A Telecom Generative AI Marketplace for 6G Open Networks: Architecture and Early Evaluation

Peizheng Li, Adrián Sánchez-Mompó, Tim Farnham, Aftab Khan, Adnan Aijaz

Bristol Research and Innovation Laboratory, Toshiba Europe Ltd., U.K.
Email: {firstname.lastname}@toshiba-bril.com

***Abstract*—Generative artificial intelligence (GAI) has emerged as a pivotal technology for content generation, reasoning, and decision-making, making it a promising solution on the 6G stage characterized by openness, connected intelligence, and service democratization. This paper explores strategies for integrating and monetizing GAI within future open 6G networks, mainly from the perspectives of mobile network operators (MNOs). We propose a novel API-centric telecoms GAI marketplace platform, designed to serve as a central hub for deploying, managing, and monetizing diverse GAI services directly within the network. This platform underpins a flexible and interoperable ecosystem, enhances service delivery, and facilitates seamless integration of GAI capabilities across various network segments, thereby enabling new revenue streams through customer-centric generative services. Results from experimental evaluation in an end-to-end Open RAN testbed, show the latency benefits of this platform for local large language model (LLM) deployment, by comparing token timing for various generated lengths with cloud-based general-purpose LLMs. Lastly, the paper discusses key considerations for implementing the GAI marketplace within 6G networks, including monetization strategy, regulatory, and service platform aspects.**

***Index Terms*—6G, generative AI, large language models, marketplace, monetization, open networks, platform.**

I. Introduction

Generative artificial intelligence (GAI) has emerged as a compelling and prominent research area due to its proven success in content generation services. Large GAI models, such as large language models (LLMs), image and video generation models, and multi-modality models, excel at understanding language and performing general-purpose tasks.

Models like GPT-4, Deepseek, Gemini, LLaMA, and Claude demonstrate powerful capabilities in context understanding, planning, responding, and code generation. These models can be customized for specific industries using techniques like retrieval-augmented generation (RAG), low-rank adaptation (LoRA), and prompt-tuning for cost-effective updates.

Integration of AI and communication networks is at the heart of 6G evolution, as highlighted in the "IMT-2030 Framework" [1]. GAI, especially through LLMs [2], is seen as a key enabler of this integration, enhancing wireless communication systems with advanced understanding, reasoning, and generating capabilities – essential for developing in-network intelligence. GAI is increasingly being adopted as a service in the telecoms sector. For example, LLMs are used for analyzing 3GPP specifications, data synthesis, and goal-oriented and semantic communication for efficiency improvement [3].

GAI services are expected to significantly impact the socioeconomics of the telecoms industry [4], by delivering customer-centric services, improving operational efficiency, creating new products and revenue streams, reducing costs, and fostering innovation. It should be noted that the role and practices of mobile network operators (MNOs) are crucial in this evolution.

However, integrating GAI with telecoms systems and networks remains challenging for MNOs. Ongoing GAI-focused research is driven by computer sciences rather than telecoms and is generally confined to academic and AI circles. While MNOs can access GAI capabilities for potential services like conversational chatbots [4] via cloud platforms (e.g., Google's), such solutions are not attractive due to limited prospects of intrinsic control and customization of GAI services. Three key issues need to be addressed in this respect.

- *Lack of fine-grained control.* Third-party cloud solutions often function as black boxes, limiting an MNO's ability to enforce domain-specific policies, guarantee low-latency responses, or optimize model usage.
- *Scalability and cost structure.* Large GAI models can be very costly to train and deploy at scale. Offloading tasks to generic cloud LLMs may not produce the desired return on investment for telecom-specific usage.
- *Fragmented business models.* As GAI moves in-network, new revenue streams arise (e.g., AR/VR experiences, generative chatbots for enterprise customers), necessitating robust billing and service orchestration frameworks.

The key principles of open radio access network (Open RAN), i.e., openness, disaggregation, cloudification, and programmability, hold transformative potential for 6G architectural evolution. The traditional monolithic RAN is embracing disaggregated RAN components [5] while MNOs are increasingly adopting cloud-neutral platforms supporting multi-cloud, private cloud, and hybrid configurations. While the shift towards openness complicates network management, especially with AI integration across RAN, edge, and cloud layers, it also provides opportunities for innovation [6]. As AI evolves into GAI and 6G networks become more open, there is a growing need to address the following issues and explore monetization strategies for GAI within open networks.

1) Monetization and ROI: Large-scale GAI deployment entails high CAPEX and OPEX. Sustainable *revenue models*, such as integrating GAI into premium plans or enterprise services, are essential to ensure ROI.

978-3-903176-77-5© 2025 IFIP

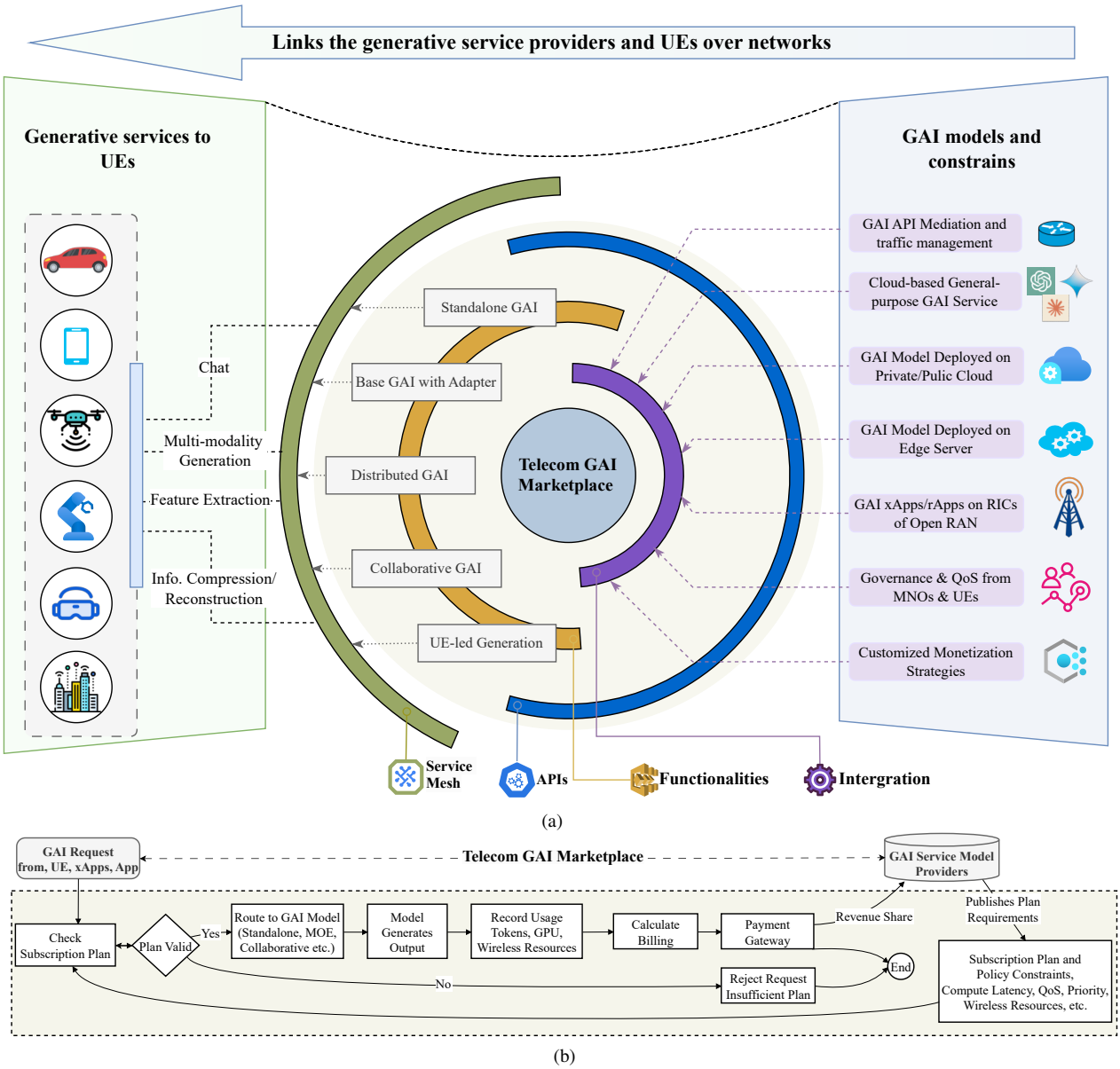


Fig. 1. (a) Illustration of our conceptual telecom marketplace platform for GAI model integration and service delivery. It supports model registration, container-based deployment, multi-cloud integration, and open APIs for usage and billing; (b) Flowchart illustrating the monetization process in the telecom GAI marketplace.

2) Control and Management: MNOs require fine-grained control to meet telecom-grade reliability, security, and QoS standards, and dynamic orchestration across RAN, edge, and core capabilities often lacking in conventional cloud solutions.

3) Service Quality and Latency: Interactive and real-time AI applications demand ultra-low latency. Deploying GAI at the network edge and enabling flexible, edge-aware, multi-cloud orchestration are key to maintaining predictable performance.

4) Regulatory and Privacy Compliance: Telecom data is sensitive and subject to strict regulations (e.g., GDPR, data sovereignty laws). MNOs must control inference locations and enforce robust data anonymization and protection mechanisms.

Addressing these issues requires a framework that supports diverse GAI models, monetization strategies, strong management capabilities, and policy compliance.

To our knowledge, this is one of the first studies proposing a monetization strategy for GAI models integrated into 6G networks, with telecoms GAI marketplace in the spotlight. The main contributions of this paper are summarized as follows:

- We design and implement an API-centric telecoms GAI marketplace platform, serving as the entry point for heterogeneous GAI services deployed across various network segments and the exit for integrated and meshed GAI services. This marks a significant step toward a (G)AI-native network, enabling seamless AI integration within telecom infrastructure.
- We demonstrate an in-network GAI deployment use case

within an end-to-end Open RAN network, wherein an LLMs-based generative service is examined, highlighting the benefits of this approach in terms of reduced service latency compared to general-purpose cloud GAI services.
- We provide a detailed discussion on the marketplace framework, covering aspects like service access, monetizing, regulation, management, and open service platform.

II. Marketplace Solution for GAI and Open Network Integration

The telecom GAI marketplace is the core integration and governance layer that connects heterogeneous generative AI services to disaggregated 6G networks. As shown in Fig. 1a, it establishes a robust ecosystem enabling MNOs, third-party developers, and enterprise customers to seamlessly deploy and consume GAI services across disaggregated networks.

A. Telecoms Marketplace Design Principles

- **Unified abstraction.** Standalone LLMs, adapter-based fine-tuning (e.g., LoRA), and collaborative or distributed inference pipelines are presented through a consistent API.
- **Orchestration.** Placement and scaling across edge, near-edge, and cloud layers allow operators to balance latency, cost, and compliance goals.
- **Service mesh governance.** Mesh proxies enforce secure service-to-service communication, traffic control, policy enforcement (e.g., data residency), and telemetry.
- **Metering and monetization.** Token counts, GPU time, and bandwidth usage are recorded to support pay-per-use, subscription, or revenue-sharing models.

B. Monetization Process

Fig. 1b illustrates the integration and monetization flow. A GAI request enters through the marketplace gateway, which checks subscription and policy constraints. Valid requests are routed to edge or cloud inference endpoints depending on performance and regulatory policies. The platform meters token, compute, and traffic usage during inference, then forwards the metrics to the billing engine for cost calculation and optional revenue sharing with third-party developers. This closes the loop between network policy, AI service delivery, and monetization.

C. Implementation Framework

The marketplace implementation builds on the strategies outlined in [7], originally designed for the Open RAN ecosystem. It emphasizes an API-centric integration Platform-as-a-Service (iPaaS) model to facilitate the seamless integration, deployment, and monetization of GAI models, applications (x/rApps), and other services within a 5G Open RAN infrastructure, leveraging Open RAN API standards. The key design features of this marketplace are as follows:

1) API-Centric iPaaS Model: The marketplace uses an iPaaS approach, focusing on API-based integration to enable flexible deployment and monetization across different services and environments. This model provides fine-grained access control and monitoring, supporting various business models like pay-per-use, subscriptions, and SLAs.
2) Multiple Deployment Environments: The marketplace supports various runtime environments–edge, cloud, and hybrid setups. Deployment agents automate service deployment across these environments, improving flexibility and optimization.
3) Integration Features: It utilizes WSO2 API management to enable easy integration and deployment of applications. API gateways using Choreo Connect and Envoy proxies manage access and ensure secure service communication. It implements a federated service mesh to facilitate secure interactions between cloud and edge data centers.
4) Monitoring and Reconciliation: It incorporates robust monitoring and billing mechanisms with tools like Stripe marketplace plugins, Choreo analytics, and Hyperledger blockchain for decentralized auditing, ensuring accurate performance tracking, billing, and compliance.

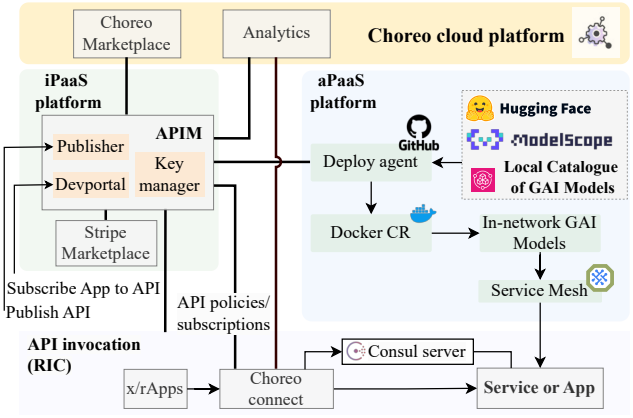


Fig. 2. Overall marketplace integration architecture with API invocation. Multiple distributed or collaborative GAI models can be handled seamlessly via centralized API publishing, management, and usage-based billing.

Fig. 2 illustrates the overall marketplace integration architecture. The WSO2 API manager handles the marketplace management APIs and portals, enabling API publishing and subscription, and the creation of API product bundles that represent integrated applications utilizing multiple GAI services. These bundles are propagated to marketplace billing platforms Stripe, allowing for the provision of optimized, ready-to-use services for specific application use cases. This is valuable for non-developer users who prefer not to handle selection, evaluation, testing, and service integration themselves.

GAI services and product bundles are deployed using YAML scripts from GitHub and containers from Docker Hub, Hugging Face, or other repositories. The Kubernetes-based YAML scripts used by the deployment agents are designed for low complexity and flexibility, featuring annotations for automated injection of service mesh sidecars and security configurations for different deployment environments. This approach allows users to integrate services without needing to be code developers.

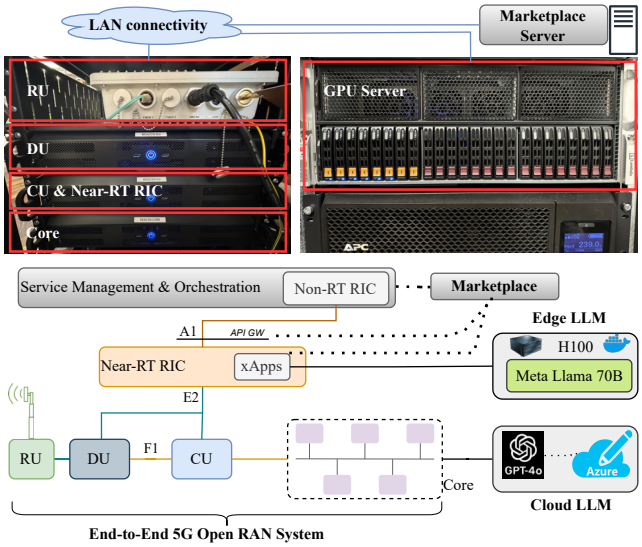


Fig. 3. Illustration of the experimental setup. The local LLM container is deployed on the edge server near the Open RAN CU. Requests are managed via the marketplace's orchestrator.

The Federated Consul service mesh is used to securely manage service interactions, with Consul server instances deployed for each environment. This lightweight and highly scalable solution is available as an open-source addition. For cross-tenant integration between isolated service mesh clusters deployed in different environments, the Choreo microgateway is used, leveraging the lightweight Envoy proxy.

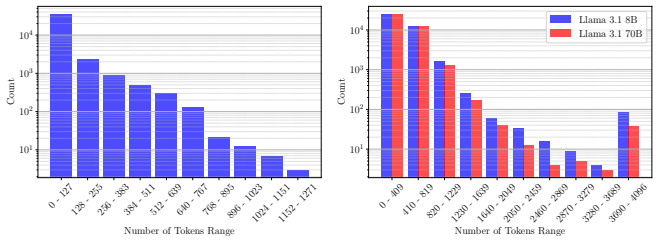
This GAI marketplace is highly scalable, leveraging cloud-based service distribution to dynamically allocate resources across edge, near-edge, and central cloud environments.

III. Early Evaluation: Edge vs. Cloud LLM through Marketplace

To validate the marketplace concept, we conducted a proof-of-concept experiment on an end-to-end Open RAN testbed [8], where a local LLM service was deployed at the edge linked to the Centralized Unit (CU) of the Open RAN and compared against cloud-hosted models under realistic loading.

The edge node (2×H100 GPU) hosted *Llama 3.1 8B/70B* models via a vLLM agent [9], while the cloud baselines used GPT-3.5 Turbo and GPT-4o. A background load was applied to emulate realistic usage. Fig. 4 shows the input/output token distribution of the prompts, concentrated at low token counts with mild long-tail outputs. Notably, the 8B model exhibited a tendency to produce longer responses. This may stem from model-specific prompt handling and tokenization behavior, and it has direct implications for token-based billing and marketplace cost control.

The LLM models in both the edge and cloud setups were configured to handle requests with specific parameters. The input tokens were set to 10, while the maximum output tokens allowed were 1000. Streaming functionality was enabled, and where applicable, the models were configured to ignore the end of sequence (EOS) token.



(a) (b)

Fig. 4. Histograms of Input and Output token lengths for the Chatbot arena dataset with the Llama 3.1 8B and 70B models, wherein (a) indicates Log-scaled Histogram of Input Tokens, (b) show Log-scaled Histogram of Output Tokens for Llama 3.1 Models.

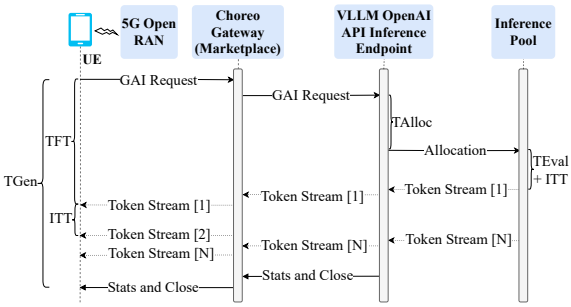


Fig. 5. Illustration of the process of GAI content generation over marketplace platform and Open RAN testbed and time measurements.

A. Testing Procedure

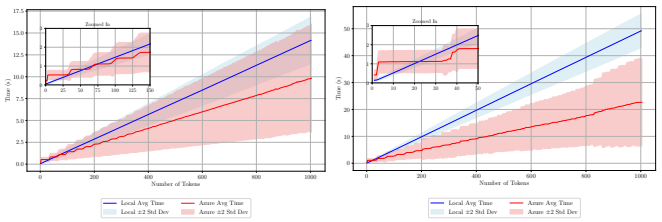
To compare cloud-based and edge-based LLM deployments in terms of latency during content generation tasks, identical requests were sent to both under controlled conditions, and key latency metrics were assessed. Fig. 5 illustrates the measurement procedure.

- **Time-to-First-Token (TFT)** — latency between request submission and the first token arrival.
- **Inter-Token Time(ITT)** — average interval between streamed tokens.

B. Results

Fig. 6 compares edge-based and cloud-based LLM deployments for both model sizes. Across both scenarios, the edge-deployed Llama models consistently achieved lower TFT than their cloud counterparts, enabling faster initial responses suitable for interactive and short-text services. Conversely, cloud-based GPT models delivered lower and more stable ITT, offering better sustained throughput for longer generations.

This complementary latency behavior highlights a key design implication: the telecom GAI marketplace can dynamically route requests according to application demands—e.g., edge for latency-sensitive or short interactions, cloud for long-form or compute-heavy tasks—enabling differentiated service quality and monetization. While this early evaluation focuses on latency, additional performance dimensions such as throughput, concurrent user capacity, and resource efficiency are also critical for MNO deployment decisions and will be explored in future work.



(a) (b)

Fig. 6. Token timing for Edge vs. cloud comparison: (a) GPT-3.5 Turbo vs. Llama 3.1 8B; (b) GPT-4o vs. Llama 3.1 70B.

IV. Discussion

A. Monetization and Business Models

A telecom-native marketplace enables business models that go far beyond generic cloud billing. Rather than a one-size-fits-all *per-token* or *per-API* charge, operators can tie pricing directly to network and compute usage. This enables more transparent, flexible, and incentive-aligned cost structures:

- **Resource-based accounting:** metering GPU time, bandwidth, and memory usage ensures fair and predictable billing.
- **Vertical bundling:** generative services can be packaged with 5G/6G slices or enterprise data plans, turning AI into a native network feature rather than an add-on.
- **Revenue sharing:** third-party developers can receive royalties when their adapters or models are invoked, stimulating an open AI-service economy.

For example, revenue can be shared between MNOs and third-party model providers based on token usage or SLA tiers, ensuring both infrastructure owners and service innovators benefit. This also provides MNOs with pricing control, which is essential given the potential misalignment of incentives between telecom operators and external AI vendors.

B. Regulation and Governance

Because the marketplace sits at the network's control plane, it naturally becomes a policy enforcement point. This allows operators to align GAI usage with existing telecom-grade compliance frameworks. Key governance requirements include: *(1) Privacy protection and anonymization* of sensitive user or enterprise data; *(2) Data residency enforcement* within regulated geographic boundaries; *(3) Fairness and auditability* to mitigate bias and harmful outputs. Unlike external cloud APIs, such controls can be embedded directly into the service fabric, offering both regulatory confidence and operational transparency.

C. Operations and Multi-Operator Ecosystem

For marketplaces to scale, operational agility is as critical as technical capability. Embedding monitoring, automated updates, and resource orchestration ensures that services remain reliable and cost-efficient. Moreover, coupling the marketplace with an open service platform (OSP) enables *federation across multiple operators*, supporting shared service catalogs and collaborative deployments. This ecosystem view turns individual networks into part of a larger, policy-driven AI infrastructure, paving the way for cross-operator services in future 6G deployments.

V. Conclusions and Future Directions

In this paper, we have presented a telecoms GAI marketplace solution aimed at maximizing the benefits of large generative AI models in open 6G networks. From an MNO standpoint, our approach addresses key hurdles in *monetization, management,* and *service quality*. By hosting, orchestrating, and billing GAI services directly within the network, operators can reduce dependencies on external cloud providers, gain finer control over latency and data governance, and co-create novel service bundles with third-party developers.

Our proof-of-concept deployment in an Open RAN testbed shows promising latency advantages for local LLMs. Cloud-based solutions may retain strengths in sustained token throughput, underscoring the need for an intelligent marketplace that routes requests according to specific application demands. The marketplace paradigm is readily extensible to future multi-modal models and advanced inference pipelines.

Future work can delve into collaborative AI approaches (e.g., mixture-of-experts or hierarchical inference across multiple edge servers), multi-operator interoperability, and advanced data privacy frameworks such as homomorphic encryption or differential privacy for GAI tasks. Close alignment with evolving 6G standards and open network specifications (e.g., O-RAN Alliance, 3GPP, AI-RAN Alliance) will be critical to ensure that telecoms GAI marketplaces seamlessly integrate into next-generation architectures. We hope this paper lays a foundation for broader adoption of generative AI in open 6G networks, ultimately driving both technical innovation and new revenue opportunities for MNOs.

Acknowledgment

This work was supported partially by the 6G-GOALS project under the 6G SNS-JU Horizon program, n.101139232.

References

[1] "IMT-2030 Vision - International Telecommunication Union (ITU)," https://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2030/Pages/default.aspx, accessed: August 31, 2024.
[2] L. Bariah *et al.*, "Large Generative AI Models for Telecom: The Next Big Thing?" *IEEE Commun. Mag.*, 2024.
[3] P. Li and A. Aijaz, "Task-Oriented Connectivity for Networked Robotics with Generative AI and Semantic Communications," in *IEEE INFOCOM Workshops*, 2025, pp. 1–6.
[4] A. Maatouk *et al.*, "Large Language Models for Telecom: Forthcoming Impact on the Industry," *IEEE Commun. Mag.*, 2024.
[5] M. Polese *et al.*, "Empowering the 6G Cellular Architecture With Open RAN," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 2, pp. 245–262, 2024.
[6] E. C. Strinati *et al.*, "Goal-oriented and semantic communication in 6g ai-native networks: The 6g-goals approach," in *Proc. of 2024 EuCNC/6G Summit*, 2024, pp. 1–6.
[7] T. Farnham *et al.*, "Demo: Integration of Marketplace for the 5G Open RAN Ecosystem," in *Proc. of IEEE ICNP*, 2023, pp. 1–2.
[8] A. Aijaz *et al.*, "Open RAN for 5G Supply Chain Diversification: The BEACON-5G Approach and Key Achievements," in *Proc. of IEEE CSCN*. IEEE, 2023, pp. 1–7.
[9] W. Kwon *et al.*, "Efficient Memory Management for Large Language Model Serving with PagedAttention," 2023. [Online]. Available: https://arxiv.org/abs/2309.06180