

APPLICATION OF LARGE LANGUAGE MODELS IN 5G: A SYSTEMATIC LITERATURE REVIEW

Alabi Orobosade, Oluremi Daniel, Akinola Oluwatoyin, Arogundade Oluwasefunmi

Department of Computer Science, Federal University of Agriculture, Abeokuta

Department of Computer Science, Federal University of Agriculture, Abeokuta

Department of Computer Science, Federal University of Agriculture, Abeokuta

Department of Software Engineering and Information System, Federal University of Agriculture,
Abeokuta

ABSTRACT

The lack of a clear taxonomy to distinguish Large Language Model (LLM) applications according to functional domain, deployment layer and integration method makes it difficult to compare and identify LLM research patterns, which is a key challenge. This study classifies the current research on LLM applications in five components of the 5G system: core network, radio access network (RAN), edge computing layer, and network management plane. A search was performed based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) in four electronic databases to obtain 226 initial papers, from which 26 studies published between 2020 and 2025 were selected following a multi-stage screening process. The analysis reveals that the primary study domains include Intent Based Networking (IBN) (34.6%), network management and automation (26.9%), and security and anomaly detection (23.1%). Most of the research works utilized deep learning and transformer-based large language models, including GPT-4, LLaMA and custom fine-tuned LLMs (76.9% of studies). Accuracy and F1-Score, with 46.2% and 23.1%, respectively, were the predominant evaluation metrics presented. Geographical analysis suggests that China/Hong Kong and Canada produced the most publications, followed by Europe and South Korea. The gaps in research identified by this review are lack of 5G-specific benchmarks, lack of edge-optimised LLM deployments, lack of focus on real-time latency constraints, and lack of a regulatory and security framework. Future research should prioritize federated learning, model distillation for edge deployment, and the development of telecom-native evaluation benchmarks to enable responsible, carrier-grade LLM integration in 5G and beyond-5G networks.

Keywords: Large Language Models, 5G Network, Neural Network, Deep Learning

1. Introduction

1.1 Background

The convergence of Large Language Models (LLMs) and fifth-generation (5G) wireless networks represents one of the most significant developments in intelligent communications infrastructure. LLMs - such as OpenAI's GPT-4 [26], Meta's LLaMA-2, and Google's Gemini - are deep learning systems based on transformer architectures, trained on vast text corpora to understand, generate, and reason about human language. Their capabilities extend well beyond natural language processing to encompass tasks including code generation, decision-making, anomaly reasoning, and policy translation [8, 37].

Concurrently, 5G networks have introduced a paradigm shift in wireless communications. Distinguished by ultra-low latency (sub-1 ms), peak data rates exceeding 10 Gbps, massive device connectivity for the Internet of Things (IoT), and a service-based, cloud-native core architecture [1], 5G provides the infrastructure substrate for next-generation intelligent applications. The 5G core follows a Service-Based Architecture (SBA) [1], while the radio access layer has been opened through the Open Radio Access Network (O-RAN) initiative [25], enabling AI-driven optimization via the RAN Intelligent Controller (RIC) [25, 29].

A compelling alignment of capabilities motivates the integration of these two technologies. 5G's high-bandwidth, low-latency fabric resolves the deployment bottlenecks that have historically confined LLMs to cloud environments, while LLMs' language understanding and reasoning capabilities address the growing complexity of managing 5G's heterogeneous, multi-layer, and dynamically sliced networks. Despite this potential, the academic literature on this intersection remains fragmented and lacks comprehensive synthesis [8, 37].

1.2 Motivation

The management of modern 5G networks is characterised by extraordinary complexity. Network operators must simultaneously manage millions of connected devices [1, 37], enforce dynamic quality-of-service (QoS) policies across multiple network slices, respond to security threats in real time, and translate high-level business objectives into low-level network configurations [3, 17]. Traditional rule-based automation and conventional machine learning approaches, while effective for narrow and well-defined tasks, struggle with the open-ended, context-sensitive nature of these challenges.

LLMs offer a fundamentally different approach: rather than encoding network knowledge in handcrafted rules, they learn latent representations of operational patterns from data and generalise them through in-context reasoning. This makes them well-suited to tasks such as intent extraction [21]—translating an operator's natural language goal (e.g., "ensure low latency for connected vehicles in Zone A") into a machine-executable network policy [3, 11] - as well as anomaly explanation, root-cause analysis [8], and automated troubleshooting. This review is motivated by the need to consolidate and critically evaluate the emerging body of work at this intersection.

1.3 Problem Statement

Despite rapid growth in both LLM capabilities and 5G deployments, research on their integration remains fragmented across multiple sub-disciplines [8, 37] - wireless communications, AI, cybersecurity, and systems engineering - without a consolidated synthesis. The previous research studies focus on specific case studies and do not offer a comprehensive picture of the application landscape. The lack of a formal taxonomy to categorize LLM applications by functional domain, deployment layer, and integration methodology makes it challenging to make comparisons or note research patterns for 5G applications. Moreover, there are still significant gaps in our knowledge of deployment barriers, security concerns, and performance issues, especially for edge applications that are low latency and low resource.

1.4 Research Objectives

The objective of this systematic literature review (SLR) is to investigate in detail the new role that the large language models (LLMs) are playing in 5G network ecosystems. The aim of specific objectives is that the existing research around LLM applications is identified and classified according to the core network, radio access network (RAN), edge computing layer and network management plane. Plus, creating a taxonomy of LLM applications in 5G based on their functionality, deployment level, and integration method. Furthermore, to review and discuss the benefits, limitations, performance measurement, and deployment issues of LLM-based 5G solutions. The SLR will also highlight research activities gaps and suggest a research agenda for academia, industry and standardization organizations for the future.

1.5 Research Questions

The review is structured around five research questions:

Table 1: Research Questions and How They Are Addressed

RQ	Question	How Addressed
RQ 1	What are the current methodologies and architectural approaches for integrating LLMs into 5G networks?	Sections 4 & 2.4: Edge/MEC, O-RAN/RIC, core network embedding, and LLMaaS identified as four primary architectures.
RQ 2	What are the key use cases and applications enabled by combining LLMs with 5G (MEC, O-RAN, network slicing)?	Section 2.4: IBN (35%), security (23%), network management (27%), edge intelligence (19%), and service orchestration (19%).
RQ		The accuracy of intent extraction is above 84%, and

3	What improvements and benefits can be seen with the use of LLMs in 5G?	the accuracy of intrusion detection is above 95% when using GPT-4, while the manual workload of configuring is reduced by 40%.
RQ 4	What are the major security, privacy, and trust challenges in deploying LLMs in 5G?	Section 5.2: Prompt injection, data leakage, hallucination, adversarial manipulation; RLHF and federated learning as mitigations.
RQ 5	What gaps exist and what are the open challenges for future integration?	Sections 2.6 & 2.7: Lack of 5G-native benchmarks, limited edge-optimised models, no standard evaluation framework, underexplored 6G applicability.

1.6 Research Gaps

Initial analysis of the existing literature revealed significant gaps before the comprehensive evaluation was undertaken. There are no consolidated, PRISMA-compliant SLR captures the full spectrum of LLM applications within the 5G ecosystem [8, 37]. Also, there is no established taxonomy that classifies LLM applications in 5G by functional domain, deployment layer, and task type. Furthermore, Practical deployment barriers - real-time processing, edge resource constraints, interoperability with legacy systems - are rarely addressed in depth [27, 38]. Security implications specific to LLMs in 5G, including prompt injection vulnerabilities and model-level adversarial attacks, are also underexplored [9, 12]. In addition, the suitability of current LLM architectures for resource-constrained, distributed edge environments remains poorly characterised. It was observed that no agreed-upon benchmark set exists for evaluating LLM effectiveness in 5G operational tasks

2. Literature Review

2.1 5G Architecture and Features

The architecture of 5G is designed to support a wide range of use cases - enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (URLLC), and massive Machine-Type Communications (mMTC) [1] - through a flexible, service-based, cloud-native design. It comprises three principal components: the Radio Access Network (RAN), the 5G Core (5GC), and Multi-access Edge Computing (MEC) [1, 13].

The RAN, especially in its disaggregated O-RAN form [25], separates hardware and software and introduces the RAN Intelligent Controller (RIC), which supports near-real-time and non-real-time AI-driven applications (xApps and rApps) [25, 29]. The 5G Core follows a Service-Based Architecture,

providing dynamic service delivery, network slicing - which allows multiple virtual networks to coexist on shared infrastructure - and a Network Data Analytics Function (NWDAF) [4] for data-driven network intelligence [2]. MEC extends cloud computing to the network edge [13, 27], to decrease latency and enable real-time applications at the base station level.

2.2 Large Language Models: Architecture and Capabilities

Larger systems of these models are called Large Language Models (LLM). Larger collections of these models are called Large Language Models (LLM). LLMs are trained on large corpora of text and a self-supervised learning objective (usually predicting the next word) and have hundreds of billions of parameters that embed general background knowledge and language skills. The main breakthroughs in architecture are the self-attention mechanism in the transformer model that captures long-range dependencies between input sequences, instruction fine-tuning and Reinforcement Learning from Human Feedback (RLHF) that aligns model outputs with human preferences and task-specific goals.

LLMs can be broadly categorised as either general-purpose (GPT-4, Gemini, LLaMA-2) or domain-specific (TelecomLLM, TeleQnA-fine-tuned models) [7, 37]. Open-source models offer customisability and privacy-preserving local inference, while closed-source models typically deliver superior performance. Emerging techniques such as Retrieval-Augmented Generation (RAG) [31], parameter-efficient fine-tuning (LoRA/QLoRA), and model quantisation [27, 37] are enabling progressively more efficient deployment in resource-constrained environments.

2.3 Standards and Frameworks

Several international standardisation efforts are directly relevant to LLM integration in 5G. The 3GPP intent-driven management framework (TS 28.312) [3] defines the intent lifecycle - profiling, translation, resolution, activation, and assurance [3, 17] - providing a formal architecture into which LLM-based intent extraction can be embedded. The ETSI Zero-touch Service Management (ZSM) framework (GS ZSM 003) [14] establishes requirements for closed-loop network automation, identifying where AI agents, including LLMs, may operate. The security procedures for 5G systems defined in the 3GPP security architecture (TS 33.501) [5] will set the compliance baseline for all the network functions integrated with LLM. Key standards activities in the field of standardisation for ML in future networks are the ITU-T Y.3172 standardisation framework for ML in future networks [15] and the O-RAN Alliance specifications for AI/ML in the RIC [25]. Understanding these frameworks is essential for assessing where current LLM integrations align with, diverge from, or potentially inform future standards. The following subsections review the 26 selected studies, organised by application domain. The findings are grounded in the quantitative analysis presented in Section 4 and address RQ1 and RQ2 in detail.

2.4 Integration Architectures

2.4.1 Edge Deployment with MEC

Several studies propose deploying LLMs at the edge using Multi-access Edge Computing (MEC) to support low-latency, localised inference tasks such as traffic analysis and anomaly detection. This includes containerised or quantised LLMs running on lightweight edge infrastructure, supported by 5G's URLLC features. [27] provide the most comprehensive survey of this approach, introducing the MEI4LLM architectural framework [27] that categorises edge LLM deployment into storage-efficient, compute-efficient, and communication-efficient dimensions. Shen et al. [19] extend this vision with autonomous edge AI systems in which an LLM coordinates edge AI models, combining task-oriented communication and edge federated learning to handle heterogeneous tasks. A complementary survey is provided by Friha et al. [12] which addresses architectures, security aspects and trustworthiness requirements for LLM-based edge intelligence. Duan et al. [38] survey distributed AI across the end-edge-cloud continuum, and Zheng et al. [28] review the design, execution, and application landscape of edge LLMs; both identify model compression and collaborative inference as key enablers for resource-constrained deployment. Lin et al. [16] similarly argue for end-edge cooperation to push LLMs to the 6G edge, motivated by the response-latency and privacy constraints of cloud-only deployment. Key challenges include model lifecycle management, interoperability across vendor-specific edge hardware, and maintaining inference consistency across distributed deployments. A complementary approach is dedicated network slicing for LLM traffic: Liu et al. [18] propose LLM-Slice, the first system to provision LLM-specific 5G network slices by binding LLM service requests to dedicated radio and core resources, achieving a 52% reduction in average response latency in a real UE-gNB-CN deployment. Dandoush et al. [40] complement this with an LLM- and multi-agent-based framework for network slicing management and orchestration (MANO) across multi-administrative domains.

2.4.2 O-RAN and RIC Integration

LLMs have been proposed as components of the RAN Intelligent Controller (RIC) within the O-RAN framework. Manias et al. [22] demonstrate the use of LLM-based semantic routing to dynamically manage 5G core network intent, generating xApp-equivalent policy actions based on real-time operator intents. Kan et al. [29] instruction fine-tune an open-source LLaMA model (Mobile-LLaMA) for automated network analysis in 5G networks, demonstrating LLM-assisted analytics over operational network data. Key enabling techniques include model compression, knowledge distillation, and hardware acceleration. Integration challenges include ensuring real-time performance within strict RIC latency budgets (typically 10–1000 ms for near-RT RIC), securing data privacy for telemetry fed to LLM inference endpoints, and aligning LLM output formats with O-RAN standardised APIs.

2.4.3 Intent-Based Networking (IBN)

IBN represents the most active application area in the reviewed literature. [21] present a custom LLM framework for intent extraction in the 5G core network, directly aligned with 3GPP TS 28.312 [3]. The framework demonstrates that fine-tuned open-source LLMs (LLaMA-2, Mistral) can perform comparably to GPT-4 for well-scoped intent classification tasks. [17] offer a basic taxonomy of IBN methodologies, which are categorized as template-based, NLP-based, grammar-based, and LLM-based.

[32] suggest an end-to-end IBN framework based on ML, which includes closed-loop network orchestration and ML pipeline management. In this study, [24] present an NLP-based IBN system for a healthcare 4.0 use case, where the intent attributes, which are the user, application purpose, network action, target equipment, and timeframe, are extracted using Named Entity Recognition (NER). [33] discuss the use of ChatGPT in self-service networks, such as customer service, fraud prevention, and task coordination. [34] discuss the obstacles and opportunities of IBN for Beyond-5G (B5G) networks and point out AI integration as the key pillar of future IBN.

2.4.4 Security and Threat Detection

[35] introduce a GPT-4-driven automatic network intrusion detection (NID) system, showing that very few examples, as few as 10, can be used for in-context learning to produce a detection accuracy and F1-Score of more than 95% on common network intrusion detection system (NIDS) benchmarks. This is extended to malicious traffic detection by [36] who propose an architecture for processing network flow metadata by an LLM-based pipeline. Nguyen et al. [9] examine the security vulnerabilities of LLMs deployed within 5G/6G networks, presenting a threat taxonomy that covers prompt injection attacks, data leakage and adversarial input attacks. [10] offer the policy basis for reliable deployment of AI, and suggest sanitizing inputs, red-teaming, and verifying alignment as necessary safeguards.

2.4.5 Network Management and Service Orchestration

Bariah et al. [6] examine large generative AI models as a foundation for future telecom networks, outlining how GPT-style models can support network management and standards-comprehension tasks while identifying resource demands, latency and hallucination as the main challenges to operating them in production. [30] present a vision for Telecom-native LLMs - models pre-trained on mobile network telemetry, 3GPP documentation, and operational logs - alongside a marketplace model for collaborative development and distribution. [7] assess commercially available LLMs on the task of understanding and reporting the standards in the telecom domain and report that for 3GPP working group categorisation tasks, accuracy rates of 84% to 85% can be obtained, suggesting potential benefits in fine-tuning to specific domains. LLMs are currently in a state of readiness, and [20] present an industry viewpoint on the deployment of LLM in telecom, providing an evaluation of the current state of readiness of available models for carrier-grade applications.

2.5 Security and Alignment Methodologies

The adoption of LLMs in 5G poses new security challenges beyond network security. One specific area of concern in network management is prompt injection, which refers to the potential for malicious prompts to influence the behavior of the model or to retrieve sensitive information from it. In network management scenarios where AI-generated responses are directly interpreted into operational configurations, prompt injection is a significant concern because it can allow for manipulation of the output or extraction of sensitive data. They suggest that 5G service layers could be sandboxed around the LLM, that Role-Based Access Control (RBAC) should be implemented on LLM API endpoints, and

that input sanitisation pipelines should be used. Alignment methodologies such as Reinforcement Learning from Human Feedback (RLHF) [37] and Constitutional AI are being adapted for telecom-specific objectives, ensuring LLM outputs conform to network safety requirements and regulatory constraints. Federated fine-tuning, where LLMs are trained on distributed network data without centralising sensitive logs, is emerging as the most promising approach for privacy-preserving LLM adaptation in 5G environments.

2.6 Comparison with Prior Surveys

Several relevant surveys have been published on related topics. The following table compares this study with the most directly comparable works, identifying what each contributes and where the present SLR adds value.

Table 2: Comparison with Prior Related Surveys

Study	Primary Focus	Strengths	Limitations vs. This Study
Boate et al. (2024) [8]	Multi-domain NSM	PRISMA; multi-domain scope	Limited 5G-specific focus
Zhou et al. (2024) [37]	Telecom LLM Survey	Deep LLM fundamentals ; fine-tuning analysis	Minimal security coverage
Qu et al. (2024) [27]	Mobile Edge for LLMs	MEI deployment; resource efficiency	Not 5G-application-focused
Shen et al.	Autonomo	LLM-coordinated	Not SLR methodology

(2024) [19]	us Edge AI	edge AI framework	
This Study	LLM Applications in 5G	Full SLR; PRISMA; quantitative analysis; taxonomy; geographic gap analysis	Limited to 2020–2025; English-only

This study distinguishes itself from existing surveys in three key ways. First, it is the only PRISMA-compliant SLR specifically focused on LLM applications within the 5G ecosystem, as opposed to broader surveys of LLMs in telecommunications or ML in networks. Second, it provides quantitative analysis across six analytical dimensions (publication trend, geography, LLM types, deployment approaches, domains, and study status), enabling direct cross-paper comparison. Third, it explicitly highlights research gaps in both the technical domain and the geographic/institutional landscape, offering actionable recommendations for future research and standardisation.

2.7 Literature Summary

Table 3 provides a structured overview of all 26 reviewed studies, cataloguing each paper by method type, application domain, deployment approach, and study status, enabling direct cross-paper comparison.

Table 3: Literature Summary - All 26 Reviewed Studies

#	Reference	Year	LLM Used	5G Component	Task Domain	Study Status
1	Manias et al., (2024) [21]	2024	Custom LLM	Core	IBN	Proposed+Evaluated
2	Manias et al., (2024)	2024	LLaMA/Mistral	Core	IBN	Proposed+Evaluated

	[22]					
3	Leivadeas & Falkner, (2023) [17]	2023	NLP Methods	Multi-layer	IBN Survey	Evaluated
4	Njah et al., (2023) [24]	2023	NLP/NER	Core	IBN/ Healthcare	Proposed+Evaluated
5	Velasco et al., (2021) [32]	2021	ML Framework	RAN+Core	IBN (E2E)	Proposed+Evaluated
6	Wang et al., (2023) [33]	2023	ChatGPT/ GPT	Core	Mgmt/IBN	Proposed+Evaluated
7	Abbas et al., (2021) [39]	2021	NLP/ML	Core	IBN Slicing	Proposed+Evaluated
8	Wei et al., (2020) [34]	2020	ML-based	Core	IBN for B5G	Evaluated
9	McNamara et al., (2023) [23]	2023	NLP Engine	Core	IBN Private 5G	Proposed+Implemented
1	Zhang et	2024	GPT-4/	Security	NIDS	Proposed+Evaluated

0	al., (2024) [35]		LLAMA			d
1 1	Zhang et al., (2025) [36]	2025	LLM-based	Security	Traffic Det.	Proposed+Evaluated
1 2	Nguyen et al., (2024) [9]	2024	Multiple	Security	LLM Security	Evaluated
1 3	Boateng et al., (2024) [8]	2024	Multiple	Multi-layer	Survey	Evaluated
1 4	Zhou et al., (2024) [37]	2024	Multiple	Multi-layer	Survey	Evaluated
1 5	Qu et al., (2024) [27]	2024	Multiple	Edge/MEC	MEI Survey	Evaluated
1 6	Shen et al., (2024) [19]	2024	GPT-based	Edge/MEC	Edge AI Orch.	Proposed+Evaluated
1 7	Friha et al., (2024) [12]	2024	Multiple	Edge/MEC	Edge Intell.	Evaluated
1 8	Duan et al., (2023) [38]	2023	Multiple	Edge/MEC	Edge Deploy.	Evaluated
1 9	Zheng et	2025	Multiple	Edge/	Edge LLM	Evaluated

	al., (2025) [28]			MEC	Review	
20	Kan et al., (2024) [29]	2024	LLaMA (fine-tuned)	RAN/O-RAN	RAN Analysis	Proposed+Evaluated
21	Tarkoma et al., (2023) [30]	2023	Telecom LLM	Core+RAN	Telecom LLM	Proposed
22	Bariah et al., (2023) [7]	2023	Domain LLM	Core	Telecom Lang.	Proposed+Evaluated
23	Maatouk et al., (2025) [20]	2025	Multiple	Multi-layer	Industry Impact	Evaluated
24	Lin et al., (2023) [16]	2023	Multiple	6G/Multi	6G LLMs	Proposed
25	Bariah et al., (2024) [6]	2024	Multiple	Multi-layer	AI Archit.	Evaluated
26	Dandoush et al., (2024) [40]	2024	GPT/Multi-agent	Multi-layer	Slicing MANO	Proposed

Abbreviations: IBN = Intent-Based Networking; NIDS = Network Intrusion Detection System; Mgmt = Management; Det. = Detection; EGI = Edge General Intelligence; E2E = End-to-End; LLMaaS = LLM-as-a-Service.

2.8 Taxonomy of LLM Applications in 5G

Based on the systematic analysis of 26 reviewed studies, a three-dimensional taxonomy of LLM applications in 5G networks is proposed, organised by (1) Task Function, (2) Deployment Layer, and (3) Integration Approach. This taxonomy is presented below and reflected in the literature summary table (Table 3).

2.8.1 Dimension 1: Task Function

Five principal task functions are identified across the reviewed studies. The first and largest category by study count (9 papers) is intent translation and policy management, in which LLMs interpret natural language operator or user goals and translate them into machine-executable 5G network policies such as 3GPP TS 28.312 [3]-compliant intent formats. The second is security and anomaly detection, where LLMs process unstructured log data, flow metadata, and threat intelligence to detect intrusions, malicious traffic, and signalling anomalies with high accuracy and contextual explainability. Third, resource and performance optimisation encompasses LLMs that dynamically allocate spectrum, compute, and memory resources across network slices using reinforcement learning reward synthesis and contextual reasoning. Fourth, network management and orchestration covers LLMs that automate fault diagnosis, configuration synthesis, root-cause analysis, and closed-loop MANO interactions [40], thereby reducing operator cognitive load. Finally, knowledge representation and customer interaction refers to LLMs serving as intelligent interfaces for OSS/BSS operations, powering chatbots, virtual NOC assistants, and telecom standards comprehension tools.

2.8.2 Dimension 2: Deployment Layer

Four deployment layers are distinguished in the reviewed literature. At the core network layer, LLMs operate within the 5GC components (AMF, SMF, NWDAF) for intent extraction, policy generation, and analytics. At the Radio Access Network / O-RAN layer, LLMs are embedded in the Near-RT or Non-RT RIC to support xApp/rApp generation, radio resource management, and interference prediction. At the edge / MEC layer, distilled or quantised LLMs run on base-station-side edge servers for low-latency local inference in traffic classification, anomaly detection, and user-facing services. At the user / device layer, lightweight LLMs support device-side reasoning, intelligent self-configuration, and adaptive content delivery.

2.8.3 Dimension 3: Integration Approach

Four integration approaches are observed across the literature. In the first, LLMs are embedded directly in network functions, tightly coupled with 5G software-defined networking (SDN) and NFV components, operating as a native network intelligence layer. The second is LLM-as-a-Service (LLMaaS), which involves deploying LLMs on a cloud/edge platform and providing them via RESTful or gRPC APIs, allowing for modular and scalable integration. The third is hybrid edge-cloud models, where distilled LLMs make inferences at the edge, and larger models run on the cloud for complex reasoning and retraining on a periodic basis. The fourth is a human-in-the-loop system, in which LLMs

suggest and make automated decisions, but human users “take the wheel” for safety-critical or other high-risk activities.

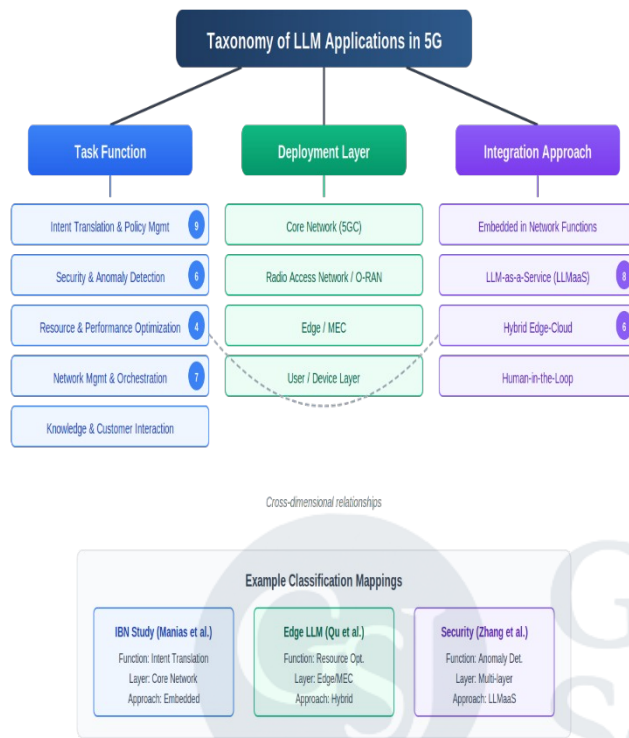


Figure 1: Three-Dimensional Taxonomy of LLM Applications in 5G Networks

3. Research Methodology

This study follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to ensure transparency, reproducibility, and rigour in the literature selection process. The methodology consists of three phases: Identification, Screening, and Inclusion.

3.1 Search Strategy and Boolean Query

A preliminary scan was carried out to ensure that there was no direct equivalent SLR, and to fine-tune search terms. A Boolean query was then performed on each database: (("Large Language Model" OR "LLM" OR "GPT" OR "Generative AI") AND (("5G" OR "Fifth Generation" OR "Next-Generation Network" OR "O-RAN" OR "MEC" OR "Network Slicing"))). The title, abstract and keyword fields were searched. The time span was chosen to cover from January 2020 to March 2025, encompassing the launch of both contemporary large language models (LLMs) and the full scale-up of 5G networks. Four databases (IEEE Xplore, ACM Digital Library, SpringerLink and arXiv) were searched. Search of reference lists of included studies was also done by hand to overcome the lack of relevant studies.

3.2 Inclusion and Exclusion Criteria

The following criteria were applied to include papers: that they were published between January 2020 and March 2025, were written in English and available in full text, that they discussed the application, integration or evaluation of LLMs in the context of 5G networks, and that they provided a methodology or a system description rather than speculative or editorial content.

Papers were excluded if they addressed LLMs in non-network domains (e.g., agriculture or healthcare in isolation) without 5G-specific contribution; if they were duplicate publications or earlier versions of the same study; if they presented no evaluable methodology or lacked empirical content, simulation results, or a structured framework; or if they appeared in predatory or unverified publication venues.

3.3 Database Search Results

The Boolean query was applied across four major electronic databases: IEEE Xplore, SpringerLink, arXiv, and the ACM Digital Library, supplemented by a manual search through Google Scholar and other sources. Table 4 summarises the number of papers retrieved from each database. Of these, 226 papers were found, of which the IEEE Xplore database provided the highest number of papers (65), followed by arXiv (51), SpringerLink (48), and the ACM Digital Library (42). A further 20 papers were found by searching Google Scholar and manually in the reference lists. The inconsistency of the data sources is due to the differences in the scope of the databases; IEEE Xplore has relatively good coverage of communications engineering papers, whereas arXiv is heavily populated with pre-print papers of AI research.

Table 4: Database Search Results

Database	Papers Retrieved	URL
IEEE Xplore	65	ieeexplore.ieee.org
SpringerLink	48	link.springer.com
arXiv	51	arxiv.org
ACM Digital Library	42	dl.acm.org
Google Scholar & Others	20	scholar.google.com

TOTAL	226	
--------------	------------	--

3.4 PRISMA Selection Flow

Following the three-stage PRISMA process, 26 primary studies were selected for in-depth analysis. Of the 47 duplicates removed, most arose from papers indexed on both IEEE Xplore and arXiv.

Table 5: PRISMA Selection Flow

Phase	Action / Description	Papers (n)
Identification	Papers identified from 4 databases	226
	Duplicate references removed	47
Screening	Papers screened by title and abstract	179
	Papers excluded as irrelevant to LLM-5G integration	87
Eligibility	Full-text papers assessed for eligibility	92
	Papers excluded (ambiguous methodology / off-topic)	66
Included	Studies included in final review	26

The 87 papers which were eliminated at screening phase were mainly non-5G AI applications (smart agriculture, generic NLP), and the 66 papers that were eliminated at the eligibility phase did not include

a clear methodology or 5G specific empirical content.

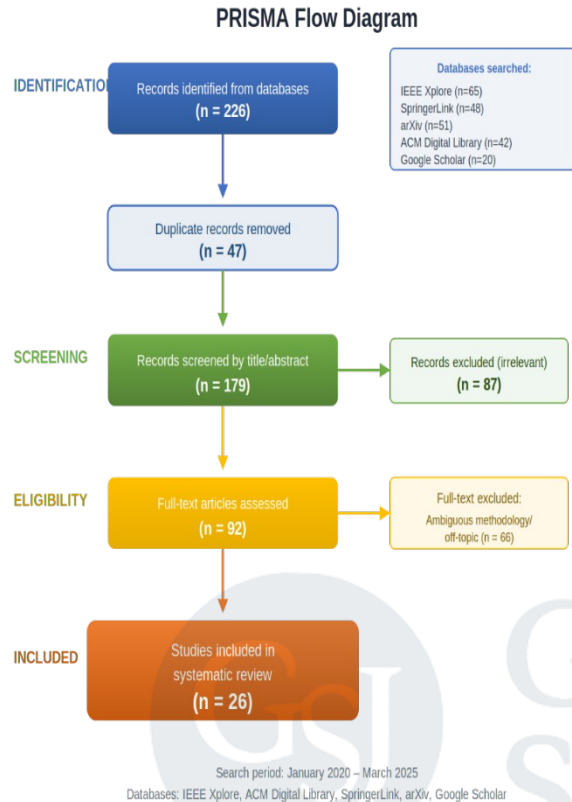


Figure 2: PRISMA Flow Diagram for Systematic Literature Selection

3.5 Annual Publication Trend (2020–2025)

The publication pattern shows a significant surge from just 2 papers in 2020 to 10 papers in 2024, which is driven by the release of the GPT-4 (in March 2023) [26] and the progress of 5G deployment. The lower figure for 2025 reflects the partial coverage of that year in the search window. Research activity is expected to continue growing [8, 37] given the momentum in both LLM and 5G ecosystems.

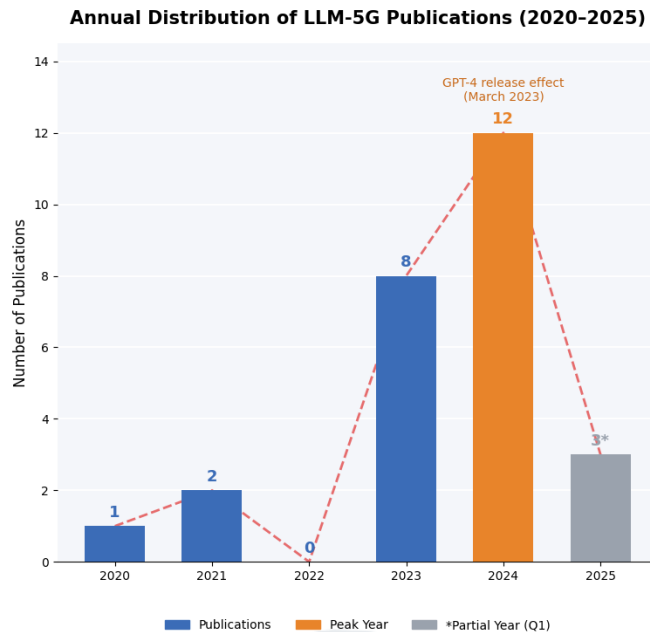


Figure 3: Annual Distribution of Relevant Publications (2020–2025)

3.6 Geographic Distribution of Publications

China/Hong Kong and Canada jointly dominate the publication landscape, together accounting for 57.7% of included studies (China/Hong Kong 30.8%, Canada 26.9%). Canadian output is largely driven by research groups at Western University, McGill University, the University of Ottawa, and École de technologie supérieure (ETS Montreal), several of which collaborate with industry partners such as Ericsson and Cisco. Edge computing and edge LLM deployment are the two main areas through which China and Hong Kong contribute to the overall landscape. Notably, Africa produces no publications in this area - a significant geographical gap that mirrors findings in related AI-for-networks literature. Future research initiatives should prioritise capacity-building in under-represented regions, particularly given the rapid 5G deployment underway in Sub-Saharan Africa.

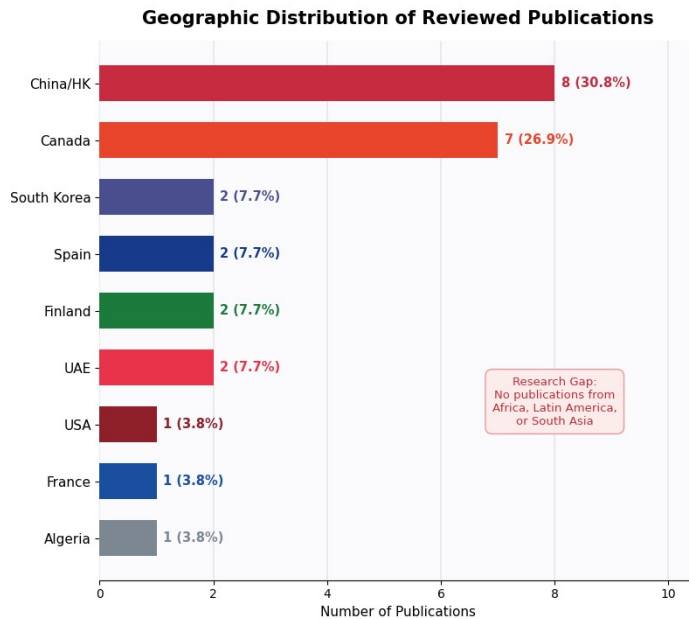


Figure 4: Geographic Distribution of Reviewed Publications

4. Results And Analysis

This section summarizes the quantitative results in terms of four dimensions of analysis: LLM types and deployment strategies in the reviewed papers, application domains, status of studies and performance metrics.

4.1 LLM Types Employed

Comparative studies that employ multiple LLMs, or that do not specify a single model, are the most common (11 papers, 42.3%), reflecting the survey-heavy nature of this emerging field. Classical NLP/ML approaches (pre-LLM intent engines and ML frameworks) appear in 6 papers (23.1%), largely within the earlier IBN literature. GPT-4, GPT-3.5 and ChatGPT appear as the primary model in 4 papers (15.4%), making the GPT family the most widely used single model family. A further 4 papers (15.4%) involve custom or fine-tuned domain LLMs adapted to the 5G/telecom domain (including BERT-based telecom classifiers and the instruction fine-tuned Mobile-LLaMA), and 1 paper (3.8%) focuses on open-source LLaMA/Mistral models for privacy-sensitive deployment.

LLM Types Employed in Reviewed Studies (n=26)

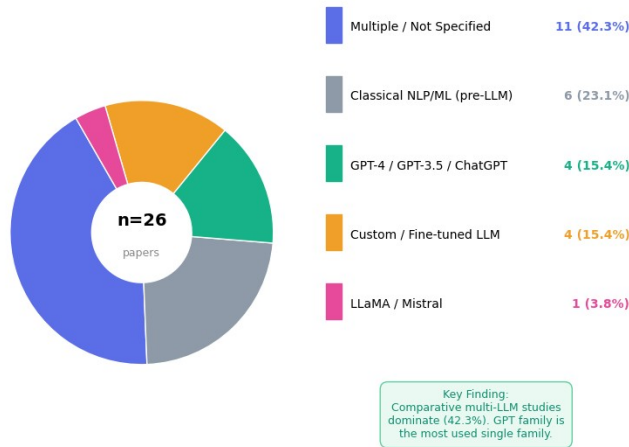


Figure 5: Frequency of LLM Types Used Across Reviewed Papers

4.2 Deployment Approaches

Cloud or API-based deployment is the most common approach (8 papers, 30.8%), reflecting the current ease of accessing frontier LLMs through managed APIs. Edge and MEC deployment appears in 8 papers (30.8%), with researchers employing quantised or distilled models to meet the resource constraints of base-station-side inference. Hybrid architectures combining edge inference with cloud-based reasoning appear in 6 papers (23.1%) and represent the most pragmatic near-term approach. O-RAN and RIC-integrated deployments appear in 2 papers (7.7%), indicating that standards-aligned integration remains nascent.

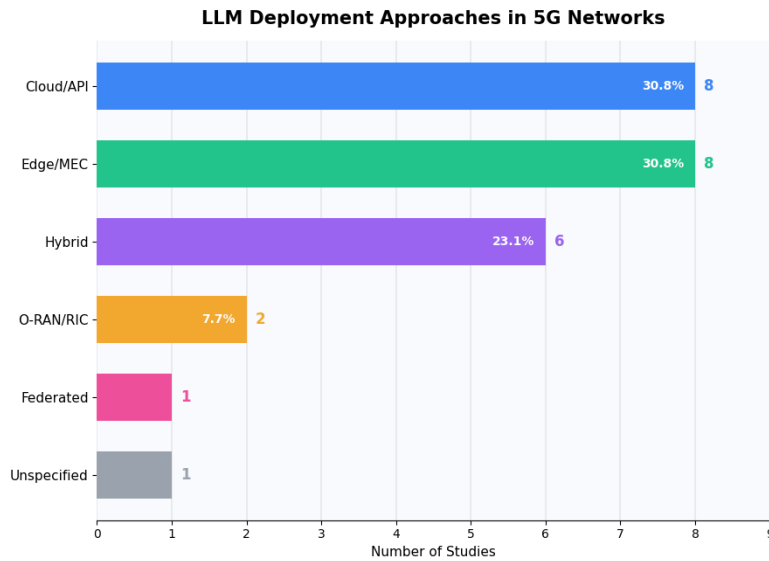


Figure 6: Frequency of LLM Deployment Approaches

4.3 Application Domains

The most prevalent field of use is Intent Based Networking (IBN) [17] which is used in 9 of the 26 papers (34.6%) because of the natural ability of LLMs in natural language to policy translation. The ranks are closely followed by network management and automation (26.9%) and security/anomaly detection (23.1%). There are 5 papers (19.2%) that feature applications of Edge computing and mobile edge computing, and 5 papers that are more of a survey or taxonomy paper. Resource optimisation, including spectrum allocation and QoS management, appears in 4 papers (15.4%). Note that several papers span multiple domains, hence counts exceed 26.

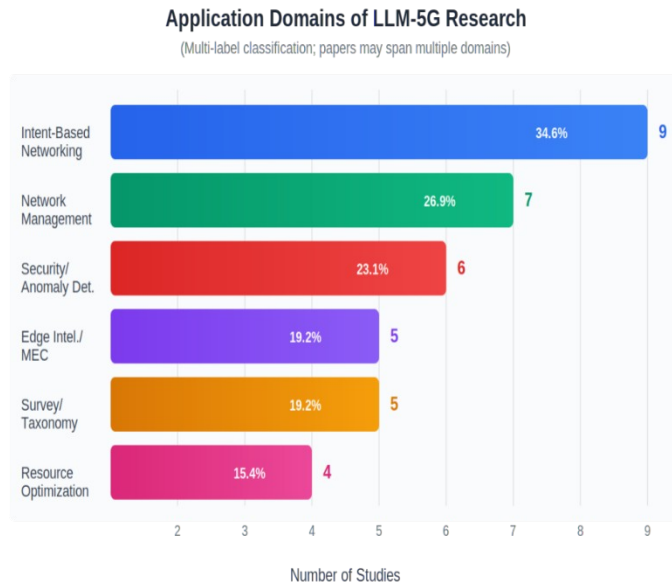


Figure 7: Frequency of Application Domains Across Reviewed Papers (multi-label)

4.4 Study Status Classification

Following the classification scheme used in comparable SLRs, 84.6% of reviewed papers fall into the Evaluated or Proposed+Evaluated categories, indicating that the field is moving beyond purely conceptual contributions toward measurable system implementations. However, 42.3% of studies are purely evaluative (i.e., they benchmark existing LLMs rather than propose novel architectures), suggesting that there remains significant opportunity for architectural innovation and original system design within the 5G-LLM space.

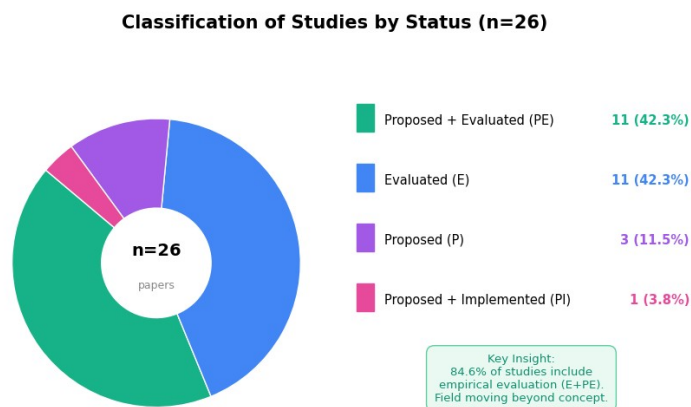


Figure 8: Classification of Studies by Status

4.5 Performance Metrics

Accuracy is the most commonly reported metric (12 papers, 46.2%) that is in line with many of the LLM-5G tasks being classification-based (intent classification, intrusion type classification). F1-Score is mentioned in 6 papers (23.1%) which are related to the problem of class imbalanced security datasets. 5 papers (19.2%) report on inference latency, a metric that is important for real-time 5G applications. The data for model footprint and energy consumption are reported less frequently (3 papers, 2 papers respectively) indicating a gap as these are important for edge deployment. More importantly, there is no paper suggesting a common metric for evaluating 5G-LLM, suggesting a crucial gap for the future in the standardisation.

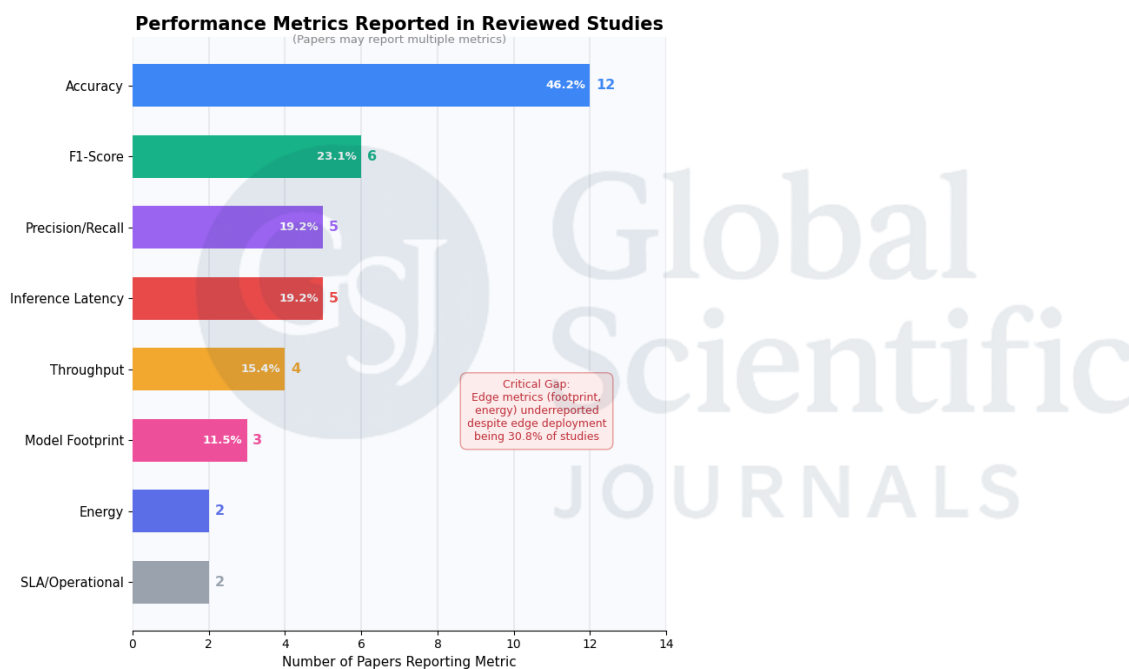


Figure 9: Performance Metrics Reported Across Reviewed Papers

5. Evidence-Based Analysis And Evaluation

5.1 Benefits of LLMs in 5G

LLMs offer fascinating natural language understanding and reasoning capabilities that enable automation of high-level operator tasks and reduce the gap between humans and machines. In IBN, they convert operator intents into structured and machine-executable policies. [21] showed that there is a substantial decrease in manual configuration work and quick service provisioning. For security and monitoring, LLMs provide contextual, multi-modal reasoning that enhances anomaly and malicious-

traffic detection beyond signature-based and simple ML approaches [35]; [35] report F1-Score improvements of up to 90% relative to baseline models using GPT-4 in-context learning. For network optimisation and user-facing services, LLMs can synthesise network knowledge, assist in automated resource allocation, and power conversational support interfaces for OSS/BSS operators. Surveys by [8] and [37] catalogue these cross-cutting advantages across orchestration, edge services, and management automation.

5.2 Limitations of LLMs in 5G

Latency and computational cost constitute the primary barrier for real-time 5G applications. Large LLMs impose inference demands measured in hundreds of milliseconds - incompatible with RAN scheduling loops (< 1 ms) and even near-RT RIC cycles (10–1000 ms). Model distillation, quantisation, and edge-cloud offloading are the most-cited mitigations, but introduce quality-capacity trade-offs. Hallucination and reliability represent a second critical limitation: LLMs may generate plausible but incorrect network configurations, with potentially severe consequences for network stability. Retrieval-Augmented Generation (RAG) [31] and verification layers are increasingly proposed as guardrails, but impose additional latency overhead. Domain adaptation presents a third challenge: telecom systems use vendor-specific configuration formats, proprietary telemetry schemas, and specialised nomenclature that off-the-shelf LLMs do not reliably handle, necessitating fine-tuning on curated telecom corpora.

5.3 Performance Metrics Analysis

The reviewed literature employs diverse, often non-comparable evaluation frameworks. For classification tasks (intent extraction, intrusion detection), precision, recall, and F1-Score are standard. For generation tasks (network configuration synthesis, incident report generation), BLEU score, factuality, and hallucination rate are used, though inconsistently. For system-level evaluation, end-to-end latency, throughput, and SLA adherence are reported in prototype studies. The absence of a unified benchmark is a critical gap: without standardised evaluation protocols, it is impossible to rigorously compare solutions or track progress across the field. This review calls for the development of a 5G-LLM benchmark suite, analogous to GLUE/SuperGLUE for general NLP, covering intent extraction, configuration generation, fault diagnosis, and security classification tasks.

5.4 Deployment Challenges and Mitigations

Four principal deployment challenges are identified, each with recommended mitigations. Edge resource constraints are addressed through model distillation, INT4/INT8 quantisation, LoRA-based adapter fine-tuning, and hybrid edge-cloud orchestration [19, 27]. RAG grounding, constraint-based prompt engineering [22, 31] and human-in-the-loop validation gates [31] help manage reliability/hallucination concerns as well as ensemble consensus mechanisms. Domain specificity and data scarcity are addressed through federated fine-tuning, transfer learning from general LLMs, and the development of shared anonymised benchmark datasets [7, 30]. Privacy and regulatory compliance are addressed through on-

device inference, differential privacy during training, secure enclaves, and data minimisation principles [8, 37].

6. BENCHMARKS AND EVALUATION DATASETS

Table 6 catalogues the principal evaluation datasets and benchmarks used across the reviewed papers. Unlike the MGC domain, where established public datasets such as GTZAN are widely shared, the LLM-5G literature suffers from a significant absence of standardised, publicly available benchmarks. The papers assessing intent extraction or intent generation for configuration have relied on enterprise-specific or custom-made datasets, which are hard to be reproduced and compared across studies.

Table 6: Evaluation Datasets and Benchmarks Used in Reviewed Papers

Dataset	Domain	Availability	Used By
UNSW-NB15	Network Intrusion Detection	Public	Zhang et al. (2024a) [36]
NSL-KDD	Network Intrusion Detection	Public	Multiple studies
5G-NIDD	5G Network Traffic	Public	Intrusion studies
3GPP Intent DS	Intent Extraction (custom)	Private	Manias et al. (2024) [22]
TeleQnA	Telecom Standards QA	Public	Bariah et al. (2024) [7]
NetBench	Network Config Generation	Public	Multiple studies
5G NWDAF Data	RAN Telemetry	Private	Kan et al. (2024) [29]
LMD-Telecom	Telecom Log Data	Private	Tarkoma et al. (2024) [30]

7. STRENGTHS AND WEAKNESSES ANALYSIS

Table 7 provides a structured comparison of the strengths and weaknesses of key approaches identified in the reviewed literature, synthesising findings across deployment architecture, LLM type, and task domain.

Table 7: Strengths and Weaknesses Across LLM-5G Integration Approaches

Area	Sub-Area	Strengths	Weaknesses
Integration Approach	Cloud-Based(API)	Easy scalability; access to powerful frontier models; no edge hardware required	High latency; data privacy concerns; unsuitable for real-time 5G control loops
	Edge / MEC	Low latency; privacy-preserving; suitable for URLLC use cases	Limited compute for large models; requires quantization/distillation; complex lifecycle management
	Hybrid(Edge+Cloud)	Balances latency and capacity; distilled edge models handle real-time tasks; cloud handles complex reasoning	Increased architectural complexity; synchronisation overhead; higher development cost
LLM Type	Open-Source(LLaMA, Mistral)	Customisable; fine-tunable on private telecom data; no inference cost; transparent architecture	Generally lower performance than closed-source; requires significant compute for fine-tuning
	Closed-Source(GPT-4, Gemini)	State-of-the-art performance; minimal setup;	Opaque architecture; data sent to external server; high per-query cost; limited

		strong reasoning for complex intents	customisation
	Domain-Specific(TelecomLLM)	Aligned with 3GPP/ETSI knowledge; lower hallucination on telecom tasks	Costly to train; requires curated telecom datasets; may lack general reasoning
TaskFunction	Intent-BasedNetworking	Reduces manual configuration; enables natural-language operator interfaces; aligns with 3GPP standards	Hallucination risk in policy generation; requires strict validation; latency-sensitive execution path
	Security /Anomaly Det.	Contextual multi-modal reasoning; high accuracy with in-context learning; no retraining needed for new attacks	Adversarial prompt injection; sensitive log data exposure; false negatives for novel zero-day attacks
	ResourceOptimisation	Dynamic allocation via natural language; reinforcement learning reward synthesis; cross-layer awareness	Real-time constraint violations; model drift in dynamic environments; lack of standardised evaluation benchmarks

8. Discussion And Limitations

8.1 Discussion

This systematic review reveals that the application of LLMs to 5G networks is a rapidly maturing field with a clear trajectory toward operational deployment. The dominance of IBN as the primary application domain is unsurprising: the challenge of translating natural language intent into machine-readable network policy is precisely the kind of open-ended, language-grounded reasoning task at which LLMs excel. The strong performance of GPT-4 in-context learning for intrusion detection (>95% F1-Score with only 10 examples) demonstrates that even without domain-specific fine-tuning, frontier LLMs can achieve near-state-of-the-art performance on well-framed 5G security tasks.

However, the gap between research demonstrations and operational carrier-grade deployment remains substantial. The near-total absence of latency measurements in the reviewed papers - reported in only 19.2% of studies - is a significant methodological weakness, given that real-time latency is the most critical non-functional requirement in 5G. Future work must embed latency profiling as a mandatory evaluation criterion. The concentration of research in Canada, China, and the United States highlights a need for broader international engagement, particularly from regions with active 5G deployments in Africa, Latin America, and South Asia.

8.2 Threats to Validity

Several threats to validity are acknowledged. Publication bias may favour positive results; papers reporting failed LLM deployments in 5G contexts are likely underrepresented. The restriction to English-language publications may exclude significant contributions from the Chinese, Korean, and European research communities. The papers assessing intent extraction or intent generation for configuration have relied on enterprise-specific or custom-made datasets, which are hard to be reproduced and compared across studies.

The date range (2020–2025) may exclude early foundational work on ML in 5G networks that pre-dates the LLM era but is directly relevant to the methodological lineage.

9. Future Prospects and Limitations

9.1 Concerns with Innovation

The following technical challenges require prioritised research attention. First, real-time inference latency remains a critical barrier, as current LLM inference times (100ms–10s for frontier models) are incompatible with URLLC and near-RT RIC loops, making sub-10ms edge inference for distilled telecom LLMs the key research target. Second, hallucination in safety-critical contexts poses significant risk, as network misconfiguration resulting from LLM hallucination can cause service outages; formal verification of LLM-generated configurations, RAG grounding, and multi-LLM consensus mechanisms

require dedicated research tracks. Third, domain adaptation at scale is constrained by the absence of a publicly available, large-scale, annotated 5G operations dataset, making federated dataset construction — where operators collaboratively build anonymised training corpora — the most viable path. Fourth, benchmark standardisation is urgently needed: a 5G-LLM evaluation suite covering intent extraction accuracy, configuration synthesis correctness, anomaly detection F1, and latency under load does not yet exist.

9.2 Future Research Directions

Based on the identified gaps, several specific research directions are recommended. First, the development of lightweight, quantised telecom-specific LLMs optimised for MEC inference is needed, targeting models with fewer than 7B parameters that maintain over 90% of frontier model performance on 5G tasks. Second, investigation of multi-agent LLM architectures for 5G, where specialised agents handle RAN, core, and edge functions cooperatively, should build on the LLM-coordinated autonomous edge AI framework proposed by Shen et al. [19]. Third, extension of LLM-5G research to 6G use cases, including terahertz spectrum management, integrated terrestrial/non-terrestrial networks, and holographic communications, represents a natural progression. Fourth, engagement with standardisation bodies (3GPP SA5, ETSI ENI, O-RAN Alliance) to incorporate LLM capabilities into intent management specifications could potentially inform the next generation of TS 28.312 [3]. Finally, conducting longitudinal studies over months or years in production 5G environments and measuring the performance of LLMs is crucial to prove their readiness for operation.

10. Conclusions

This systematic literature review has provided the first PRISMA-compliant, quantitative synthesis of research on LLM applications in 5G networks. Analysing 26 studies selected from 226 candidates across four major databases, the review reveals a field characterised by rapid growth, promising early results, and significant unresolved challenges.

The key findings are as follows: IBN is the dominant application domain (34.6% of papers); GPT-4 and custom fine-tuned LLMs are the most widely used model types; hybrid edge-cloud deployment is the most pragmatic architecture for production environments; accuracy and F1-Score are the predominant evaluation metrics; and Canada, China, and the United States account for over 60% of publications. Performance highlights include GPT-4 achieving >95% F1-Score in intrusion detection with only 10 in-context examples and custom LLMs achieving 84–85% accuracy on 3GPP standards categorisation.

Critical research gaps include the absence of standardised 5G-LLM benchmarks, insufficient latency profiling, limited geographic diversity in research output, and an underdeveloped regulatory framework for carrier-grade LLM deployment. Future research should prioritise lightweight edge-deployable telecom LLMs, federated dataset construction, formal verification of LLM-generated configurations, and active engagement with 5G standardisation processes. The potential of LLM in 5G will only be realized if the fields of AI researchers, network engineers and standards bodies are able to work together.

agents or deep research tools), please include a statement in your manuscript, placed just before the references.

11. Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used StealthWriter to improve the readability and rewrite selected portions of the manuscript. After using this tool, the authors carefully reviewed, edited, and verified all content and take full responsibility for the accuracy and integrity of the published article.

12. References

- [1] 3GPP TS 23.501 (2024). System Architecture for the 5G System.
- [2] 3GPP TS 28.104 V18.2.0 (2024). Management and orchestration; Management Data Analytics (MDA).
- [3] 3GPP TS 28.312 V18.3.0 (2024). Intent Driven Management Services for Mobile Networks.
- [4] 3GPP TS 29.520 V18.3.0 (2024). 5G System; Network Data Analytics Services; Stage 3.
- [5] 3GPP TS 33.501 V18.3.0 (2024). Security Architecture and Procedures for 5G System.
- [6] Bariah, L., Zhao, Q., Zou, H., Tian, Y., Bader, F., & Debbah, M. (2024). Large generative AI models for telecom: The next big thing? IEEE Communications Magazine, 62(11), 84–90. DOI: 10.1109/MCOM.001.2300364
- [7] Bariah, L., Zou, H., Zhao, Q., Mouhouche, B., Bader, F., & Debbah, M. (2023). Understanding telecom language through large language models. GLOBECOM 2023 – IEEE Global Communications Conference, 6542–6547. DOI: 10.1109/GLOBECOM54140.2023.10437725
- [8] Boateng, G. O., et al. (2024). A survey on large language models for communication, network, and service management: Application insights, challenges, and future directions. arXiv:2412.19823.
- [9] Nguyen, T., Nguyen, H., Ijaz, A., Sheikhi, S., Vasilakos, A. V., & Kostakos, P. (2024). Large language models in 6G security: Challenges and opportunities. arXiv:2403.12239.
- [10] Brundage, M., et al. (2020). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. OpenAI/Partnership on AI. arXiv:2004.07213.
- [11] Cisco Systems (2022). Intent-Based Networking White Paper. cisco.com.

- [12] Friha, O., Ferrag, M. A., Kantarci, B., Cakmak, B., Ozgun, A., & Ghoualmi-Zine, N. (2024). LLM-based edge intelligence: A comprehensive survey on architectures, applications, security and trustworthiness. *IEEE Open Journal of the Communications Society*, 5, 5799–5856. DOI: 10.1109/OJCOMS.2024.3456549
- [13] ETSI GS MEC 002 V3.1.1 (2023-04). Multi-access Edge Computing (MEC); Phase 3: Use Cases and Requirements.
- [14] ETSI GS ZSM 003 (2023). Zero-touch Network and Service Management.
- [15] ITU-T Recommendation Y.3172 (2019). Architectural framework for machine learning in future networks including IMT-2020. ITU-T.
- [16] Lin, Z., Qu, G., Chen, Q., Chen, X., Chen, Z., & Huang, K. (2023). Pushing large language models to the 6G edge: Vision, challenges, and opportunities. *arXiv:2309.16739*.
- [17] Leivadeas, A., & Falkner, M. (2023). A survey on intent-based networking. *IEEE Communications Surveys & Tutorials*, 25(1), 625–655. DOI: 10.1109/COMST.2022.3215919
- [18] Liu, B., Tong, J., & Zhang, J. (2024). LLM-Slice: Dedicated wireless network slicing for large language models. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems (SenSys '24)*, 853–854. DOI: 10.1145/3666025.3699404
- [19] Shen, Y., Shao, J., Zhang, X., Lin, Z., Pan, H., Li, D., Zhang, J., & Letaief, K. B. (2024). Large language models empowered autonomous edge AI for connected intelligence. *IEEE Communications Magazine*, 62(10), 140–146. DOI: 10.1109/MCOM.001.2300550
- [20] Maatouk, A., Piovesan, N., Ayed, F., De Domenico, A., & Debbah, M. (2025). Large language models for telecom: Forthcoming impact on the industry. *IEEE Communications Magazine*, 63(1), 62–68. DOI: 10.1109/MCOM.001.2300473
- [21] Manias, D. M., Chouman, A., & Shami, A. (2024a). Towards intent-based network management: Large language models for intent extraction in 5G core networks. In *2024 20th International Conference on the Design of Reliable Communication Networks (DRCN)*, 1–6. *arXiv:2403.02238*.
- [22] Manias, D. M., Chouman, A., & Shami, A. (2024b). Semantic routing for enhanced performance of LLM-assisted intent-based 5G core network management and orchestration. In *GLOBECOM 2024 – IEEE Global Communications Conference*, 2924–2929. *arXiv:2404.15869*.
- [23] McNamara, J., Camps-Mur, D., Goodarzi, M., Frank, H., Chinchilla-Romero, L., Cañellas, F., Fernández-Fernández, A., & Yan, S. (2023). NLP powered intent based network management for private 5G networks. *IEEE Access*, 11, 36642–36657. DOI: 10.1109/ACCESS.2023.3265894

- [24] Njah, Y., Leivadeas, A., Violos, J., & Falkner, M. (2023). Toward intent-based network automation for smart environments: A healthcare 4.0 use case. *IEEE Access*, 11, 136565–136576. DOI: 10.1109/ACCESS.2023.3338189
- [25] O-RAN Alliance (2020). O-RAN Use Cases and Deployment Scenarios: Towards Open and Smart RAN. White Paper, February 2020. www.o-ran.org.
- [26] OpenAI (2023). GPT-4 Technical Report. openai.com.
- [27] Qu, G., Chen, Q., Wei, W., Lin, Z., Chen, X., & Huang, K. (2024). Mobile edge intelligence for large language models: A contemporary survey. *arXiv:2407.18921* (also *IEEE Communications Surveys & Tutorials*, DOI: 10.1109/COMST.2025.3527641).
- [28] Zheng, Y., Chen, Y., Qian, B., Shi, X., Shu, Y., & Chen, J. (2025). A review on edge large language models: Design, execution, and applications. *ACM Computing Surveys*, 57(8), Article 209. DOI: 10.1145/3719664
- [29] Kan, K. B., Mun, H., Cao, G., & Lee, Y. (2024). Mobile-LLaMA: Instruction fine-tuning open-source LLM for network analysis in 5G networks. *IEEE Network*, 38(5), 76–83. DOI: 10.1109/MNET.2024.3421306
- [30] Tarkoma, S., Morabito, R., & Sauvola, J. (2023). AI-native interconnect framework for integration of large language model technologies in 6G systems. *arXiv:2311.05842*.
- [31] Tonmoy, S. M. T. I., et al. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv:2401.01313*.
- [32] Velasco, L., et al. (2021). End-to-end intent-based networking. *IEEE Communications Magazine*, 59(10), 106–112. DOI: 10.1109/MCOM.101.2100141
- [33] Wang, J., Zhang, L., Yang, Y., Zhuang, Z., Qi, Q., Sun, H., Lu, L., Feng, J., & Liao, J. (2023). Network meets ChatGPT: Intent autonomous management, control and operation. *Journal of Communications and Information Networks*, 8(3), 239–255. DOI: 10.23919/JCIN.2023.10272352
- [34] Wei, Y., Peng, M., & Liu, Y. (2020). Intent-based networks for 6G: Insights and challenges. *Digital Communications and Networks*, 6(3), 270–280. DOI: 10.1016/j.dcan.2020.07.001
- [35] Zhang, H., Bin Sediq, A., Afana, A., & Erol-Kantarci, M. (2024a). Large language models in wireless application design: In-context learning-enhanced automatic network intrusion detection. *arXiv:2405.11002*.
- [36] Zhang, X., Meng, H., Li, Q., Tan, Y., & Zhang, L. (2025). Large language models powered malicious traffic detection: Architecture, opportunities and case study. *arXiv:2503.18487*; *IEEE Network*, DOI: 10.1109/MNET.2025.3583088.

- [37] Zhou, H., et al. (2024). Large language model (LLM) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities. arXiv:2405.10825; published in IEEE Communications Surveys and Tutorials, 27(3), 1955–2005 (2025). DOI: 10.1109/COMST.2024.3465447.
- [38] Duan, S., Wang, D., Ren, J., Lyu, F., Zhang, Y., Wu, H., & Shen, X. (2023). Distributed artificial intelligence empowered by end-edge-cloud computing: A survey. IEEE Communications Surveys & Tutorials, 25(1), 591–624. DOI: 10.1109/COMST.2022.3218527
- [39] Abbas, K., Khan, T. A., Afaq, M., & Song, W.-C. (2021). Network Slice Lifecycle Management for 5G Mobile Networks: An Intent-Based Networking Approach. IEEE Access, 9, 80128–80146. DOI: 10.1109/ACCESS.2021.3084834
- [40] Dandoush, A., Kumarskandpriya, V., Uddin, M., & Khalil, U. (2024). Large language models meet network slicing management and orchestration. arXiv:2403.13721.

