timeline reconstruction, and hallucination-aware causal modeling [72] stand out as a research gap.

TABLE 5. Summary of research questions (RQs), motivations, and corresponding sections.

NO	RQ	Motivation	Section
1	How can LLMs semantically use fragmented provenance data from 6G sources in APT detection?	The aim of this RQ is to investigate how fragmented resource records can be attributed by LLMs in 6G.	5.1
2	What are the limitations of encrypted 6G channels in APT detection and what features of LLM can solve this problem?	The aim of this RQ is to examine the potential challenges posed by 6G networks in APT detection and how LLMs can solve them.	5.2
3	What are the resource constraints in LLMs deployment to 6G edge devices and what techniques can be used to mitigate these limitations?	The aim of this RQ is to explore which compression strategies can be applied when deploying LLMs models to 6G edge devices.	5.3
4	What are the datasets and modeling methods available for LLM-driven APT detection approaches within 6G wireless networks?	The aim of this RQ is to investigate suitable datasets and dataset generation methods for LLM-focused APT detection approaches in 6G wireless networks.	5.4
5	Where are existing LLM-based APT studies published and do they support reproducibility?	The aim of this RQ is to evaluate the reproducibility of the dataset and model usability of the reviewed studies.	5.5

B. LIMITATIONS OF ENCRYPTED 6G CHANNELS AND LLM-DRIVEN SOLUTIONS (RQ2)

The widespread use of some communication protocols, such as DNS-over-HTTPS (DoH) and end-to-end encrypted tunnels, in the transition to 6G wireless networks has made great contributions to security and user privacy. In addition to these contributions, it also brings disadvantages like blind spots, such as traffic semantics obscurity for AI-supported detection systems. Figure 14 shows a taxonomy of LLM-focused solutions offered to address the challenges, limitations, and risks of 6G networks due to encrypted channels.

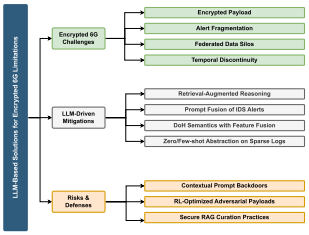

FIGURE 14. RQ2 taxonomy: LLM-based solutions for encrypted 6G limitations.

1) TECHNICAL LIMITATIONS IMPOSED BY ENCRYPTION

Encrypted 6G traffic channels limit the visibility of attack surfaces due to the techniques used (such as DoH). Recent studies have shown that advanced DL models fail to detect malicious traffic because semantic payloads become ambiguous while being encrypted [83]. In addition, advanced attack methods such as APT try to avoid detection by using encrypted channels such as DoH and embedding the C2 infrastructure in HTTPS payloads [84]. Edge-based data isolation, whose main purpose is privacy, prevents correlation (temporal and spatial) between devices. For example, since fragmented traffic logs are produced in UAV-based 6G networks, anomaly monitoring becomes very difficult [85]

2) LLM-DRIVEN MECHANISMS TO ADDRESS THESE GAPS

To overcome all these limitations, LLMs are promising by making meaningful inferences with their capabilities in semantic reasoning and contextual abstraction.

TABLE 6. Mapping encrypted 6G challenges to LLM-driven solutions.

Work	Limitation	LLM-Based Technique
Xu et al. [87]	Payload Obfuscation	Retrieval-Augmented Inference
Du et al. [79]	Alert Fragmentation	Multi-stage Reasoning via Prompt Engineering
Diao et al. [85]	Covert DoH C2 Channels	LLM + Expert Features for Tunnel Detection
Cheng et al. [89]	Contextual Reasoning in Sparse Logs	Log Fusion and Interpretation via Few-shot Learning
Sun et al. [84]	LLM Model Poisoning via Traffic	Adversarial Sample Generation with Reinforcement Learning (RL)
Liu et al. [88]	Contextual Logic Corruption	In-context Backdoor Prompt Manipulation

A recent study, APTSniffer, is a framework that detects APTs in encrypted channels by converting flow features into textual prompts [86]. The results confirm that the framework is successful with a 97% F1 score. Another study, MAD-LLM, is a framework that reconstructs APT chains by semantically collecting them through LLMs despite fragmented IDS alerts and encryption at the network layer [78]

Some malware (such as DoHunter, Godlua) are difficult to detect by detection systems because they use encrypted channels, so researchers track some technical features of the traffic, such as timing, length, and target domain structure, in addition to raw data with LLM models [84].

3) EMERGING CHALLENGES AND THREATS

Although using LLM models in encrypted channels is a promising solution, it is important to consider LLM-based vulnerabilities. One of these vulnerabilities is that LLM behavior can be manipulated by hostile requests and poisoned rollbacks. Studies have confirmed that fine-tuned LLM models based on RL can generate malicious traffic [83]. Furthermore, literature confirms that LLM models inject hidden logic into LLM models that are activated by benign triggers in encrypted channels [87]

In conclusion, while LLM models provide an advantage, such as semantic visibility for attack detection in encrypted channels, they also inherently introduce attack surfaces.

Table 6 summarizes the main limitations of encrypted 6G environments in light of the current literature reviewed.

C. DEPLOYING LLMs AT THE EDGE: CONSTRAINTS AND OPTIMIZATION TECHNIQUES (RQ3)

Despite the advantages of high speed and low latency offered by 6G networks, they consist of many different

distributed nodes and heterogeneous structures, such as edge devices. Therefore, LLM models to be used for security, privacy, and context-adaptive smart applications should also take into account the major computational, architectural, and security-related challenges when deployed in 6G networks. This research question (RQ3) examines optimization techniques for edge scenarios by categorizing these constraints. Figure 15 shows edge-oriented LLM optimization strategies.

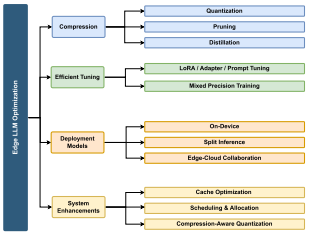

FIGURE 15. RQ3 taxonomy: Edge-oriented LLM optimization strategies.

Resource Constraints in Edge Environments: Edge devices (IoT, smartphones, etc.) consist of devices with limited processing power and storage capabilities, and even very small LLM models require more than 7GB RAM, which is usually beyond the capacity of edge computing devices [89]. Furthermore, due to the nature of LLMs (autoregressive), sequential token generation may cause latency bottlenecks [90].

Security and Fairness Considerations: Since LLM models deployed on edge nodes typically handle user data, privacy concerns may arise if this data is compromised [3]. Additionally, recent studies have reported that compression techniques used to make LLM models lighter for edge nodes may increase bias against underrepresented groups [91]. Therefore, the issue of fairness and reliability in compression cases is an open research gap.

Model Compression Techniques: These are techniques used to reduce the memory and processing load of LLM models, and one of the most popular methods is quantization. In this method, the weights (such as FP32) are converted to lower bit representations (INT8 or FP4) to reduce the model size and optimize the hardware speed [92], [93]. Another method where the distribution is optimized is pruning, and in this method, the weights are rescaled before quantization [89]. Distillation and low-rank approximation methods aim to provide additional performance gains on the inference quality [91], [92].

Parameter-Efficient Fine-Tuning (PEFT): Fine-tuning of LLM models cannot be done at edge nodes, and therefore PEFT methods (such as LoRA, Adapters, and Prompt-Tuning) apply them to local tasks by updating a subset of the models' parameters [94]. Another recent research introduces a new collaborative training model that optimizes fine-tuning of early layers at the edge device (mobile) and deep layers at the edge server [95]. In this model, the aim is to reduce communication and energy costs while keeping the personalized performance constant.

Collaborative and Split Inference: Another rational approach is to distribute the LLM model across the device-edge-cloud heterogeneity to achieve the balance of performance and resource utilization. Yang et al. [96] propose a structure in which the cloud and edge are jointly used to offload LLM inference with a UCB-based scheduler. The results show that the energy usage is halved and the efficiency is doubled. Another approach is to share the converter layers among the heterogeneous edge devices using the matching theory [97].

System-Level Runtime Optimizations: Another method of increasing efficiency is runtime and architecture-based approaches:

- **KV-cache compression**: It is the process of reorganizing memory to save RAM on edge devices [93]
- **Contextual sparsity and batch-aware scheduling**: In contextual sparsity, the process of reducing the processing load by looking only at the important tokens of the model, while in batch-aware scheduling, the process of running multiple tasks in the most efficient order without blocking each other [93], [95].
- **Speculative decoding**: It is the process of predicting multiple tokens at the same time, and the aim is to reduce the autoregressive latency bottleneck [90].

Table 7 summarizes the main optimization techniques developed against resource constraints on edge devices, their usage scenarios, and the cost/benefit balances they bring.

1) FEDERATED OR DISTRIBUTED TRAINING OF LLMs AT THE EDGE

With the proliferation of user-centric applications, it is expected that deployment strategies of LLM models will be developed on edge devices as well [87]. Cloud-based systems cause additional latency and bandwidth loads in 6G environments compared to edge computing [1]. Therefore, Federated Learning (FL) and distributed tuning paradigms can be used to reduce these loads by processing data on edge devices [1].

a: MOTIVATION FOR FEDERATED EDGE TRAINING

Compared to traditional cloud-based systems, training on edge devices provides improvements in the following limitations [3], [90], [98]:

- **Privacy**: Since sensitive data, such as biometric data, is processed on edge devices, there are fewer privacy concerns than cloud systems that use central servers.
- **Latency**: Since data is processed close to the data source, there is less latency than cloud-based systems.
- **Bandwidth**: Since cloud-based systems are used only for operations that require large processing power, unnecessary communication bandwidth is not used.

TABLE 7. Optimization techniques for deploying LLMs under edge constraints.

Reference	Technique	Constraint Addressed	Use Case	Trade-off
Wang et al. [92]	INT4/INT8 Quantization	Memory, Compute	Mobile Edge Inference	↓Accuracy, ↑Speed
Wang et al. [89]	Compression-aware Quantization	Memory, Latency	Smartphone + AI Assistant	Low overhead
Qin et al. [94]	LoRA / Adapters	Fine-Tuning cost, Storage	Personalized edge assistants	↓Flexibility, ↑Privacy
Yang et al. [96]	PerLLM Scheduler	QoS-aware Scheduling	Edge-Cloud Mixed Load	↑Efficiency, ↑Throughput
Zhao et al. [90]	Token Parallel Decoding	Latency (token gen)	Edge-terminal co-inference	Complex sync
Picano et al. [97]	Matching-based Layer Placement	Device Heterogeneity	Heterogeneous Edge Inference	↑Accuracy

TABLE 8. Federated and distributed LLM training: Constraints and LLM-Based Trade-offs.

Work	Constraint Addressed	LLM-Based Technique / Trade-off
Liu et al. [95]	Bandwidth, Memory	LoRA-Based FL for Lightweight Personalization (Lower Global Accuracy)
Qu et al. [98]	Compute Offloading	Split Learning (MEI4LLM) with Multi-layer Collaboration (Sync Overhead)
Zhao et al. [90]	Latency, Energy	Parallel Token Learning at Edge-Terminals (Complex Scheduling)
Qin et al. [94]	Fine-tuning Privacy	Federated Prompting + PEFT (Bias Sensitivity in User-Tuning)
Yang et al. [97]	Device Reliability	Trust-Aware Aggregation in FL (Risk of Model Divergence)

b: FEDERATED FINE-TUNING TECHNIQUES FOR LLMS

In constrained environments (such as communication and computation), LLM model personalization schemes can be summarized as follows:

- **Parameter efficient**: In LoRA-based work, low-rank matrices are fine-tuned among clients to reduce transmission volume [95].
- **Split Federation Learning**: Qu et al. [98] propose to train the first layers of the model on the device and optimize the deep layers on edge nodes in their proposed framework called Mobile Edge Intelligence (MEI).
- **Inter-device gradient fusion**: In order to dynamically balance the update frequency and energy budgets, distributed scheduling algorithms are proposed in [90].

c: CHALLENGES IN FEDERATED LLM TRAINING

Despite the advantages of low latency and low bandwidth overhead, federated training at the edge also suffers from statistical heterogeneity [92], system heterogeneity [99], and security risks [3]. To overcome these challenges, recent research focuses on model-system co-design based techniques. These techniques include adaptive aggregation [96] where clients are weighted according to their trust scores, compression-aware updates [89] where updates are sparse before transmission, and energy-aware scheduling [93] where the training frequency is dynamically adjusted to preserve battery and network life. Table 8 provides a comparative overview of federated and distributed education strategies.

D. DATASETS AND MODELING TECHNIQUES FOR LLM-DRIVEN APT DETECTION (RQ4)

The quality of the datasets to be used to train models in LLM-based APT detection in 6G networks directly affects the success rate. The datasets created as a result of examining 32 different studies and the results of the systematic and taxonomy study on modeling techniques are examined in this subsection. Figure 16 shows this taxonomy and its subsections.

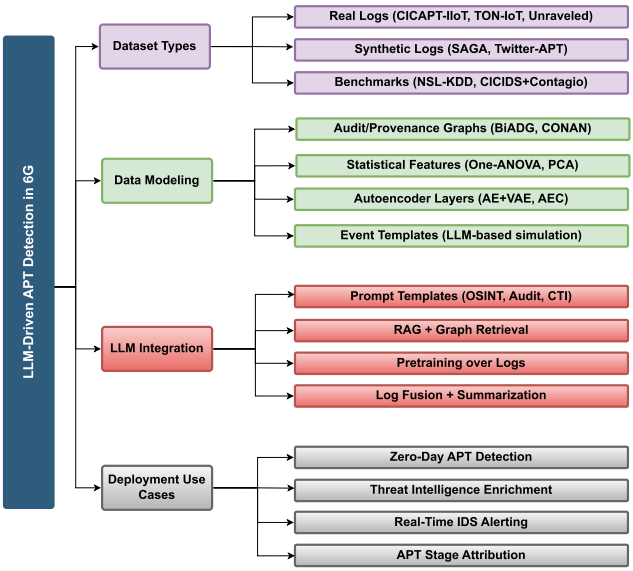

FIGURE 16. Taxonomy of Dataset-Model-LLM alignment in 6G APT detection pipelines.

1) DATASET TYPES FOR LLM-BASED APT DETECTION

When the literature is examined, it is seen that APT datasets can be examined under three main headings:

- **Semi-Synthetic Datasets**: Semi-synthetic datasets that model APT attacks are as follows: (i) Unraveled dataset [100], which combines real cloud infrastructure logs and simulated APT stages, and (ii) edge-based CICAPT-IIoT dataset [101], which includes UAV and ICS smart environment logs.
- **Synthetic and Augmented Logs**: Synthetic datasets that model APT attacks are as follows: (i) SAGA [102], which consists of audit logs compatible with the ATT&CK matrix, (ii) Twitter-APT [103], which is created by applying LLMs to OSINT-based threats, (iii) a dataset where labeled attacks are created using pcap filters, log records and IDS simulation [104].
- **Merged Benchmark Corpora**: For APT detection, datasets that are based on a real organization's network traffic and model attacks such as trojans and spyware can be used [105].

NSL-KDD or CICIDS data outside these categories are now outdated and fail to model real APTs (stealth lateral movement or long-term dormancy strategies) [106], [107]

2) DATA MODELING TECHNIQUES AND REPRESENTATIONS

Data modeling strategies that can be used to combine 6G network data with LLM models can be summarized as follows:

- **Behavioral Graph Profiling**: In this modeling method, BiADG and MIG models are obtained by applying Graph Convolutional Network (GCN) on IP flow graphs and behavior patterns [108], [109]. In addition, there is the CONAN model that provides low-latency matching for APT stages using a Finite State Machine (FSM) [110].
- **Statistical + Feature Engineering Pipelines**: As an example of these strategies, two separate studies that apply preprocessing such as One-ANOVA based cleaning, decomposition, and boosting by synthetic generation [104], [105] can be given as examples.
- **Multi-Stage Autoencoders**: In APTSID [105], where this strategy is applied, standard and variational autoencoders are combined with statistical feature selection to achieve high accuracy anomaly detection.
- **ML + Expert System Hybrids**: In the CDT system [104], where this technique is used, an attack detection prediction is taken with an ML model and transmitted to the rules used by systems such as SMORT.

Table 9 shows the comparison of datasets and modeling techniques in LLM-Driven APT detection studies

3) LLM INTEGRATION STRATEGIES

These are the methods used when integrating LLM models into various systems, and the main purpose is to enable LLM models to be used with various data.

- **Prompt Templates + Simulation**: LLM prompts are the methods used to generate attack data and multiply training data, and SAGA and CyExec are two examples of this in academia [102], [111].
- **OSINT + NER Pipelines**: Although LLM models are successful in detecting threats in open source articles, fine-tuning is required for small details. Shafee et al. [103] tries to find threats from open source information with LLM.
- **Fusion Architectures**: In these methods, after the data is processed with other models and made meaningful, it is given to the LLM model to process, and thus it is expected that LLM will perform a more successful analysis. Models such as AE+VAE and AE-CNN use this method [105]

E. REPRODUCIBILITY AND PUBLICATION TRENDS IN LLM-BASED APT STUDIES (RQ5)

This research question questions the reproducibility and other statistical information of LLM-based APT detection studies. In order to provide a comprehensive assessment of the 142 recent studies utilized throughout the paper, we have classified all papers in **Appendix A** according to *Code Availability*, *Dataset Evaluation*, *Protocol Venue/Platform*, and *Year*. The description of these features and the resulting statistical information are as follows:

1) CODE AVAILABILITY

This feature was used to classify studies according to their reproducibility. Figure 17 shows how many percentage of the studies shared their source code (YES/NO), and on which platform (Github, etc.) they were published. As can be seen from the figure, only a very small portion of the examined studies shared their source code, while most of their code was published on the GitHub platform.

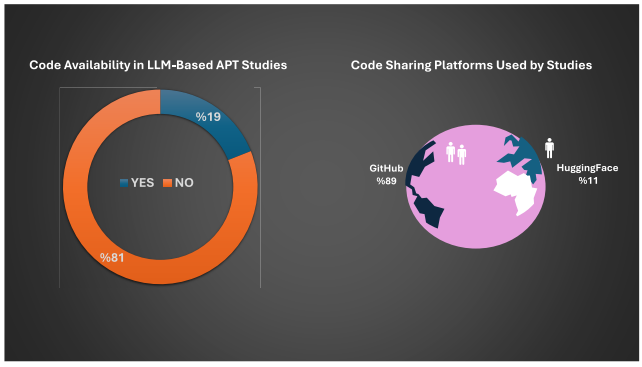

FIGURE 17. The percentage of LLM-based APT studies that shared source code (YES/NO) and the platforms where the code was hosted.

2) DATASET

This column was used to measure the diversity of datasets used in the studies and to determine how many of them used real-world data. Figure 18 shows the percentage of datasets shared by year and the percentage of articles using synthetic-public datasets. The results confirm that datasets used in APT detection studies tend to be shared and that the most used dataset is synthetic dataset.

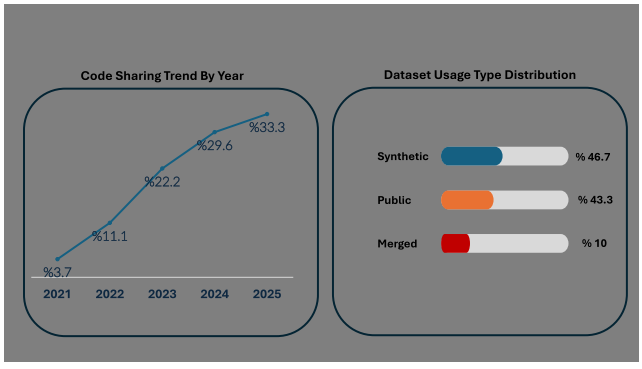

FIGURE 18. The trend of dataset usage over the years and most commonly used datasets in LLM-based APT studies.

TABLE 9. Comparison of datasets and modeling techniques in LLM-driven APT detection studies.

Reference	Dataset	Modeling Technique	LLM Use	Key Insight
Huang et al. [102]	SAGA (Synthetic)	Prompt-Based Log Generation	Training Input	ATT&CK-aligned, synthetic audit logs for APT stages
Neuschmied et al. [105]	CICIDS + Contagio	AE + VAE Stack	Feature Compression	Multi-stage anomaly detection with zero-day support
Al-Aamri et al. [104]	Custom Logs	CDT + SNORT Rule Feed	Manual LLM Friendly	Time-series + journaling logs with rule generation
Ghiasvand et al. [101]	CICAPT-IIoT	Provenance + Network Flow Fusion	LLM-Graph Possible	Audit trails + flow logs for IIoT APT detection
Shafee et al. [103]	Twitter Corpus	OSINT Classification + NER	NER and Prompt Evaluation	LLMs need domain adaptation for threat-level NER
Xuan et al. [108]	BiADG	Behavioral GCN + LSTM	Graph-to-LLM Potential	IP-node behavior modeling over graph structures
Oleiwi et al. [106]	NSL / CICIDS / UNSW	Meta-Model Voting Ensemble	Pre-LLM Classifier Layer	Traditional ML stack for high-precision filtering

3) EVALUATION PROTOCOL

This column is to evaluate the level of empirical validity of the reviewed articles based on whether they use robust protocols such as cross-validation. Figure 19 shows the frequency of the protocols used.

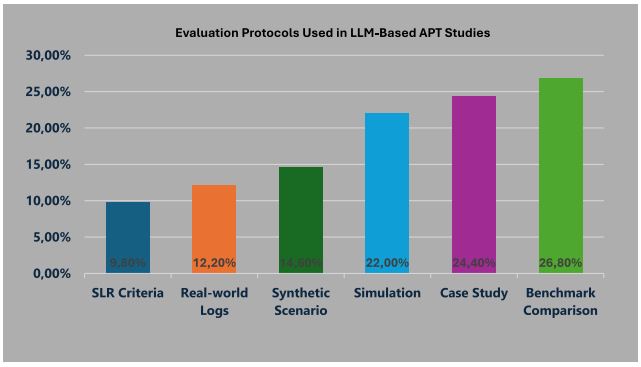

FIGURE 19. Evaluation protocols used in LLM-based apt studies.

4) VENUE / PLATFORM

This column examines the publication quality and field spread by examining the venue/platforms and types (conference/journal) where the reviewed studies were published. Figure 20 shows a summary of this statistic.

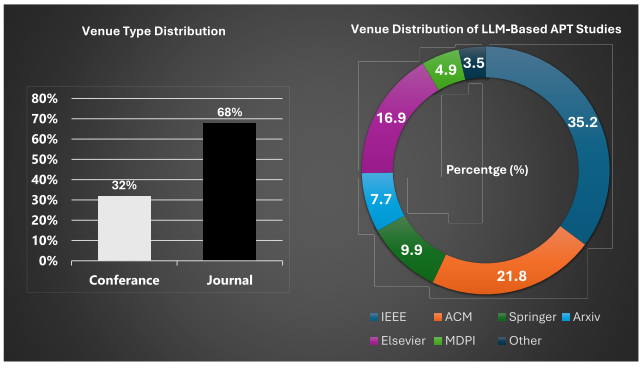

FIGURE 20. Distribution of venues where LLM-based APT studies were published and their types.

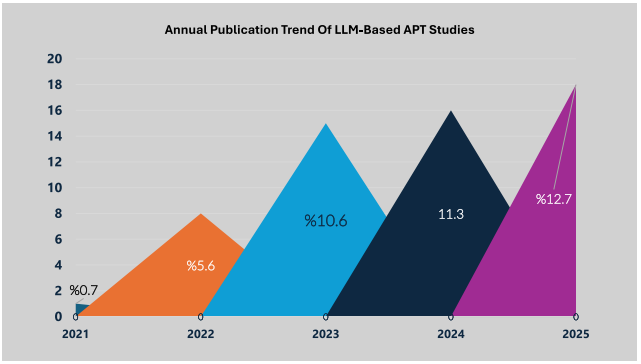

FIGURE 21. The annual trend showing the percentage of LLM-based APT studies published each year.

5) YEAR

The last column shares the publication dates of the reviewed studies and evaluates the increase in LLM-focused APT papers as we move towards the 6G wireless networks era. Figure 21 shows the change in LLM-focused APT papers by year.

VI. OPEN CHALLENGES AND FUTURE DIRECTIONS

As the use of 6G networks and LLM deployments in 6G becomes widespread, many research gaps and new open challenges to be solved will emerge for researchers. These challenges include architecture and security issues, and we discuss the open challenges, a taxonomy of which is given in Figure 22, in this section.

A. SEMANTIC-AWARE REASONING AND LIMITED CONTEXTUAL MEMORY

LLMs are promising for APT detection in 6G networks with their high performance in understanding causal relationships and threat contexts using data such as system logs and audit trails [106]. However, LLM models have limited performance in long-term and fragmented event sequences because their architectures offer limited window sizes and context management. Therefore, it makes detection difficult in multi-stage APTs with long processes such as infiltration and reconnaissance.

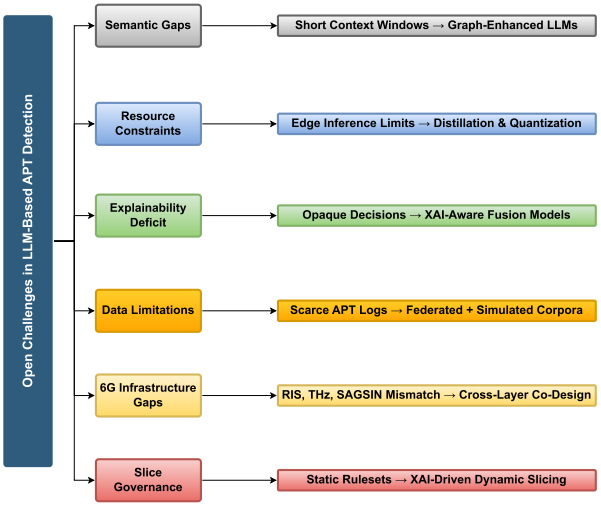

FIGURE 22. Taxonomy of open challenges in LLM-based APT detection.

Future Directions: Future researchers can overcome these limitations by focusing on the integration of memory modules and hierarchical memory structures. These structures make it easier for the model to learn long-term correlations between events. An example of this is the modeling of the relationship between system input and data exfiltration behavior to understand the holistic behavior of an attack. In addition, models that can provide transformer-graph synergy, such as GNNs, can effectively establish topological or temporal relationships between events [18]. Thus, event traces can be modeled over a graph structure, enabling LLM models to learn event dependencies in a scalable manner.

B. REAL-TIME PROCESSING UNDER EDGE CONSTRAINTS

6G lines provide great advantages for time-constrained scenarios such as autonomous vehicles by offering high speed and low latency [63]. In particular, since edge devices in their heterogeneous structure bring the processing power closer to the data source, it will enable real-time data processing with low latency and low bandwidth usage [43]. However, since LLM models require high processing power and most edge devices have low capacity and low processing power, this poses a serious challenge.

Future Directions: Future researchers can work on threat-aware and edge-adaptive LLM models to overcome this challenge. For this, some techniques such as knowledge distillation, quantization, and edge-aware fine-tuning come to the fore. These techniques are explained in detail in section V-C.

C. LACK OF GROUNDED EXPLAINABILITY

Although LLM-based models have great potential in cybersecurity, such as APT detection, most models can cause serious security vulnerabilities in mission-critical tasks due to their non-interpretable model nature [24]. For this reason, it is necessary to understand the inputs and probabilities that the model uses when making this decision. However, there is no system that shows the necessary causal traceability and root cause reasoning information in LLM models to make this understanding. This causes gaps in forensic analysis, such as explaining attacks and auditing sources.

Future Directions: Future researchers can design more transparent systems by examining network slicing and decision-making processes for LLM models with new Explainable Artificial Intelligence (XAI) frameworks. More transparent information can be obtained with techniques that explain the training phase of models and the output phase of models, especially pre-hoc XAI and post-hoc XAI.

D. SCARCITY OF FINE-GRAINED LLM TRAINING DATA

Data quality has a major impact on the predictive performance of LLM models, and the volume and variety of APT-related data used in the existing literature are limited in terms of real-world representation [101]. Studies (see section V-E) show that most studies rely on synthetic audit logs or CTIs with limited content. This limits the generalizability of LLM models and their ability to detect threats in different environments.

Future Directions: To overcome this data limitation issue, steps can be taken such as collaboration between researchers and organizations (public-private), development of benchmark datasets, and modeling of attack progression scenarios.

E. INTEGRATION WITH EMERGING 6G TECHNOLOGIES

In addition to high data rates in 6G networks, new generation technologies such as intelligent reflective surfaces (IRS) and terahertz (THz) band communication will also provide more dynamic and uninterrupted communication opportunities [23]. However, this also brings some challenges such as synchronization, spectrum sharing, and secure orchestration. Adaptation of current LLM-based APT systems to such complex and multi-layered environments requires great attention.

Future Directions: To overcome these challenges, researchers can develop multi-layered security protocols. In this way, 6G networks will gain threat perception and response cycle capability in ultra-dynamic and variable environments.

F. UNDEREXPLORED ROLE OF NETWORK SLICING AND XAI FUSION

With the network slicing feature, 6G networks can run mission-critical scenarios such as autonomous vehicle communication and industrial control on dedicated and isolated resources [24]. However, incorrect resource allocations and unexpected load shifts that may occur during slicing operations can cause serious security problems [36]. An example is the autonomous vehicle experiencing signal delays that are beyond the delay tolerance due to network slicing.

Although XAI techniques provide dynamic adaptation capabilities in network slicing, the use of these capabilities in real-time environments is still limited [24]. Since most of the research is theory or simulation-based, its use with real-world data from SDN infrastructures needs to be investigated.

Future Directions: To overcome these limitations, researchers can develop slice-aware and network state-oriented LLM models, and these models should be able to dynamically adjust network slice configurations and allocation policies by continuously monitoring real-time data. In addition, XAI techniques can be integrated with the obtained decisions to provide traceable and reliable information for network operators.

VII. CONCLUSION

This paper presents a comprehensive systematic review and taxonomy, the first of its kind, for LLM-based Advanced Persistent Threat (APT) detection in 6G networks. Findings from 142 recent papers examine the interaction between the capabilities (semantics) of LLM models and the challenges (architecture, privacy, etc.) of 6G environments. We aim to provide new insights for future research by presenting a taxonomy covering input types, model techniques, deployment settings, and threat lifecycle stages. Although LLM has great potential in APT attack detection, it also has limitations such as limited context memory, opaque decision processes, and real-time inference at the edge. In addition, reproducibility and dataset generalizability stand out as important obstacles for research in this area. Based on the findings, we call for joint efforts in the following research areas:

- Designing lightweight, unified LLMs for edge devices in 6G networks,
- Investigating new XAI-driven decision monitoring mechanisms to increase transparency of LLMs,
- Enriching datasets used for APT detection using fine-grained, multimodal, and real-world data,
- Integrating LLMs with slicing-aware orchestration systems in 6G for dynamic demands on 6G links.

REFERENCES

[1] M. A. Ferrag, O. Friha, B. Kantarci, N. Tihanyi, L. Cordeiro, M. Debbah, D. Hamouda, M. Al-Hawawreh, and K.-K.-R. Choo, "Edge learning for 6G-enabled Internet of Things: A comprehensive survey of vulnerabilities, datasets, and defenses," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 4, pp. 2654–2713, 4th Quart., 2023.

[2] D. P. M. Osorio, I. Ahmad, J. D. V. Sánchez, A. Gurtov, J. Scholliers, M. Kutila, and P. Porambage, "Towards 6G-enabled Internet of Vehicles: Security and privacy," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 82–105, 2022.

[3] S. Ali Khowaja, P. Khuwaja, K. Dev, H. A. Hamadi, and E. Zeydan, "Pathway to secure and trustworthy ZSM for LLMs: Attacks, defense, and opportunities," 2024, *arXiv:2408.00722*.

[4] O. G. Lira, A. Marroquín, and M. A. To, "Harnessing the advanced capabilities of LLM for adaptive intrusion detection systems," in *Proc. Int. Conf. Adv. Inf. Netw. Appl.*, 2024, pp. 453–464.

[5] B. Gulbay and M. Demirci, "APT-scope: A novel framework to predict advanced persistent threat groups from enriched heterogeneous information network of cyber threat intelligence," *Eng. Sci. Technol., Int. J.*, vol. 57, Sep. 2024, Art. no. 101791.

[6] M. A. Rahman and M. S. Hossain, "A deep learning assisted software defined security architecture for 6G wireless networks: IIoT perspective," *IEEE Wireless Commun.*, vol. 29, no. 2, pp. 52–59, Apr. 2022.

[7] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering—A systematic literature review," *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, Jan. 2009.

[8] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *Proc. Electron. Workshops Comput.*, Jun. 2008.

[9] M. Hassanin and N. Moustafa, "A comprehensive overview of large language models (LLMs) for cyber defences: Opportunities and directions," 2024, *arXiv:2405.14487*.

[10] H. Xu, S. Wang, N. Li, K. Wang, Y. Zhao, K. Chen, T. Yu, Y. Liu, and H. Wang, "Large language models for cyber security: A systematic literature review," 2024, *arXiv:2405.04760*.

[11] H. Kheddar, "Transformers and large language models for efficient intrusion detection systems: A comprehensive survey," 2024, *arXiv:2408.07583*.

[12] S. Tian, T. Zhang, J. Liu, J. Wang, X. Wu, X. Zhu, R. Zhang, W. Zhang, Z. Yuan, S. Mao, and D. In Kim, "Exploring the role of large language models in cybersecurity: A systematic survey," 2025, *arXiv:2504.15622*.

[13] Y. Chen, M. Cui, D. Wang, Y. Cao, P. Yang, B. Jiang, Z. Lu, and B. Liu, "A survey of large language models for cyber threat detection," *Comput. Secur.*, vol. 145, Oct. 2024, Art. no. 104016.

[14] F. Zuo, J. Rhee, and Y. Ryn Choe, "Knowledge transfer from LLMs to provenance analysis: A semantic-augmented method for APT detection," 2025, *arXiv:2503.18316*.

[15] Z. Liu, "A review of advancements and applications of pre-trained language models in cybersecurity," in *Proc. 12th Int. Symp. Digit. Forensics Secur. (ISDFS)*, Apr. 2024, pp. 1–10.

[16] S. Singh, P. K. Sharma, S. Y. Moon, D. Moon, and J. H. Park, "A comprehensive study on APT attacks and countermeasures for future networks and communications: Challenges and solutions," *J. Supercomput.*, vol. 75, no. 8, pp. 4543–4574, Aug. 2019.

[17] S. Krishnapriya and S. Singh, "A comprehensive survey on advanced persistent threat (APT) detection techniques," *Comput., Mater. Continua*, vol. 80, no. 2, pp. 2675–2719, 2024.

[18] Z. Xu, Y. Wu, S. Wang, J. Gao, T. Qiu, Z. Wang, H. Wan, and X. Zhao, "Deep learning-based intrusion detection systems: A survey," 2025, *arXiv:2504.07839*.

[19] L.-H. Shen, K.-T. Feng, and L. Hanzo, "Five facets of 6G: Research challenges and opportunities," *ACM Comput. Surveys*, vol. 55, no. 11, pp. 1–39, Nov. 2023.

[20] A. Blika, S. Palmos, G. Doukas, V. Lamprou, S. Pelekis, M. Kontoulis, C. Ntanos, and D. Askounis, "Federated learning for enhanced cybersecurity and trustworthiness in 5G and 6G networks: A comprehensive survey," *IEEE Open J. Commun. Soc.*, vol. 6, pp. 3094–3130, 2025.

[21] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, D. Niyato, O. Dobre, and H. V. Poor, "6G Internet of Things: A comprehensive survey," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 359–383, Jan. 2022.

[22] R. Alghamdi, R. Alhadrami, D. Alhothali, H. Almorad, A. Faisal, S. Helal, R. Shalabi, R. Asfour, N. Hammad, A. Shams, N. Saeed, H. Dahrouj, T. Y. Al-Naffouri, and M.-S. Alouini, "Intelligent surfaces for 6G wireless networks: A survey of optimization and performance analysis techniques," *IEEE Access*, vol. 8, pp. 202795–202818, 2020.

[23] Y. Zhao, W. Zhai, J. Zhao, T. Zhang, S. Sun, D. Niyato, and K.-Y. Lam, "A comprehensive survey of 6G wireless communications," 2020, *arXiv:2101.03889*.

[24] H. Sun, Y. Liu, A. Al-Tahmeesschi, A. Nag, M. Soleimanpour, B. Canberk, H. Arslan, and H. Ahmadi, "Advancing 6G: Survey for explainable AI on communications and network slicing," *IEEE Open J. Commun. Soc.*, vol. 6, pp. 1372–1412, 2025.

[25] R. Ross, V. Pillitteri, G. Guissanie, R. Wagner, R. Graubart, and D. Bodeau, "Enhanced security requirements for protecting controlled unclassified information: A supplement to nist special publication 800–171," Nat. Inst. Standards Technol., Tech. Rep., 2020.

[26] K. Xing, A. Li, R. Jiang, and Y. Jia, "A review of APT attack detection methods and defense strategies," in *Proc. IEEE 5th Int. Conf. Data Sci. Cyberspace (DSC)*, Jul. 2020, pp. 67–70.

[27] D. Arulkumar, K. Kartheeban, and G. Arulkumaran, "The APT cyber warriors with TTP weapons to battle: An review on IoT and cyber twin," in *New Approaches To Data Analytics and Internet of Things Through Digital Twin*. IGI Global, 2023, pp. 211–225.
[28] Y. Lv, S. Qin, Z. Zhu, Z. Yu, S. Li, and W. Han, "A review of provenance graph based APT attack detection: Applications and developments," in *Proc. 7th IEEE Int. Conf. Data Sci. Cyberspace (DSC)*, Jul. 2022, pp. 498–505.
[29] S. Ullah, J. Li, J. Chen, I. Ali, S. Khan, A. Ahad, F. Ullah, and V. C. M. Leung, "A survey on emerging trends and applications of 5G and 6G to healthcare environments," *ACM Comput. Surveys*, vol. 57, no. 4, pp. 1–36, Apr. 2025.
[30] Q. Song and Z. Wang, "An empirical study on teaching innovation of traditional wushu culture from the perspective of 6G network," *Int. J. e-Collaboration*, vol. 21, no. 1, pp. 1–18, Jan. 2025.
[31] S. Ariyanti and M. Suryanegara, "Visible light communication (VLC) for 6G technology: The potency and research challenges," in *Proc. 4th World Conf. Smart Trends Syst., Secur. Sustainability (WorldS4)*, Jul. 2020, pp. 490–493.
[32] M. Elamassie and M. Uysal, "Free space optical communication: An enabling backhaul technology for 6G non-terrestrial networks," *Photonics*, vol. 10, no. 11, p. 1210, Oct. 2023.
[33] B. Mao, J. Liu, Y. Wu, and N. Kato, "Security and privacy on 6G network edge: A survey," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 1095–1127, 2nd Quart., 2023.
[34] P. Bhide, D. Shetty, and S. Mikkili, "Review on 6G communication and its architecture, technologies included, challenges, security challenges and requirements, applications, with respect to AI domain," *IET Quantum Commun.*, vol. 6, no. 1, Jan. 2025.
[35] N. Singh, R. Devi, J. Singh, B. Rajalakshmi, M. W. Mohammed, N. Thandra, and C. Vidyadhari, "AI-powered 6G networks: Transforming wireless communication with intelligence and terahertz waves," in *Proc. Int. Conf. Intell. Control, Comput. Commun. (IC3)*, Feb. 2025, pp. 744–750.
[36] I. H. Abdulqadder and S. Zhou, "SliceBlock: Context-aware authentication handover and secure network slicing using DAG-blockchain in edge-assisted SDN/NFV-6G environment," *IEEE Internet Things J.*, vol. 9, no. 18, pp. 18079–18097, Sep. 2022.
[37] Y. Siriwardhana, P. Porambage, M. Liyanage, and M. Ylianttila, "AI and 6G security: Opportunities and challenges," in *Proc. Joint Eur. Conf. Netw. Commun. 6G Summit (EuCNC/6G Summit)*, Jun. 2021, pp. 616–621.
[38] F. Naeem, M. Ali, G. Kaddoum, C. Huang, and C. Yuen, "Security and privacy for reconfigurable intelligent surface in 6G: A review of prospective applications and challenges," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1196–1217, 2023.
[39] J. Suomalainen, I. Ahmad, A. Shajan, and T. Savunen, "Cybersecurity for tactical 6G networks: Threats, architecture, and intelligence," *Future Gener. Comput. Syst.*, vol. 162, Jan. 2025, Art. no. 107500.
[40] C. Smiliotopoulos, G. Kambourakis, and C. Kolias, "Detecting lateral movement: A systematic survey," *Heliyon*, vol. 10, no. 4, Feb. 2024, Art. no. e26317.
[41] R. Mishra, N. Chauduray, and G. Kaur, "Harnessing machine learning for APTs detection and mitigation in large-scale networks," in *Proc. 7th Int. Conf. Contemp. Comput. Informat. (IC3I)*, Sep. 2024, pp. 565–571.
[42] S. M. Milajerdi, R. Gjomemo, B. Eshete, R. Sekar, and V. N. Venkatakrishnan, "HOLMES: Real-time APT detection through correlation of suspicious information flows," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2019, pp. 1137–1152.
[43] J. Yang, J. Zheng, Z. Zhang, Q. Chen, D. S. Wong, and Y. Li, "Security of federated learning for cloud-edge intelligence collaborative computing," *Int. J. Intell. Syst.*, vol. 37, no. 11, pp. 9290–9308, 2022.
[44] A. Bertrand, R. Belloum, J. R. Eagan, and W. Maxwell, "How cognitive biases affect XAI-assisted decision-making: A systematic review," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Jul. 2022, pp. 78–91.
[45] J. Zhang, H. Wen, L. Li, and H. Zhu, "UniTTP: A unified framework for tactics, techniques, and procedures mapping in cyber threats," in *Proc. IEEE 23rd Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom)*, Dec. 2024, pp. 1580–1588.
[46] R. Agrawal, H. Kumar, and S. R. Lnu, "Efficient LLMs for edge devices: Pruning, quantization, and distillation techniques," in *Proc. Int. Conf. Mach. Learn. Auto. Syst. (ICMLAS)*, Mar. 2025, pp. 1413–1418.
[47] D. Wu, L. Nie, R. A. Mumtaz, and K. Agarwal, "A LLM-based hybrid-transformer diagnosis system in healthcare," *IEEE J. Biomed. Health Informat.*, pp. 1–12, 2024.
[48] N. Karanikolas, E. Manga, N. Samaridi, E. Tousidou, and M. Vassilakopoulos, "Large language models versus natural language understanding and generation," in *Proc. 27th Pan-Hellenic Conf. Prog. Comput. Informat.*, Nov. 2023, pp. 278–290.
[49] C. Di Sipio, R. Rubei, J. Di Rocco, D. Di Ruscio, and L. Iovino, "On the use of LLMs to support the development of domain-specific modeling languages," in *Proc. ACM/IEEE 27th Int. Conf. Model Driven Eng. Lang. Syst.*, Sep. 2024, pp. 596–601.
[50] B. Padiu, R. Iacob, T. Rebedea, and M. Dascalu, "To what extent have LLMs reshaped the legal domain so far? A scoping literature review," *Information*, vol. 15, no. 11, p. 662, Oct. 2024.
[51] N. Shenoy and A. V. Mbaziira, "An extended review: LLM prompt engineering in cyber defense," in *Proc. Int. Conf. Electr., Comput. Energy Technol. (ICECET*, Jul. 2024, pp. 1–6.
[52] W. U. Hassan, A. Bates, and D. Marino, "Tactical provenance analysis for endpoint detection and response systems," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2020, pp. 1172–1189.
[53] Y. Liu, X. Shu, Y. Sun, J. Jang, and P. Mittal, "RAPID: Real-time alert investigation with context-aware prioritization for efficient threat discovery," in *Proc. 38th Annu. Comput. Secur. Appl. Conf.*, Dec. 2022, pp. 827–840.
[54] R. Buchta, G. Gkoktsis, F. Heine, and C. Kleiner, "Advanced persistent threat attack detection systems: A review of approaches, challenges, and trends," *Digit. Threats, Res. Pract.*, vol. 5, no. 4, pp. 1–37, Dec. 2024.
[55] M. Golec, S. S. Gill, R. Bahsoon, and O. Rana, "BioSec: A biometric authentication framework for secure and private communication among edge devices in IoT and industry 4.0," *IEEE Consum. Electron. Mag.*, vol. 11, no. 2, pp. 51–56, Mar. 2022.
[56] M. Golec, S. S. Gill, A. K. Parlikad, and S. Uhlig, "HealthFaaS: AI-based smart healthcare system for heart patients using serverless computing," *IEEE Internet Things J.*, vol. 10, no. 21, pp. 18469–18476, Nov. 2023.
[57] H. Hafi, B. Brik, P. A. Frangoudis, A. Ksentini, and M. Bagaa, "Split federated learning for 6G enabled-networks: Requirements, challenges, and future directions," *IEEE Access*, vol. 12, pp. 9890–9930, 2024.
[58] M. Tao, Y. Zhou, Y. Shi, J. Lu, S. Cui, J. Lu, and K. B. Letaief, "Federated edge learning for 6G: Foundations, methodologies, and applications," *Proc. IEEE*, pp. 1–39, 2024.
[59] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly," *High-Confidence Comput.*, vol. 4, no. 2, Jun. 2024, Art. no. 100211.
[60] X. Wang, L. Xu, L. Zhou, Y. Liu, N. Xiong, and K.-C. Li, "Large language model-driven probabilistic trajectory prediction in the Internet of Things using spatio-temporal encoding and normalizing flows," *Digit. Commun. Netw.*, Mar. 2025.
[61] O. Friha, M. A. Ferrag, B. Kantarci, B. Cakmak, A. Ozgun, and N. Ghoualmi-Zine, "LLM-based edge intelligence: A comprehensive survey on architectures, applications, security and trustworthiness," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 5799–5856, 2024.
[62] M. Golec, G. K. Walia, M. Kumar, F. Cuadrado, S. S. Gill, and S. Uhlig, "Cold start latency in serverless computing: A systematic review, taxonomy, and future directions," *ACM Comput. Surv.*, vol. 57, no. 3, pp. 1–36, Mar. 2025.
[63] C. Sergiou, M. Lestas, P. Antoniou, C. Liaskos, and A. Pitsillides, "Complex systems: A communication networks perspective towards 6G," *IEEE Access*, vol. 8, pp. 89007–89030, 2020.
[64] A. K. Sood, S. Zeadally, and E. Hong, "The paradigm of hallucinations in AI-driven cybersecurity systems: Understanding taxonomy, classification outcomes, and mitigations," *Comput. Electr. Eng.*, vol. 124, May 2025, Art. no. 110307.
[65] H. Alturkistani and S. Chuprat, "Artificial intelligence and large language models in advancing cyber threat intelligence: A systematic literature review," Tech. Rep., 2024.
[66] S. Benabderrahmane, P. Valtchev, J. Cheney, and T. Rahwan, "APT-LLM: Embedding-based anomaly detection of cyber advanced persistent threats using large language models," 2025, *arXiv:2502.09385*.
[67] P. Atulbhai Gandhi, P. N. Wudali, Y. Amaru, Y. Elovici, and A. Shabtai, "SHIELD: APT detection and intelligent explanation using LLM," 2025, *arXiv:2502.02342*.

[68] M. He, T. Jia, C. Duan, H. Cai, Y. Li, and G. Huang, "LLMeLog: An approach for anomaly detection based on LLM-enriched log events," in *Proc. IEEE 35th Int. Symp. Softw. Rel. Eng. (ISSRE)*, Oct. 2024, pp. 132–143.

[69] S. Shan, Y. Huo, Y. Su, Y. Li, D. Li, and Z. Zheng, "Face it yourselves: An LLM-based two-stage strategy to localize configuration errors via logs," in *Proc. 33rd ACM SIGSOFT Int. Symp. Softw. Test. Anal.*, Sep. 2024, pp. 13–25.

[70] D. R. Arikkat, R. Rehiman, S. Nicolazzo, M. Arazzi, A. Nocera, and M. Conti, "DroidTTP: Mapping Android applications with TTP for cyber threat intelligence," 2025, *arXiv:2503.15866*.

[71] Y. Zhang, X. Zhou, H. Wen, W. Niu, J. Liu, H. Wang, and Q. Li, "Tactics, techniques, and procedures (TTPs) in interpreted malware: A zero-shot generation with large language models," 2024, *arXiv:2407.08532*.

[72] V. N. Ignatyev, N. V. Shimchik, D. D. Panov, and A. A. Mitrofanov, "Large language models in source code static analysis," in *Proc. Ivannikov Memorial Workshop (IVMEM)*, May 2024, pp. 28–35.

[73] H. Ji, J. Yang, L. Chai, C. Wei, L. Yang, Y. Duan, Y. Wang, T. Sun, H. Guo, T. Li, C. Ren, and Z. Li, "SEvenLLM: Benchmarking, eliciting, and enhancing abilities of large language models in cyber threat intelligence," 2024, *arXiv:2405.03446*.

[74] X. Li, Y. Huo, C. Mao, S. Shan, Y. Su, D. Li, and Z. Zheng, "AnomalyGen: An automated semantic log sequence generation framework with LLM for anomaly detection," 2025, *arXiv:2504.12250*.

[75] J. Huang, Z. Jiang, Z. Chen, and M. R. Lyu, "LUNAR: Unsupervised LLM-based log parsing," 2024, *arXiv:2406.07174*.

[76] J. Koenders, "Advancing cyberdefense through bert: A natural language processing approach for vulnerability to attack mapping within a responsible artificial intelligence framework," Tech. Rep., 2024.

[77] J. Wang, T. Zhu, C. Xiong, and Y. Chen, "MultiKG: Multi-source threat intelligence aggregation for high-quality knowledge graph representation of attack techniques," 2024, *arXiv:2411.08359*.

[78] D. Du, X. Guan, Y. Liu, B. Jiang, S. Liu, H. Feng, and J. Liu, "MAD-LLM: A novel approach for alert-based multi-stage attack detection via LLM," in *Proc. IEEE Int. Symp. Parallel Distrib. Process. with Appl. (ISPA)*, Oct. 2024, pp. 2046–2053.

[79] L. Wang, Z. Li, Y. Jiang, Z. Wang, Z. Guo, J. Wang, Y. Wei, X. Shen, W. Ruan, and Y. Chen, "From sands to mansions: Towards automated cyberattack emulation with classical planning and large language models," 2024, *arXiv:2407.16928*.

[80] S. Mitra, S. Neupane, T. Chakraborty, S. Mittal, A. Piplai, M. Gaur, and S. Rahimi, "LOCALINTEL: Generating organizational threat intelligence from global and local cyber knowledge," 2024, *arXiv:2401.10036*.

[81] M. L. Diakhame, C. Diallo, and M. Mejri, "MCM-llama: A fine-tuned large language model for real-time threat detection through security event correlation," in *Proc. Int. Conf. Electr., Comput. Energy Technol. (ICECET*, Jul. 2024, pp. 1–6.

[82] K. Yang, V. Kindratenko, and C. Zhai, "TinyHelen's first curriculum: Training and evaluating tiny language models in a simpler language environment," 2024, *arXiv:2501.00522*.

[83] P. Sun, X. Yun, S. Li, T. Yin, C. Si, and J. Xie, "AdvTG: An adversarial traffic generation framework to deceive DL-based malicious traffic detection models," in *Proc. ACM Web Conf.*, Apr. 2025, pp. 3147–3159.

[84] J. Diao, S. Zhao, J. Xie, R. Xie, and G. Shi, "Poster: DoHunter: A feature fusion-based LLM for DoH tunnel detection," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Dec. 2024, pp. 5012–5014.

[85] H. J. Hadi, Y. Cao, S. Li, L. Xu, Y. Hu, and M. Li, "Real-time fusion multi-tier DNN-based collaborative IDPS with complementary features for secure UAV-enabled 6G networks," *Expert Syst. Appl.*, vol. 252, Oct. 2024, Art. no. 124215.

[86] H. Xu, C. Si, C. Wang, P. Sun, and Q. Liu, "APTSniffer: Detecting APT attack traffic using retrieval-augmented large language models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2025, pp. 1–5.

[87] A. Liu, Y. Zhou, X. Liu, T. Zhang, S. Liang, J. Wang, Y. Pu, T. Li, J. Zhang, W. Zhou, Q. Guo, and D. Tao, "Compromising LLM driven embodied agents with contextual backdoor attacks," *IEEE Trans. Inf. Forensics Security*, vol. 20, pp. 3979–3994, 2025.

[88] W. Cheng, T. Zhu, S. Jing, J.-P. Mei, M. Ma, J. Jin, and Z. Weng, "OMNISEC: LLM-driven provenance-based intrusion detection via retrieval-augmented behavior prompting," 2025, *arXiv:2503.03108*.

[89] W. Wang, Y. Mao, D. Tang, H. Du, N. Guan, and C. Jason Xue, "When compression meets model compression: Memory-efficient double compression for large language models," 2025, *arXiv:2502.15443*.

[90] W. Zhao, W. Jing, Z. Lu, and X. Wen, "Edge and terminal cooperation enabled LLM deployment optimization in wireless network," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC Workshops)*, Aug. 2024, pp. 220–225.

[91] Z. Xu, A. Gupta, T. Li, O. Bentham, and V. Srikumar, "Beyond perplexity: Multi-dimensional safety evaluation of LLM compression," 2024, *arXiv:2407.04965*.

[92] W. Wang, W. Chen, Y. Luo, Y. Long, Z. Lin, L. Zhang, B. Lin, D. Cai, and X. He, "Model compression and efficient inference for large language models: A survey," 2024, *arXiv:2402.09748*.

[93] D. Liu, Y. Yu, Y. Wang, J. Wu, Z. Wan, S. Alinejad, B. Lengerich, and Y. N. Wu, "Designing large foundation models for efficient training and inference: A survey," 2024, *arXiv:2409.01990*.

[94] R. Qin, D. Liu, C. Xu, Z. Yan, Z. Tan, Z. Jia, A. Nassereldine, J. Li, M. Jiang, A. Abbasi, J. Xiong, and Y. Shi, "Empirical guidelines for deploying LLMs onto resource-constrained edge devices," 2024, *arXiv:2406.03777*.

[95] C. Liu and J. Zhao, "Resource allocation for stable LLM training in mobile edge computing," in *Proc. 25th Int. Symp. Theory, Algorithmic Found., Protocol Design Mobile Netw. Mobile Comput.*, Oct. 2024, pp. 81–90.

[96] Z. Yang, Y. Yang, C. Zhao, Q. Guo, W. He, and W. Ji, "PerLLM: Personalized inference scheduling with edge-cloud collaboration for diverse LLM services," 2024, *arXiv:2405.14636*.

[97] B. Picano, D. T. Hoang, and D. N. Nguyen, "A matching game for LLM layer deployment in heterogeneous edge networks," *IEEE Open J. Commun. Soc.*, vol. 6, pp. 3795–3805, 2025.

[98] G. Qu, Q. Chen, W. Wei, Z. Lin, X. Chen, and K. Huang, "Mobile edge intelligence for large language models: A contemporary survey," *IEEE Commun. Surveys Tuts.*, early access, Jan. 9, 2025, doi: 10.1109/COMST.2025.3527641.

[99] M. Zhang, X. Shen, J. Cao, Z. Cui, and S. Jiang, "EdgeShard: Efficient LLM inference via collaborative edge computing," *IEEE Internet Things J.*, vol. 12, no. 10, pp. 13119–13131, May 2025.

[100] S. Myneni, K. Jha, A. Sabur, G. Agrawal, Y. Deng, A. Chowdhary, and D. Huang, "Unraveled—A semi-synthetic dataset for advanced persistent threats," *Comput. Netw.*, vol. 227, May 2023, Art. no. 109688.

[101] E. Ghiasvand, S. Ray, S. Iqbal, S. Dadkhah, and A. A. Ghorbani, "CICAPT-IIOT: A provenance-based APT attack dataset for IIoT environment," 2024, *arXiv:2407.11278*.

[102] Y.-T. Huang, Y.-R. Guo, Y.-S. Yang, G.-W. Wong, Y.-Z. Jheng, Y. Sun, J. Modini, T. Lynar, and M. C. Chen, "SAGA: Synthetic audit log generation for APT campaigns," 2024, *arXiv:2411.13138*.

[103] S. Shafee, A. Bessani, and P. M. Ferreira, "Evaluation of LLM-based chatbots for OSINT-based cyber threat awareness," *Expert Syst. Appl.*, vol. 261, Feb. 2025, Art. no. 125509.

[104] A. S. Al-Aamri, R. Abdulghafor, S. Turaev, I. Al-Shaikhli, A. Zeki, and S. Talib, "Machine learning for APT detection," *Sustainability*, vol. 15, no. 18, p. 13820, Sep. 2023.

[105] H. Neuschmied, M. Winter, B. Stojanović, K. Hofer-Schmitz, J. Božić, and U. Kleb, "APT-attack detection based on multi-stage autoencoders," *Appl. Sci.*, vol. 12, no. 13, p. 6816, Jul. 2022.

[106] H. W. Oleiwi, D. N. Mhawi, and H. Al-Raweshidy, "A meta-model to predict and detect malicious activities in 6G-structured wireless communication networks," *Electronics*, vol. 12, no. 3, p. 643, Jan. 2023.

[107] M. Saeed, R. Saeed, M. Abdelhaq, R. Alsaqour, M. Hasan, and R. Mokhtar, "Anomaly detection in 6G networks using machine learning methods," *Electronics*, vol. 12, no. 15, p. 3300, Jul. 2023.

[108] C. D. Xuan and T. T. Nguyen, "A novel approach for APT attack detection based on an advanced computing," *Sci. Rep.*, vol. 14, no. 1, p. 22223, Sep. 2024.

[109] H. C. Nguyen, C. D. Xuan, L. T. Nguyen, and H. D. Nguyen, "A new framework for APT attack detection based on network traffic," *J. Intell. Fuzzy Syst.*, vol. 44, no. 3, pp. 3459–3474, Mar. 2023.

[110] C. Xiong, T. Zhu, W. Dong, L. Ruan, R. Yang, Y. Cheng, Y. Chen, S. Cheng, and X. Chen, "Conan: A practical real-time APT detection system with high accuracy and efficiency," *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 1, pp. 551–565, Jan. 2022.
[111] M. Mudassar Yamin, E. Hashmi, M. Ullah, and B. Katt, "Applications of LLMs for generating cyber security exercise scenarios," *IEEE Access*, vol. 12, pp. 143806–143822, 2024.

MUHAMMED GOLEC received the M.Sc. and Ph.D. degrees in computer science from the Queen Mary University of London under the prestigious Turkish Ministry of National Education Scholarship. He is currently an Assistant Professor with the Department of Computer Engineering, Boğaziçi University. During his postgraduate studies, he authored more than 40 peer-reviewed publications in high-impact journals and leading international conferences, including IEEE Transactions on Industrial Informatics, IEEE Internet of Things Journal, IEEE Transactions on Sustainable Computing, IEEE Journal of Biomedical and Health Informatics, *ACM Computing Surveys*, IEEE Transactions on Cognitive Communications and Networking, and IEEE CCGrid. He gained industry experience by participating in industry 4.0 projects as an Electrical and Electronic Maintenance Engineer at Sisecam, a global glass manufacturing company that combines theoretical knowledge with applied industrial expertise. His research interests include artificial intelligence, cloud computing, and security and privacy. He has also received multiple awards in recognition of his research and reviewing contributions: Demonstrator of the Year (2024 Queen Mary University of London), Reviewer Award (2025 Wiley TETT, 2024 Elsevier IoT, and 2023 Wiley TETT), and Top Cited Article (2022–2023 Wiley SPJ).

YASER KHAMAYSEH received the Ph.D. degree from the University of Alberta, Canada. He is currently an Associate Professor with the College of Technological Innovation, Zayed University, Abu Dhabi Campus. He has more than 15 years of experience ranging from university teaching to research and leadership roles. He is an accomplished Teacher and a Researcher with a proven record of more than 80 publications in international journals and conferences. He taught many computer science courses for both undergraduate and graduate students and supervised more than 25 master's students' theses. His research interests include various aspects of networking and its possible applications, such as the Internet of Things (IoT) and smart spaces.

SUHIB BANI MELHEM received the M.Eng. and Ph.D. degrees in electrical and computer engineering from Concordia University, Canada. He completed a Postdoctoral Fellowship at York University in collaboration with the National Research Council Canada. In 2023, he earned a Certificate in cybersecurity from Polytechnique Montréal. From May to August 2023, he was a Scientific Researcher of cybersecurity with Zayed University, United Arab Emirates, where he focused on integrating cutting-edge advancements in cybersecurity. He is currently an Assistant Professor with Al Ain University. His research interests include cloud computing, cybersecurity, information security, machine learning, the IoT, and 5G networks.

ABDULMALIK ALWARAFY (Member, IEEE) received the Ph.D. degree in computer science and engineering from Hamad Bin Khalifa University, Doha, Qatar. He is currently an Assistant Professor with the Department of Computer and Network Engineering, College of Information Technology, United Arab Emirates University (UAEU), United Arab Emirates. With over 25 publications in high-impact peer-reviewed journals and premier international conferences, his research specializes in AI-driven optimization for self-organizing heterogeneous wireless networks, intelligent radio resource management, and security and privacy in next-generation wireless systems. His work advances the theoretical and practical foundations of autonomous network orchestration, machine learning-based resource allocation, and resilient communication protocols for 6G and beyond.