Large Generative AI Models meet Open Networks for 6G: Integration, Platform, and Monetization

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

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 article 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 cloudbased general-purpose LLMs. Lastly, the article discusses key considerations for implementing the GAI marketplace within 6G networks, including monetization strategy, regulatory, management, 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, 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

The authors are with the Bristol Research and Innovation Laboratory, Toshiba Europe Ltd., U.K. (e-mail: peizheng.li@toshiba-bril.com).

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

3GPP specifications, data synthesis, anomaly detection, network modeling, and personalized service generation. Image generation models are applied in the goal-oriented and semantic communication paradigm for image reconstruction.

GAI services are expected to significantly impact the socioeconomics of the telecoms industry [3], by delivering customercentric 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 generally confined to academic and AI circles. Hence, there is a lack of clear pathways for MNOs to effectively integrate GAI and monetize services. Three key issues need to be addressed in this respect.

- Integration of Diverse GAI Models: Large GAI models, which often have billions of parameters, require substantial storage and computational resources. Besides, multiple models are required for various task types.
- Customized Billing Strategies: The training costs of the telecoms vertical models are high, necessitating the development of appropriate billing mechanisms to recoup investments.
- 3) Building a GAI-Focused Revenue Sharing Ecosystem: This involves leveraging the collaborative efforts of stakeholders, including MNOs and third-party GAI developers.

While MNOs can access GAI capabilities for potential services like conversational chatbots [3] via cloud platforms (e.g., Google's), such solutions are not attractive due to limited prospects of intrinsic control and customization of GAI services.

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 [4] 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.

As AI evolves into GAI and 6G networks become more open, there is a growing need to address these issues and explore monetization strategies for GAI within open networks. This article investigates a marketplace strategy for integrating GAI in networks, focusing on collaboration, service access

and management, monetization, and the development of GAInative 6G open networks. 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 article are summarized as follows:

- We design and implement an API-centric telecoms GAI marketplace platform. This marketplace serves as the entry point for heterogeneous GAI services deployed across various network segments and the exit for integrated and meshed GAI services.
- We demonstrate an in-network GAI deployment use case within an end-to-end Open RAN network, highlighting the benefits of this approach in terms of reduced service latency compared to general-purpose cloud-based GAI services.
- We provide a detailed discussion on the marketplace framework, covering aspects such as service access, monetizing, regulation, management, and open service platform.

II. PREMILIARIES

A. Open Networks

In 6G context, an open network is characterized by an architecture that promotes interoperability and flexibility by adhering to open standards and interfaces across the RAN, core network, and cloud infrastructure. This approach facilitates the seamless integration of components from various vendors, supports modular and programmable network architectures, and fosters a competitive ecosystem by avoiding vendor lockin. Open networks include the following features:

- Vendor-Neutral Infrastructure: Supports interoperability between equipment and software from multiple vendors, allowing MNOs to build diversified networks.
- Multi-Cloud Flexibility: Enables the network to operate seamlessly across multiple cloud service providers, including public, private, and hybrid environments. This enhances network resilience, optimizes performance through strategic workload distribution, and offers cost management benefits through varied pricing models.
- Third-Party App Deployment: Allows external developers to deploy applications on the network, diversifying the application ecosystem. This enhances customization, meets specific needs, and opens new revenue opportunities through third-party partnerships.

B. Overview of the SOTA Large Models

Large GAI models are advanced neural networks with billions of parameters, trained on vast datasets using powerful computational resources. They generate human-like data (text, images, videos) using transformer and attention mechanisms. These state-of-the-art models are actively researched and applied in language, image/video, and multi-modality generation.

Early milestones in GAI include OpenAI's GPT-3, which significantly advanced natural language processing and generation, showcasing the potential of LLMs across various applications. Building on this foundation, GPT-4 further enhanced

the field by improving the model's ability to understand and generate human-like text.

In image generation, a major breakthrough was achieved with DALL-E, which introduced the capability to generate images from textual descriptions. The subsequent release of DALL-E 3 excelled in creating detailed and coherent images, making it an invaluable tool for artists and designers. Stable Diffusion further refined text-to-image generation by producing high-quality, detailed images from textual inputs. Other notable GAI services available in the market include Claude, Gemini, and LLaMA.

These generative models are generally designed for generalpurpose use, targeting a wide range of users and applicable to various tasks across different fields, often without substantial modification and fine-tuning.

Meanwhile, purpose-specific GAI models are emerging, trained or fine-tuned with specific knowledge in vertical domains. For instance, Codex powers GitHub Copilot, providing developers with intelligent code completion and suggestions. In the communication and network domain, NetGPT [5] has been developed for understanding and generating network traffic; Autonomous edge AI is raised in [6] for connected intelligence by using LLMs.

C. Cloud-based GAI Pricing Schemes

Cloud-based GAI services, including the models discussed earlier, typically use pricing schemes based on API call frequency, token count, subscription levels, or the need for customized services. Schemes are briefly discussed below:

- 1) Charges are based on the number of API calls made to generate a response or handle a request. This model is often preferred for services with predictable costs.
- Pricing is based on the amount of data processed, measured in tokens or characters. Charges depend on the number of tokens generated or the input text length.
- 3) Providers may offer subscription plans, typically for chatoriented LLMs. These plans might include a set number of API calls, data processing limits, or additional features. Higher-tier subscriptions can offer benefits like priority support, customized models, or higher concurrent request handling.
- 4) Additional charges apply for customized model training, specific domain adaptations, or the use of adapters for customized models. These fees vary depending on the provider and specific use case.

D. Marketplace

In the context of technology and digital services, a marketplace refers to an online platform that connects buyers and sellers, facilitating the discovery, purchase, and delivery of products or services. These marketplaces can vary widely in their focus and structure, but they all share the common goal of simplifying and streamlining transactions between parties. In the technology sector, marketplaces have become an essential aspect of how software, cloud services, and other digital products are distributed and consumed.

Marketplaces can be tailored for various purposes. For example, cloud service marketplaces such as AWS, Microsoft Azure, and Google Cloud marketplace offer curated catalogs of software applications, infrastructure, and services that seamlessly integrate with their respective ecosystems. These platforms typically provide software-as-a-service (SaaS), platform-as-a-service (PaaS), and infrastructure-as-a-service (IaaS) solutions.

III. METHODS OF INTEGRATING LARGE GAI MODELS AND OPEN NETWORKS

Integrating GAI models into telecoms involves applying GAI capabilities to edge, RAN, core, and cloud network functions. Research in this area is progressing along various paths. The following sections explore methods for integrating GAI models with open networks.

A. AI-native Network Architecture Design

AI-native network architecture design is a key enabler for integrating GAI into 6G networks. This approach involves embedding AI capabilities into (every) stack of the network, ensuring that AI-driven functionalities are an integral part of network operations. With well-defined architecture and lifecycle management methods for AI, networks can deploy AI models flexibly and at scale, dynamically adapting to varying workloads and performance demands. For instance, one study [7] explores an architecture that coordinates cloud AI, edge AI, and network AI to deliver intelligent, customized services in future 6G networks. Another work [8] introduces an open-source edge AI framework, embedding a native AI plane within a multi-access edge computing framework. In this AI-native design, GAI can be seamlessly integrated into the network.

B. Goal-oriented and Semantic Communication

Goal-oriented and semantic communication is an emerging paradigm that prioritizes the transmission of meaning over raw data [9], enabling more efficient and context-aware communication that surpasses traditional Shannon capacity. In semantic communication, the sender and receiver exchange background knowledge relevant to specific tasks, which can be utilized in the design of semantic encoders and decoders. GAI is increasingly being considered as a backbone for semantic communication, serving as the semantic encoder/decoder and even as the knowledge base itself. For example, GAI can function as a decoder, reconstructing the original input from the semantic latent space using GAI models. In [10], a semantic communication framework based on a large AI model is proposed, where the segment anything model is adopted as the knowledge base.

However, it should be noted that the aforementioned methods fail to account for the diversity of GAI models and the contributions of the current development community. To address this, we propose a marketplace-based integration approach that leverages the programmability of open networks. By utilizing API-based subscriptions, we aim to create a new ecosystem for integrating and monetizing GAI within open networks.

IV. CHALLENGES FROM THE MNO PERSPECTIVE

This section examines the key considerations for MNOs and underscores the value of a marketplace approach for improved management, customization, and integration for innetwork GAI. This strategy allows MNOs to fully leverage AI's potential and remain competitive in the rapidly evolving telecoms landscape.

A. Monetization

GAI is expected to automate network operations and improve efficiency, resulting in cost savings, particularly through reduced energy consumption. This efficiency motivates MNOs to integrate and develop GAI within their networks. However, monetization remains a significant challenge. MNOs need to create sustainable business models that guarantee a return on investment and leverage AI-driven services to generate revenue. This includes exploring direct revenue streams and enhancing existing services and customer experiences to boost profitability.

While cloud-based GAI offers powerful capabilities, it often lacks the customization and monetization options needed by telecoms operators, as these services are generally designed for broad applications and may not fit telecoms-specific revenue models. MNOs therefore need tailored GAI solutions within their networks.

For instance, MNOs can generate revenue by operating network infrastructure and marketplace services. Developers of applications and inference engines can also earn revenue if their products are commercially used on these platforms. Similarly, owners of base models and adapters can generate income under commercial use scenarios or by deploying their models in mobile applications, sharing revenue with MNOs.

B. Control and Management Capability

Control and management are also crucial considerations. Cloud-based GAI services often function as black boxes, offering limited transparency and control over their internal workings. This lack of control is problematic for MNOs, as it hinders their ability to ensure network reliability, security, and compliance with industry regulations. Managing GAI-driven operations in the dynamic and sensitive environment of telecoms networks requires advanced tools and frameworks that provide real-time insights, granular control, and seamless orchestration – capabilities that current cloud-based solutions are not fully equipped with.

C. Service Quality

Maintaining service quality amidst GAI integration is paramount; GAI models should enhance, not compromise, network performance metrics like latency, throughput, and reliability. MNOs must balance the sophisticated demands of AI processing with the need to deliver consistent, high-quality service to their customers. Relying on remote cloud services can cause delays and degrade the quality of service (QoS), negatively affecting user experience. MNOs require GAI solutions that are tightly integrated with their network

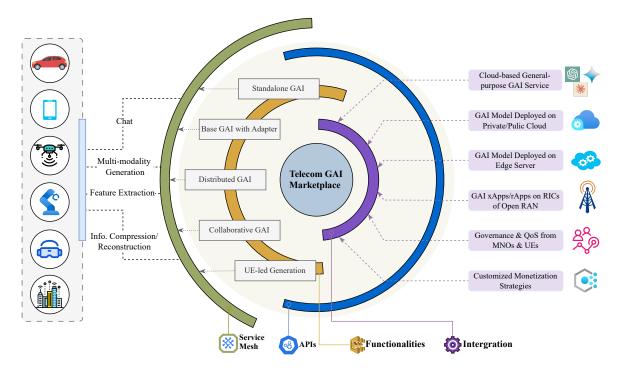


Fig. 1. Illustration of our conceptual telecom marketplace platform for GAI model integration and service delivery.

infrastructure, capable of processing data at different network segments to ensure minimal latency and high reliability, from prompt transmission to content generation.

V. Marketplace Solution for GAI and Open Network Integration

A dedicated marketplace provides a compelling solution for MNOs to integrate GAI with open networks. This marketplace would serve as a private platform for executing different types of GAI models, allowing services to be accessed, deployed, and monetized in ways tailored to telecoms operators' needs.

A. Telecoms Marketplace Design Principles

- Standalone Model Support: The platform should host and deploy independent, task-specific GAI models. With automated pipelines, it enables rapid deployment and updates, which is ideal for specific AI functionalities that operate in isolation.
- Base Model with Adapter Support: The telecoms marketplace should support adapter techniques, such as LoRA, to facilitate parameter-efficient fine-tuning (PEFT) of GAI models. This allows MNOs to customize base models for specific tasks or network environments without retraining the entire model. By selectively adjusting a small subset of parameters, MNOs can efficiently tailor AI models for different applications and services.
- Distributed Model Support: The marketplace should support distributed GAI models, enabling deployment and management across network nodes such as the cloud, RIC, and edge servers to meet different requirements in latency and processing time. It should provide an orchestration framework for managing these models, facilitating coordination and collaboration among GAI agents.

- Collaborative Model Support: The platform should enhance resource utilization and scalability by supporting collaboration among GAI models. Key collaborative modes include:
- Mixture of Experts (MoE): Divides a GAI model into specialized sub-networks ("experts"), each focused on a specific subset of the input data, working together to complete tasks. The marketplace platform should enable edge computing nodes within the open network to host MoE training and inference, maximizing the potential of distributed computing resources.
- 2) Hybrid Edge-Cloud GAI Inference: Essential for multicloud and hierarchical edge setups in open networks, this mode enables collaborative inference by partitioning models across cloud and edge environments. The marketplace should support not only the deployment of these partitioned models but also token-level interactions between the cloud and edge, ensuring seamless and efficient inference across the network.
- Service Mesh: A service mesh ensures efficient and secure operation of AI models in standalone, distributed, and collaborative environments. It manages communication between model components using proxies like Envoy, provides intelligent traffic management, enhances security with mutual transport layer security (TLS) encryption, and offers real-time observability of AI services.
 - Service mesh supports the complex interactions needed for collaborative AI models, such as MoE and hybrid edge-cloud inference, ensuring that the marketplace can deliver scalable, high-performance, and secure AI services across telecoms networks.
- UE-led Generation: Focuses on content generation driven by user equipment (UE), requiring low-latency responses and real-time processing. The marketplace should provide

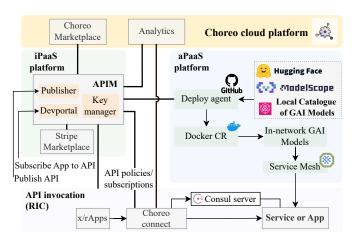


Fig. 2. Overall marketplace integration architecture with API invocation. tools for deploying AI models directly on UEs and managing billing, catering to applications like augmented reality, gaming, and personalized content delivery.

B. Marketplace Implementation

The marketplace implementation builds on the strategies outlined in [11], 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:

- 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 finegrained access control and monitoring, supporting various business models like pay-per-use, subscriptions, and service level agreements.
- 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 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

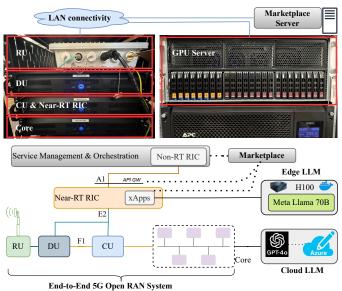


Fig. 3. Illustration of the experimental setup.

platforms Stripe, allowing for the provision of optimized, ready-to-use services for specific application use cases. This is particularly valuable for non-developer users who prefer not to handle selection, evaluation, testing, and service integration themselves.

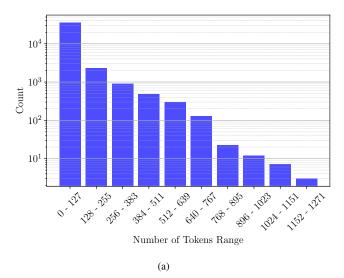
GAI services and product bundles are deployed using YAML scripts from GitHub and containers from Docker Hub, Hugging Face, or other repositories. The Kubernetes-based YAML scripts used by the deployment agents are designed for low complexity and flexibility, featuring annotations for automated injection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

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.

VI. A CASE STUDY OF IN-NETWORK GAI DEPLOYMENT THROUGH MARKETPLACE

The telecoms marketplace provides significant advantages for integrating open networks with GAI models, offering enhanced monetization and management capabilities for MNOs, and improved service quality compared to cloud-based, general-purpose GAI services.

In this section, we present a case study where a local LLM model is deployed in an Open RAN system via the developed marketplace platform. This model takes task-agnostic prompts and generates corresponding content. We compare its performance with a cloud-based LLM model, focusing on content generation latency.



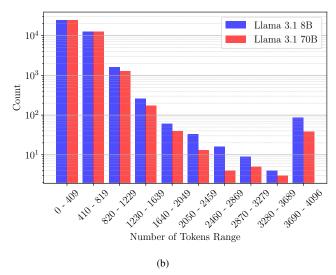


Fig. 4. Histograms of Input and Output token lengths for the Chatbot arena dataset with the Llama 3.1 8B and 70B models, wherein (a) indicates Log-scaled Histogram of Input Tokens, (b) show Log-scaled Histogram of Output Tokens for Llama 3.1 Models.

A. Experimental Setup

The experiment has been conducted using an advanced end-to-end Open RAN testbed developed under the BEACON-5G project [12]. The setup, illustrated in Fig. 3, included an Open RAN system, and the LLM service agent which is responsible for handling requests and running GAI models. The LLM agent was deployed on an edge server linked to the Centralized Unit (CU) of the Open RAN. The management of the marketplace, including deployment and API billing components, was handled by the Non-Real-Time RIC (Non-RT RIC).

1) Hardware and Software Specifications: The edge server is equipped with two Intel Xeon Platinum 8480+ CPUs (shared access) and two Nvidia H100 PCIe GPUs, each with 80 GB VRAM. It hosts the vLLM agent [13]. LLM Models are deployed in a vLLM container (one at a time) for high-performance inference compatible with OpenAI APIs: Meta Llama 3.1 8B [14] is used for fast, high-throughput processing of latency-sensitive tasks with moderate reasoning needs. Meta Llama 3.1 70B is deployed for medium-throughput processing of complex reasoning tasks to evaluate edge performance.

For the cloud deployment, the following models were deployed to Azure Cloud Cognitive Services: **GPT-3.5 Turbo** is deployed to an inference endpoint to serve as the cloud-based LLM for simple and quick tasks, equivalent in reasoning performance to the Llama 3.1 8B local model. **GPT40** is deployed to an inference endpoint to serve as the cloud-based LLM for complex tasks, slightly superior in reasoning performance to the Llama 3.1 70B local model.

2) Load Simulation: To simulate a realistic inference environment, a constant background load was applied to the edge inference server using user-generated, chat-based openresponse requests from the Chatbot Arena Dataset [15]. The system generated an average of 10 requests per second for the Llama 3.1 8B model and 3 requests per second for the Llama

3.1 70B model, with request intervals following an exponential distribution to simulate a Poisson process.

This setup resulted in an average of 44 concurrent background requests for the 8B model and 42 for the 70B model. The average input prompt size was 85 tokens, with the 8B model generating an average of 351 tokens per request and the 70B model generating 327 tokens on average.

As shown in Fig. 4a, the distribution of input token lengths from the Chatbot Arena dataset is concentrated at lower token counts. In contrast, Fig. 4b generally follows a similar trend but has a slightly higher concentration in 3690-4096 range due to the maximum token limit of 4096, which leads to output truncation and accumulation in the final bin. The 8B model also tends to generate longer responses, though this behavior was not specifically analyzed in this study.

3) Request Configuration: 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.

B. 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.

1) Measurement Metrics: Time to First Token (TFT) measures the time from when a request is sent to the API server until the first token is received by the UE. This metric is crucial for applications requiring quick response times, such as interactive systems and real-time communications.

Inter-Token Time (ITT) measures the interval between successive tokens in the response stream. Consistent ITT is essential for applications that rely on continuous data streams, ensuring stable output rates.

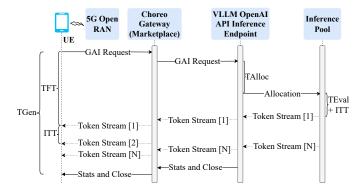


Fig. 5. Illustration of the process of GAI content generation over marketplace platform and Open RAN testbed and time measurements, wherein TFT and ITT are the key metrics for comparison.

The GAI content generation process and timing measurements are illustrated in Fig. 5.

2) Experimental Procedure: The experiment was designed with several key considerations. First, 100 identical requests were sent sequentially to both edge-based and cloud-based LLMs. The responses were streamed back to the UE, and the TFT and ITT were recorded for each token generated. To simulate real-world conditions, a continuous background load was maintained on the edge server. The TFT and ITT metrics were averaged over multiple runs, and the variations were analyzed to assess the consistency and reliability of both deployment strategies.

C. Results

The comparison focused on cloud-based versus edge-based LLM deployments through the marketplace, highlighting two key scenarios:

1) GPT-3.5 Turbo vs. Llama 3.1 8B: As illustrated in Fig. 6a, the edge-based Llama 3.1 8B model outperformed the cloud-based GPT-3.5 Turbo on TFT whilst being slightly worse on ITT. The reason for this is the harsher synthetic load that we have set for the smaller 8B model assuming a higher usage. The Llama 3.1 8B model demonstrated lower TFT, providing faster initial responses, which is critical for time-sensitive applications like real-time interactive systems. Additionally, whilst the edge deployment exhibited inferior ITT, it had a reduced variability, making it well-suited for tasks requiring quick and reliable outputs, such as short text generation and classification.

2) GPT-40 vs. Llama 3.1 70B: Fig. 6b presents a more nuanced comparison between the GPT-40 and Llama 3.1 70B models, showcasing the trade-offs between edge and cloud deployments. Similar to the smaller models, the edge-based Llama 3.1 70B exhibited a lower TFT, making it advantageous for quick, short responses, which is ideal for applications such as classification and short-form content generation.

On the other hand, the cloud-based GPT-40 excelled in ITT, delivering faster and more consistent token generation for longer prompt completions. This makes it a better choice for tasks involving complex and lengthy content generation, where sustained output is more important than initial response speed.

These findings underscore the importance of selecting the appropriate deployment strategy based on the specific requirements of the application, balancing the trade-offs between initial response speed and sustained generation performance. In this process, the GAI marketplace functions as a GAI service gateway, responsible for selecting, executing, and billing the appropriate GAI model. This process can be automated by the given application scenarios.

VII. DISCUSSIONS

A. Monetization Strategy

The marketplace enables flexible monetization strategies through API-based mediation, traffic management, and access control, accommodating various charging models to meet different service levels and user needs. For GAI models deployed within the network, in addition to standard billing based on service levels and API usage, the marketplace could provide a more granular and customized billing approach. This can be done by considering the application scenarios, inspecting data packets returned to clients, considering input and output tokens, and model configurations derived from client requests.

B. Service Access and Regulation

The telecoms GAI marketplace presents a complex yet promising framework for integrating GAI services into telecom networks. It facilitates the deployment of AI solutions directly within the network infrastructure, enabling dynamic scaling and updates based on demand and performance. This flexibility allows MNOs to adapt quickly to changing market needs and customer expectations.

The unified API-centric platform simplifies the integration of diverse GAI models and services, ensuring compatibility across network components and reducing the complexity of third-party applications.

However, integrating GAI services also presents regulatory challenges, particularly concerning data privacy, security, and compliance with regional and international standards. It is crucial to address these requirements carefully, ensuring all GAI models adhere to ethical guidelines and protect user data.

C. Service Management

Managing the lifecycle of GAI services in the marketplace requires careful planning and robust strategies. This includes monitoring performance, handling updates and patches, ensuring service continuity, and optimizing resource allocation. Effective management is essential for maintaining reliable and efficient GAI services, which directly affects customer satisfaction and operational costs. For MNOs, balancing these technical requirements with strategic goals like cost management, service innovation, and market competitiveness is vital to fully leveraging the telecoms GAI marketplace.

D. Open Service Platform for Marketplace

Open networks require the collaboration of multiple operators and service providers. In this context, the open service platform (OSP) is an option for coordinating network

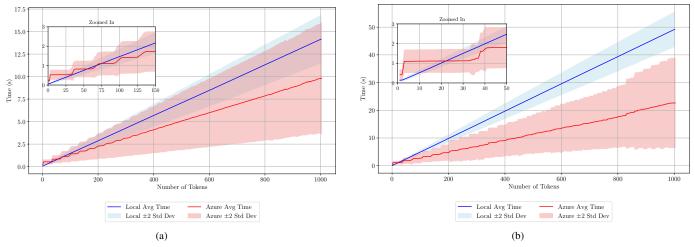


Fig. 6. Token timing comparisons for different models and deployment environments. (a) shows token timing comparison between cloud-based GPT-3.5 Turbo and edge-based Llama 3.1 8B models; (b) shows token timing comparison between cloud-based GPT-40 and edge-based Llama 3.1 70B models.

operations and GAI services. Integrating the OSP with the private marketplace provides a unified access point for service provisioning, enabling dynamic and flexible service offerings that can quickly respond to customer demands and operational needs.

Aligning the OSP with marketplace mechanisms allows customers to easily discover, purchase, and configure services across different networks. The OSP also plays a key role in ensuring fairness by balancing GAI services from various providers, enabling comparisons, and detecting anticompetitive pricing strategies. This functions like a price comparison tool but with enhanced monitoring and evaluation of integrated services.

The combination of the marketplace and OSP can enhance collaboration among service providers. This allows operators to expand their service offerings without significant infrastructure investments, enabling scalable and flexible deployment of GAI services across diverse network environments.

VIII. CONCLUSIONS

This article described the functionality of the telecoms GAI marketplace platform for integrating and monetizing GAI models from the MNOs' perspective. We designed and implemented a novel API-centric marketplace for deploying, managing, and monetizing diverse GAI services directly within the 6G open network. Our experimental results from an end-to-end Open RAN testbed validate the real-world benefits of this approach, particularly in reducing latency for LLM models deployed on local edge servers compared to cloudbased general-purpose LLMs. Furthermore, the discussion on monetization strategy, regulatory, management, and service platforms provides guidance for MNOs to navigate the complexities associated with implementing GAI in future open networks. We believe that the marketplace strategy presented in this article can help leverage GAI to enhance network capabilities and advance GAI services within 6G networks.

REFERENCES

- "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] A. Maatouk et al., "Large Language Models for Telecom: Forthcoming Impact on the Industry," IEEE Commun. Mag., 2024.
- [4] 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.
- [5] X. Meng et al., "Generative Pretrained Transformer for Network Traffic," arXiv preprint arXiv:2304.09513, 2023.
- [6] Y. Shen et al., "Large Language Models Empowered Autonomous Edge AI for Connected Intelligence," IEEE Commun. Mag., 2024.
- [7] Y. Yang et al., "6G Network AI Architecture for Everyone-Centric Customized Services," IEEE Netw., vol. 37, no. 5, pp. 71–80, 2022.
- [8] L. Zhao et al., "Open-Source Edge AI for 6G Wireless Networks," IEEE Netw., 2024.
- [9] P. Li and A. Aijaz, "Open RAN meets Semantic Communications: A Synergistic Match for Open, Intelligent, and Knowledge-Driven 6G," in Proc. of IEEE CSCN. IEEE, 2023, pp. 87–93.
- [10] F. Jiang et al., "Large AI Model-Based Semantic Communications," IEEE Wirel. Commun., vol. 31, no. 3, pp. 68–75, 2024.
- [11] T. Farnham et al., "Demo: Integration of Marketplace for the 5G Open RAN Ecosystem," in Proc. of IEEE ICNP, 2023, pp. 1–2.
- [12] 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.
- [13] W. Kwon et al., "Efficient Memory Management for Large Language Model Serving with PagedAttention," 2023. [Online]. Available: https://arxiv.org/abs/2309.06180
- [14] A. Dubey et al., "The Llama 3 Herd of Models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783
- [15] L. Zheng et al., "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena." 2023.