Intent-Based Network for RAN Management with Large Language Models

Fransiscus Asisi Bimo*, Maria Amparo Canaveras Galdon†, Chun-Kai Lai*,
Ray-Guang Cheng*, and Edwin K. P. Chong‡
* National Taiwan University of Science and Technology, Taiwan
† NVIDIA, USA
‡ Colorado State University, USA
Email: crg@mail.ntust.edu.tw

Abstract—**Advanced intelligent automation becomes an important feature to deal with the increased complexity in managing wireless networks. This paper proposes a novel automation approach of intent-based network for Radio Access Networks (RANs) management by leveraging Large Language Models (LLMs). The proposed method enhances intent translation, autonomously interpreting high-level objectives, reasoning over complex network states, and generating precise configurations of the RAN by integrating LLMs within an agentic architecture. We propose a structured prompt engineering technique and demonstrate that the network can automatically improve its energy efficiency by dynamically optimizing critical RAN parameters through a closed-loop mechanism. It showcases the potential to enable robust resource management in RAN by adapting strategies based on real-time feedback via LLM-orchestrated agentic systems.**

Index Terms—**intent-based network, agentic AI, LLM, Wireless Network**

I. Introduction

The rapid increase in diverse service types and dynamic resource demands in RAN introduces unprecedented operational complexity. Even though Open RAN enables the integration of components from multiple vendors, it can also increase configuration complexity and the risks of misconfiguration. This configuration heavily relies on low-level, manual procedures, making agility and scalability difficult to achieve [1].

Intent-Based Network (IBN) has emerged to manage networks by simplifying detailed technical configurations with minimal external intervention [2]. An intent serves as a high-level, human-expressible declaration of what a network should accomplish, rather than specifying how it should be achieved. It acts as a structured set of expectations, encompassing requirements, goals, conditions, and constraints [3]. An IBN then implements and manages these intents [4], fundamentally abstracting network complexity and allowing users and customers to request services without requiring detailed knowledge of their underlying provision.

Although IBN offers significant potential, a major challenge lies in accurately and dynamically translating various high-level user intents into precise RAN instructions. For IBN to work well and for networks to become fully autonomous, how accurately and reliably intent is translated is crucial. Understanding intents, which use natural language, requires intelligent translation. Incorrect translations can result in misconfigurations and severe network failures [5].

The advent of LLMs presents a powerful opportunity to overcome this critical translation gap. Some popular LLMs demonstrate excellent performance in processing text input, excelling in their ability to understand and generate natural language text [6], [7]. This makes them ideal as intelligent intermediaries capable of interpreting formalized intents and generating highly granular configuration strategies. Their advanced semantic understanding allows them to capture the subtle nuances and context inherent in complex service requirements, a significant leap beyond traditional rule-based or statistical methods.

As illustrated in Figure 1, the operational effectiveness of such a system hinges on the seamless and iterative interaction of its core components within a closed-loop control mechanism, with Intent Translation identified as the most critical and challenging component, particularly for the complex RAN environment. (e.g., ”desired performance objective for the RAN Energy Efficiency is expected to be greater than 10%.”) into precise, machine-executable configurations (e.g., specific parameters to be adjusted to achieve target, limitation of target RAN). The translation process demands not just a syntactic conversion but a semantic understanding to accurately capture the nuances of a user’s intent. Misinterpretations at this stage can lead to inefficient resource utilization, degraded Quality of Service (QoS), or even network instability. The effectiveness of an intent-based RAN system is thus fundamentally constrained by the accuracy and adaptability of its intent translation capability.

Existing studies on closed-loop network management, such as [8], translate intent using catalog databases. These approaches lack the advanced semantic understanding offered by LLMs. Separately, work in [9] uses Generative AI for intent management, but it misses a crucial closed-loop mechanism, preventing continuous self-adaptation.

This paper introduces a novel Intent-Based Wireless Network for RAN configuration, integrating LLMs as a core component to enhance intent translation. We elaborate on a structured prompt technique that leverages the cutting-edge semantic capabilities of LLMs. We demonstrate the system’s

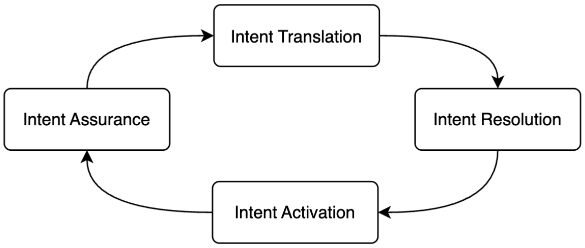


Fig. 1. Interaction of the main Intent-Based Network System components according to [2]

capability to dynamically optimize RAN parameter in energy efficiency use case through a closed-loop mechanism. The manuscript is organized as follows: Section II explains about terminology used of context within intent. Section III discuss about proposed architecture. Section IV illustrates the experiment results. Finally, Section V address the lesson learned from our experiment and potential future works.

II. Intent Terminology

Various key Internet and Telco bodies have addressed the concept of intent in their respective documents and specifications. Given that ”intent” is a fundamental term used to describe complex definitions and associated entities, establishing a common understanding of its terminology is crucial. The basic explanations from each body are described below:

A. Internet Research Task Force (IRTF)

RFC 9315 [4] is a foundational document that aims to standardize and clarify the conceptual framework for Intent-Based Networking (IBN), addressing an end-to-end management approach. It defines intent as a high-level, declarative specification of goals and desired outcomes for the network, without prescribing how to achieve them. Another term defined is Intent-Based System (IBS) which is a system that supports management functions that can be guided using intent. The document addresses critical issues concerning the understanding of IBN terms, concepts, and functionality, and resolves overlapping terminology among intent, policy, and service models. Furthermore, it describes that intent should ideally be network-wide, outcome-driven, and vendor-agnostic, enabling systems to self-manage while allowing human oversight.

B. 3GPP

In its standardization efforts for 5G networks and network slicing management, 3GPP has comprehensively defined the intent concept [3]. An intent, in this context, is understood as a set of expectations that include requirements, goals, conditions, and constraints, primarily focusing on describing ”what” needs to be achieved rather than ”how” the outcomes should be realized. As part of this framework, 3GPP further categorizes intents based on the roles related to 5G networks and network slicing management. These include Intent-CSC, which represents the declarative objectives from a Communication Service Customer (CSC) to the Communication Service Provider (CSP); Intent-CSP, conveyed from the CSP to a Network Operator (NOP) to express properties of the CSP's desired network; and Intent-NOP, which specifies characteristics of a RAN and/or 5GC network as conveyed from the NOP to a Network Equipment Provider (NEP). But 3GPP's central focus for intent-driven networks is on their Lifecycle Management, as detailed in specifications like TS 28.312 and TR 28.912, which outline comprehensive management operations and an intent information model. It also defines term for components that provide intent support called Intent Driven Management Service (MnS).

C. TMForum

TM Forum's IG1230 document [10] defines Intent as ”the formal specification of the expectations, including requirements, goals, and constraints, given to a technical system.” This document also introduces the concept of an Autonomous Network, describing it as ”a system of networks and software platforms that is capable of sensing its environment and adapting its behavior accordingly with little or no human input.”

III. Proposed Architecture

Our efforts leverage the O-RAN experimental platforms provided by [11]. Crucially, the interface establishing the connection between our system and the RAN adheres to O-RAN specifications.

Our Intent-Driven System uses an agentic approach. A Strategist Agent, responsible for intent translation, receives the formalized intent declared by the consumer. This agent leverages a LLM to perform the translation, using the user prompt structure shown in Fig. 3.

Agentic LLMs serve as the intelligent engine that translate the structured intents and generate configuration strategy. This involves a multi-stage process. First, the LLM analyzes the intent, understanding the target objects, the desired metric, the target condition, and the target value. It then decomposes this objective into a series of actionable strategies. For planning and orchestration, the agentic LLM, orchestrated through a LangGraph flow [12], determines the necessary steps to achieve the intent. This may involve interacting with the RAN simulator via O-RAN O1 interface to retrieve relevant cell details and performance metrics. The ”Strategist Agent,” as shown in Fig. 2, plays a key role in formulating a configuration strategy based on the interpreted intent and potentially leveraging past successful attempts stored by the ”History Analyzer Agent.” The ”Orchestrator Agent” then takes charge of executing these strategies by calling the appropriate configuration tools via the O-RAN O1 interface to configure the RAN simulator. While the focus of closed-loop monitoring might lean towards deterministic approaches, the initial configuration and optimization driven by the agentic LLM set the stage for efficient network operation.

To accurately simulate the complex RAN behavior for this system, we specifically use the AI RSG RAN Simulator from VIAVI Solutions.

A. *Formalized Intent Structure*

Declaration of intent by owner is formalized as a JSON object, referencing 3GPP TS 28.312 [13]. This formalized intent is then parsed to populate the Intent context, which serves as a structured input for the system. As depicted in Fig.3, the Intent context primarily comprises two distinct sections: Target KPIs and Object Target. The Target KPIs section specifies the performance metrics to be monitored and optimized (e.g., energy efficiency), along with their desired conditions and target values. Conversely, the Object Target section identifies the specific network elements or scopes to which the intent applies (e.g., a particular cell or subnetwork). For instance, an example intent might specify Object Target as following:

```
{
    "objectInstance": "SubNetwork_1",
    "ObjectTarget": ["Cell_7"]
}
```

And it declared alongside Target KPIs:

```
{
    [{
        "targetName": "
           RANEnergyEfficiency",
         "targetCondition": "
            IS_GREATER_THAN",
         "targetValue": "810000",
         "targetUnit": "bit/joule"
     }]
}
```

This comprehensive and structured representation of the intent serves as the initial input for our agentic LLM.

B. *Prompt Techniques*

We use one-shot prompting to effectively guide the LLM towards desired outputs. Specifically, we implement a five-section prompt structure to standardize the input for LLM-driven intent translation. This structured approach is crucial for ensuring consistent and predictable LLM behavior, which is essential for the reliable and automated operation of an Intent-Based Network for RAN Management. Each section of the prompt is explicitly designed to provide comprehensive context and clear directives, allowing for precise control over the LLM's decision-making process. The distinct purposes of these sections are detailed below:

- **Instruction**: Each prompt begins with an Instruction, explicitly stating the task the LLM is expected to perform.
- **Intent**: Contain the high-level objective or desired outcome from the system's perspective.
- **Current Observation**: Provides real-time or relevant operational data to give context about the current state of the network.
- **Configuration Constraints**: This section included to define any limitations for each RAN gNB to let LLM know what parameter and strategy are available to adjust.
- **Output Format**: Specifies the exact structure in which the LLM's response should be delivered. Defining structure output enables direct parsing for further computation processes.

C. *Agentic AI Framework*

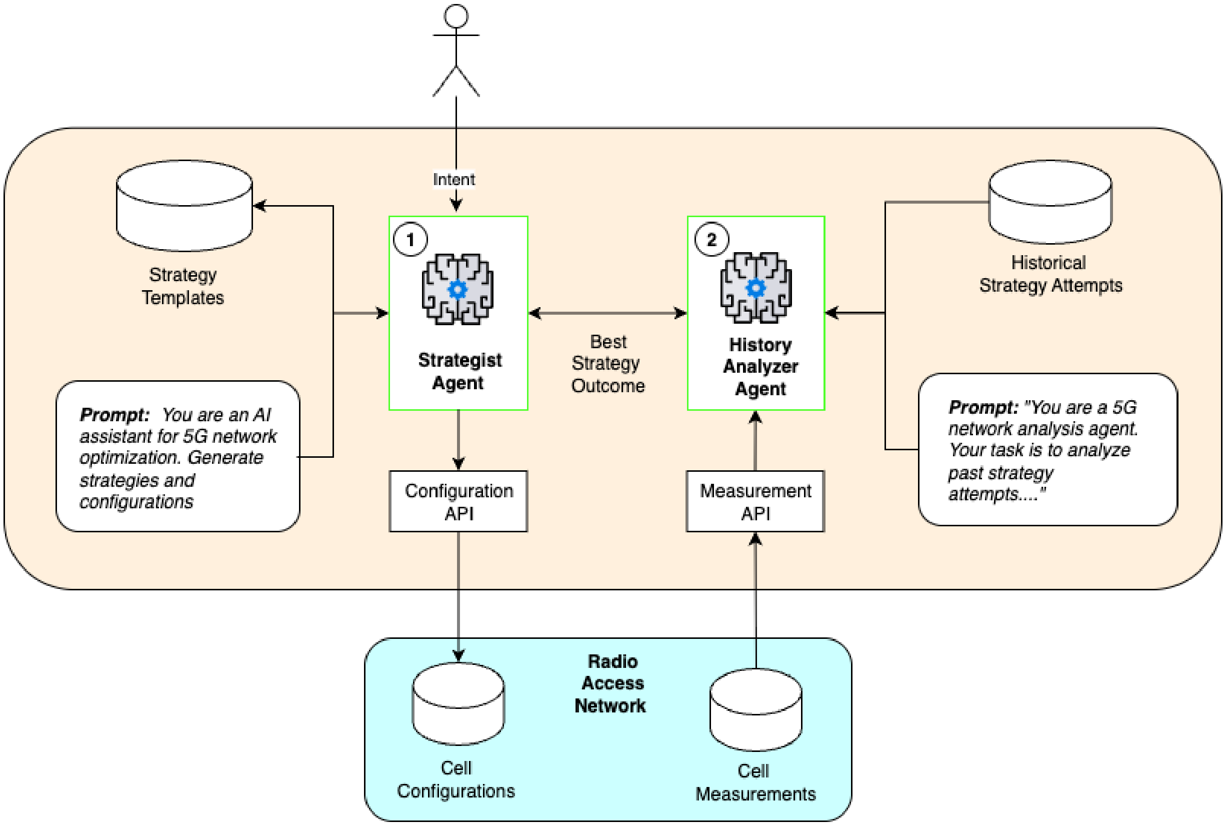
Our system is built upon two distinct agents: the History Analyzer Agent and the Strategist Agent. The History Analyzer Agent is responsible for analyzing past strategies applied in the network and obtains real-time measurements directly from the RAN as its input. This processed data, along with a number of the most relevant previous strategy attempts, is then passed to the Strategist Agent for decision-making. Crucially, both agents perform their respective inference tasks by calling the NVIDIA NIM inference API, which hosts the LLM on-cloud. The underlying intelligence for these agents is configured within the Prompt Processor component, which utilizes the LangGraph library to interface with the NVIDIA NIM endpoint, specifically leveraging the open-source Meta LLaMA 3.1 70B Instruct model.

D. *Network Configurations*

To support the agentic architecture described in our Intent-Based Network for RAN Management, we leverage both Configuration Management (CM) and Performance Management (PM) over the O-RAN O1 interface to enable closed-loop network control. The agents continuously retrieve real-time network PM data which serve as critical inputs for assessing the current operational state and evaluating the impact of previous configurations. For configuration, CM operations are performed using the NETCONF protocol. The agents first perform `<get-config>` operations to retrieve current RAN configuration parameters (e.g., transmit power levels). Leveraging the interpreted user intent and historical performance context, the Strategist Agent subsequently issues `<edit-config>` operations to apply optimized configurations to the RAN environment. This integration of CM and PM forms the backbone of the intent-based closed-loop control system.

IV. Experiment Result

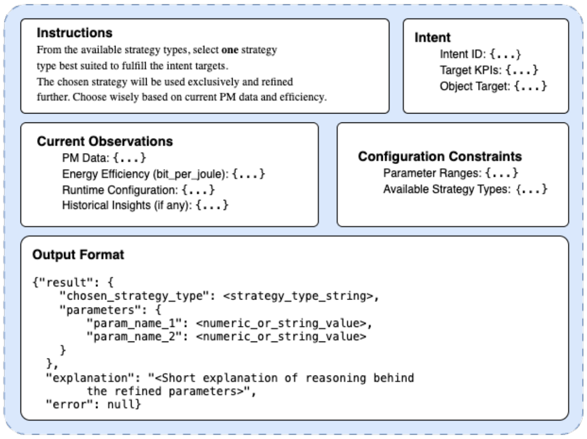
Figure 4 presents experimental results correlating `"TxPower"` with `"PEE.EnergyEfficiency"` over multiple iterations. The result obtained by performing a close-loop iteration of our system. Each iteration performed by the system by refining strategies from previous state as reference towards the required KPI set by intent. As iterations progress, there is a clear trend of the `"TxPower"` (blue line, left y-axis) decreasing significantly, starting from approximately 30 dBm at iteration 0 and converging towards 11 dBm by



Prompt: *You are an AI assistant for 5G network optimization. Generate strategies and configurations*

Prompt: *"You are a 5G network analysis agent. Your task is to analyze past strategy attempts...."*

Fig. 2. System Architecture of Intent-Based Network System



Instructions

From the available strategy types, select **one** strategy type best suited to fulfill the intent targets. The chosen strategy will be used exclusively and refined further. Choose wisely based on current PM data and efficiency.

Intent

Intent ID: {...}
Target KPIs: {...}
Object Target: {...}

Current Observations

PM Data: {...}
Energy Efficiency (bit_per_joule): {...}
Runtime Configuration: {...}
Historical Insights (if any): {...}

Configuration Constraints

Parameter Ranges: {...}
Available Strategy Types: {...}

Output Format

```
{"result": {
    "chosen_strategy_type": <strategy_type_string>,
    "parameters": {
        "param_name_1": <numeric_or_string_value>,
        "param_name_2": <numeric_or_string_value>
    }
  },
  "explanation": "<Short explanation of reasoning behind
        the refined parameters>",
  "error": null}
```

Fig. 3. User Prompt Structure for LLM Agent

iteration 4. Conversely, "PEE.EnergyEfficiency" (red dashed line, right y-axis) shows a corresponding increase, rising from around 7.7×10^5 bits/joule at iteration 0 and stabilizing near 8×10^5 bits/joule from iteration 2 onward. This inverse relationship suggests that the LLM's dynamic configuration of lower transmit power levels leads to an improved overall energy efficiency of the system, indicating a successful optimization process.

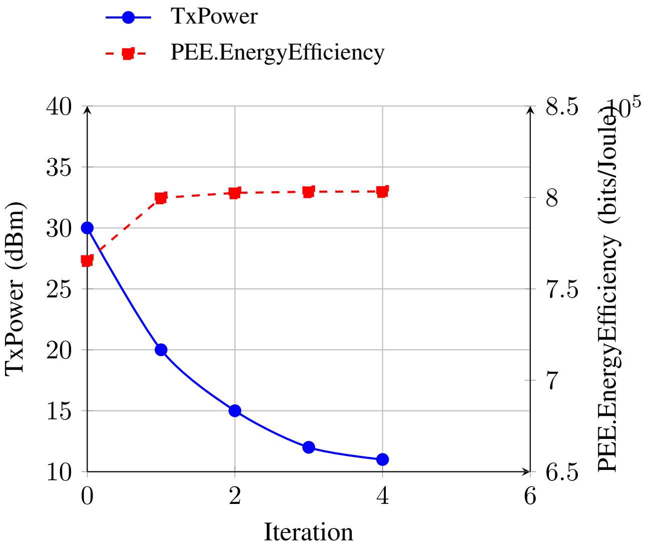


Fig. 4. Correlation between configured TxPower and PEE.EnergyEfficiency over Iteration(s).

V. Lesson Learned and Future Works

This study presents a novel Intent-Based Network for RAN Management that harnesses LLMs to optimize the management of RAN configurations. Our primary contribution lies in developing a formalized prompt engineering technique that translates high-level user intents into a structured JSON format, encapsulating intent, network topology, KPIs, and configuration parameters, thus providing a standardized interface for LLMs to interpret and achieve the desired network objectives. We further introduce an agentic structure where two LLM agents collaboratively refine network configurations through iterative optimization in a close-loop manner. This iterative process ensures the intent is achieved through gradual parameter adjustments based on real-time feedback thereby preventing network disruptions caused by abrupt configuration changes.

However, we found some key limitations:

A. Ground truth

During the preliminary evaluation of our system, we observed that feeding comprehensive PM data directly to the LLM sometimes led to hallucinations or inconsistent interpretations in the generated configuration strategy. This highlights a critical lesson: while LLMs possess remarkable capabilities in natural language understanding, their ability to generate accurate and reliable responses within highly specialized domains, such as RAN configuration, is significantly enhanced by providing them with domain-specific ground truth.

This challenge highlights the need for future research of a structured approach to serve data, explicitly detailing the relationships between entities. Such ground truth, ideally represented through formalisms like RDF (Resource Description Framework) or ontology, contains the precise nature of network entities and the fundamental properties of parameters (e.g., the exact relationship between a configuration parameter and a measurement metric). Representing and leveraging this complex domain knowledge before it is fed into the LLM is crucial for generating accurate, concise, and scientifically sound 'justification' texts that the LLM can translate effectively.

B. PM Data inaccuracy causing domino effect

The operational efficacy of LLMs within a network management framework is fundamentally predicated on its direct interaction with, and inherent assumption that, PM data represents the absolute 'ground truth.' This signifies that the LLM, by design, treats the incoming PM data—such as network traffic, latency metrics, error rates, device statuses, and resource utilization—as perfectly accurate, objective, and a faithful reflection of the real-time network state.

The LLM then consumes this PM data as critical additional context for its reasoning and decision-making processes. Unlike traditional rule-based systems, an LLM doesn't merely follow pre-programmed instructions; instead, it leverages its extensive training and in-context learning capabilities to 'understand' the intricate nuances and complex relationships within this data. This 'understanding' enables it to interpret anomalies, predict potential issues, or identify opportunities for optimization, subsequently formulating high-level operational decisions or recommendations for network actions.

However, this reliance on PM data as an unquestionable truth introduces a significant vulnerability. Should the PM data be inaccurate, incomplete, or delayed, it initiates a perilous 'domino effect' that cascades through the entire closed-loop automation process. An erroneous initial understanding of the network state, derived from flawed PM data, will inevitably lead the LLM to make suboptimal, incorrect, or even detrimental decisions. These flawed decisions, once implemented, will then generate new, equally unreliable PM data, further corrupting the LLM's subsequent interpretations and actions. This feedback loop of misinformation can rapidly destabilize the network, leading to performance degradation, service outages, or inefficient resource allocation, ultimately undermining the very purpose of the automated management system.

This challenge underscores the need for future research into reinforcement learning (RL) algorithms to achieve system resilience. RL can adapt to noisy or outlier data via iterative feedback and reward-based optimization. Pursuing this research direction will advance the development of robust and intelligent RAN management, ensuring consistent performance under diverse and imperfect data conditions.

Acknowledgment

This work was supported by the National Science and Technology Council, Taiwan, under Contract numbers 111-2221-E-011-069-MY3, 112-2218-E-011-004, and 112-2218-E-011-006.

References

[1] W. Azariah, F. A. Bimo, C.-W. Lin, R.-G. Cheng, N. Nikaein, and R. Jana, "A survey on open radio access networks: Challenges, research directions, and open source approaches," *Sensors*, vol. 24, no. 3, p. 1038, Feb. 2024. [Online]. Available: http://dx.doi.org/10.3390/s24031038

[2] A. Leivadeas and M. Falkner, "A survey on intent-based networking," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 625–655, 2023.

[3] G. W. SA5, "Intent driven management." [Online]. Available: https://www.3gpp.org/technologies/intent

[4] A. Clemm, L. Ciavaglia, L. Z. Granville, and J. Tantsura, "Intent-Based Networking - Concepts and Definitions," RFC 9315, Oct. 2022. [Online]. Available: https://www.rfc-editor.org/info/rfc9315

[5] National Telecommunications and Information Administration (NTIA), "Open RAN Security Report," National Telecommunications and Information Administration, U.S. Department of Commerce, Tech. Rep., 2023. [Online]. Available: https://www.ntia.gov/sites/default/files/publications/open_ran_security_report_full_report_0.pdf

[6] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[7] A. Fan, B. Gokkaya, M. Harman, M. Lyubarskiy, S. Sengupta, S. Yoo, and J. M. Zhang, "Large language models for software engineering: Survey and open problems," in *2023 IEEE/ACM International Conference on Software Engineering: Future of Software Engineering (ICSE-FoSE)*, 2023, pp. 31–53.

[8] T. Ahmed Khan, K. Abbas, J. J. Diaz Rivera, A. Muhammad, and W.-c. Song, "Applying routenet and lstm to achieve network automation: An intent-based networking approach," in *2021 22nd Asia-Pacific Network Operations and Management Symposium (APNOMS)*, 2021, pp. 254–257.

[9] D. Brodimas, K. Trantzas, B. Agko, G. C. Tziavas, C. Tranoris, S. Denazis, and A. Birbas, “Towards intent-based network management for the 6g system adopting multimodal generative ai,” in *2024 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, 2024, pp. 848–853.

[10] TMForum, “Tm forum introductory guide autonomous networks technical architecture;,” TMForum, Tech. Rep. Version 1.1.1, 2023.

[11] F. A. Bimo, R.-G. Cheng, C.-C. Tseng, C.-R. Chiang, C.-H. Huang, and X.-W. Lin, “Design and implementation of next-generation research platforms,” in *2023 IEEE Globecom Workshops (GC Wkshps)*, 2023, pp. 1777–1782.

[12] LangChain, “langgraph.” [Online]. Available: https://langchain-ai.github.io/langgraph/

[13] 3GPP, “Lte; 5g; management and orchestration; intent driven management services for mobile networks (3gpp ts 28.312 version 18.6.0 release 18),” 3GPP, Tech. Rep. ETSI TS 128 312 V18.6.0 (2025-01), 2025.