Telecom Foundation Models: Applications, Challenges, and Future Trends

Tahar Zanouda, Meysam Masoudi, Fitsum Gaim Gebre, Mischa Dohler Ericsson

{tahar.zanouda, meysam.masoudi, fitsum.gaim.gebre, mischa.dohler}@ericsson.com

Abstract—Telecom networks are becoming increasingly complex, with diversified deployment scenarios, multi-standards, and multi-vendor support. The intricate nature of the telecom network ecosystem presents challenges to effectively manage, operate, and optimize networks. To address these hurdles, Artificial Intelligence (AI) has been widely adopted to solve different tasks in telecom networks. However, these conventional AI models are often designed for specific tasks, rely on extensive and costlyto-collect labeled data that require specialized telecom expertise for development and maintenance. The AI models usually fail to generalize and support diverse deployment scenarios and applications. In contrast, Foundation Models (FMs) show effective generalization capabilities in various domains in language, vision, and decision-making tasks. FMs can be trained on multiple data modalities generated from the telecom ecosystem and leverage specialized domain knowledge. Moreover, FMs can be fine-tuned to solve numerous specialized tasks with minimal task-specific labeled data and, in some instances, are able to leverage context to solve previously unseen problems. At the dawn of 6G, this paper investigates the potential opportunities of using FMs to shape the future of telecom technologies and standards. In particular, the paper outlines a conceptual process for developing Telecom FMs (TFMs) and discusses emerging opportunities for orchestrating specialized TFMs for network configuration, operation, and maintenance. Finally, the paper discusses the limitations and challenges of developing and deploying TFMs.

Index Terms—Foundation Models, Foundation Models for Telecom, Wireless Networks, Artificial Intelligence, Generative Artificial Intelligence, AI for Telecom, Large Models for Telecom, TelcoAI.

I. INTRODUCTION

5G-and-beyond networks offer immense opportunities across various industries, enabling limitless connectivity and numerous emerging use cases. However, the increasing complexity of mobile networks hinders future developments, with deployment scenarios ranging from centralized to virtualized and physical implementations, multi-vendor support, and open ecosystem. To address these impediments, AI has been adopted in telecom to enhance automation and reduce manual tasks in network operations, administration, and maintenance, leading to the inception of *TelcoAI - AI for Telecom*.

As we move towards 6G [1], AI/ML is expected to be a foundational technology, driving AI-centric networks and building on the extensive use of AI in current networks throughout their design, deployment, and operational stages. The growing importance of AI in telecom is portrayed by the concept of "AI-native Telecom", where AI is at the core of network functionalities. Recently, AI has been applied to various tasks, including ML-based modulation schemes, radio

resource management (e.g., power control, scheduling, link adaptation), UE localization, mobility optimization, network slicing, network configuration, and wireless security. Graph Neural Networks (GNNs) incorporate spatio-topological dynamics in the network. Reinforcement Learning (RL) leverages feedback through exploration-exploitation learning techniques using online and offline RL architectures. Transfer Learning addresses data scarcity by transferring knowledge from related tasks. Similarly, label-efficient learning and few-shot learning mitigate data scarcity and labeling challenges.

Given data sensitivity and limited resources, distributed ML has become integral to *TelcoAI*. Federated Learning allows remote clients to collaboratively train models in a privacy-preserving manner, using isolated data and predefined aggregation strategies to update local models. Split Learning decouples models into parts trained on local data, with outputs fused into a central model on the server side. These advancements highlight the potential of AI and foundation models in telecom, paving the way for more sophisticated models to harness the full potential of 6G and beyond [2].

However, several challenges arise when applying AI/ML in real-world telecom applications, such as limited ability to generalize, inability to capture network complexities, necessity to train models on data generated by network simulators, among others. To this end, Foundational Models (FMs) can be enablers for autonomous telecom networks [3], eliminating the necessity of task-based AI models while addressing the current limitations.

FMs are large models trained on massive and heterogeneous datasets to solve a multitude of downstream tasks [4], [5]. They are versatile and can be deployed for numerous use cases in telecom and other domains. FMs can be categorized based on data modality and learning process as follows:

- Large Language Models (LLMs) are large models trained on a large corpus of textual data. Models such as LLaMA, PaLM, GPT-4, Falcon, and Mistral, among others, showed impressive results in the range of tasks. LLMs can also be fine-tuned with domain-specific data to align the performance of the LLMs to a specific domain such as Telecom [5]. LLMs have been used for different tasks in SW engineering, such as building chatbot assistants, code generation, and maintenance ticket resolution [6].
- Large Vision Models are large models designed to solve several computer vision tasks. Models such as ViT, Meta's DINO, Meta's SAM, and Florence showed

- promising results. Other models such as Prithvi were trained on geospatial and satellite imagery data.
- *Time-Series Models* are large models designed to train a unified model for time-series. Models such as LLM-Time, Lag-LLAMA, LLM4TS, and LLMTime are early efforts in this area.
- Multimodal FMs are large models designed to deal with multimodal data that refer to the ability of a model to accept different input modalities, e.g., images, texts, or audio signals. OpenAI's CLIP and DALL-E, BLIP, FILIP, PaLM, LLaVA and GPT-40 are examples of such models.
- RL FMs are large models that combine FMs and sequential decision-making to tackle complex real-world problems with better generalization. Recently, attempts have been made to build upon decision transformers or introduce tailored architectures, such as DeepMinds's Gato to achieve multi-modal, multi-task, and multi-embodiment objectives such as playing Atari, captioning images, chatting, etc.

This paper delves into the development of Telecom FMs (TFMs) and briefly summarizes the data types and modalities in the telecom ecosystem. The article describes FM use cases in telecom and sheds light on the challenges and risks of deploying TFMs.

II. AI/ML STANDARDS & ALLIANCES FOR RAN

Integrating AI into 5G and beyond networks requires developing an aligned view through standardization bodies to develop AI technology standards. Multiple alliances have been established by partnering with technology industry leaders and academic institutions to enhance RAN performance and capabilities using AI [7]. These alliances aim to capitalize on recent generative AI capabilities and accelerate the adoption of *GenAI* in the telecom sector to find new growth opportunities.

- AI RAN Alliance a collaborative initiative to develop AIdriven solutions to achieve an AI-native RAN.
- Global Telco AI Alliance (GTAA) a collaboration between key telecom operators to advance AI use cases reshaping the telecom landscape.
- Alliance for Telecommunications Industry Solutions (ATIS) ATIS-TOPS Council established a working group to evaluate GenAI/ML use cases for future networks, examining how it contributes to improving the efficiency across Telecom sectors.

Standards defining organizations (SDOs) aim to standardize the application of AI in telecom, paving the way for more efficient, secure, and intelligent network operations. Such initiatives are as follows:

- O-RAN is investigating cross-domain AI and Generative AI use cases in Open RAN architecture. O-RAN specifications are being adopted by the European Telecommunication Standardization Institute (ETSI) and published among ETSI standards.
- ETSI established a new AI Agent Core Network working group to discuss AI for next-generation telecom technologies. Key focus areas include Orchestration and management of network operations, network knowledge

- management, intelligent customer service applications, LLMs Lifecycle Management, securities, etc.
- 3GPP is now integrating AI into RAN. For instance, TS 28.105 (Rel. 17) and TR 28.908 (Rel. 18) outlines AI/ML integration in 5G systems. Ongoing discussions focus on leveraging AI for enhancements in CSI, beam management, and positioning, highlighting 3GPP's dedication to the responsible integration of AI in wireless communications.
- International Telecommunication Union (ITU with two groups (1) ITU-T to set global ICT standards, and (2) ITU-R to advance radio technologies with AI integration. ITU-T: Develops global ICT standards focusing on internet protocols, IoT, and next-generation networks (Yseries). Supplement 72 to the Y.3000-series details AI integration in ICT, featuring key standards like Y.3115 for AI-enabled cross-domain network requirements and Y.3325 for an AI-based management framework. ITU-R enhances radio technologies through AI integration, as outlined in Document M.2242. This document emphasizes Cognitive Radio Systems (CRS) with capabilities such as environment awareness, autonomous parameter adjustment, and adaptive learning.

Subsequently, we highlight opportunities and challenges for these AI-centric SDOs and Alliances to adopt TFMs.

III. TELECOM DATA ECOSYSTEM

Telecom networks generate multimodal data from spatially distributed radio nodes, capturing temporal dynamics originating from different software (SW) and hardware (HW) modules and storing them in various file formats. Figure 1 depicts a simplified overview of the telecom data ecosystem that spans phases of HW manufacturing, SW development, SW/HW testing, product knowledge management, network deployment, and network troubleshooting and healing. Telecom networks consist of a set of heterogeneous interconnected radio nodes that are distributed in a geographical region to provide connectivity services. Radio nodes encompass SW and HW components, each providing different functionality. The HW components are configured based on configuration files, and SW components cover the implementations of multiple technology standards. System configuration for both SW and HW components are captured in radio node configuration data [8], which encompass information about telecom networks configuration, including Configuration Management (CM) parameters [8], SW features, licenses, and installation setup.

The telecom infrastructure consists of a wide variety of assets, e.g., power equipment, data centers, inventory warehouses, etc, captured in telecom infrastructure inventory and configuration data along with a diary of maintenance activities such as cleaning filters, inspecting backup batteries, etc. Operators monitor the network continuously to assess the behavior of the network using Performance Management (PM) data [8] to gauge network performance. PM data is captured regularly across radio nodes in different SW and HW components.

Operators benchmark the behavior of the network and its SW using Key Performance Indicators (KPIs), which are

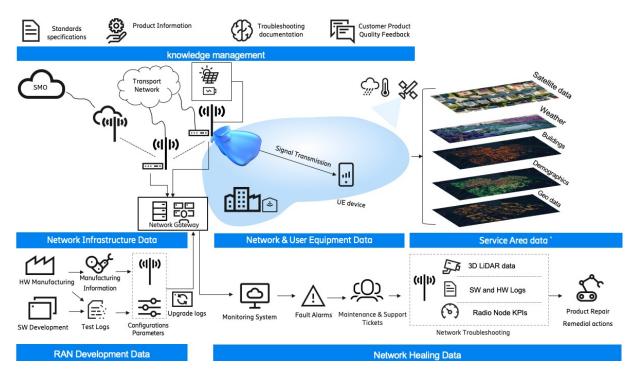


Fig. 1. The Overview of Telecom Network and Ecosystem Data.

standardized by *3GPP*. KPIs are a set of formulas to calculate performance indicators using PM and CM data. Different categories of *KPI*s exist, such as accessibility, mobility, integrity, utilization, and energy performance. This *KPI*s behavior can vary depending on service area characteristics. SW components use such *KPI*s along with other mechanisms to generate alerts for any aberrant behavior in the network. The format of the alerts is standardized by *3GPP* and known as fault management data [8].

Operators monitor network traffic usage and activities through different data sources (e.g., service usage, subscription, call records, etc). The data that characterizes the interaction between users and the network can be captured in several ways. For instance, minimization of drive test measurements, standardized by 3GPP [8], consist of user information, field measurements, radio measurements, and location information.

Developing SW-intensive HW products in the telecom typically involves HW manufacturing, SW development, HW/SW integration, testing, and field trials. Network products implement standardization guidelines that outline globally aligned solutions, functional frameworks, use cases, and test specifications captured in standardization text documents [8]. RAN HW is typically tested with simulations, and when the HW design is proven to be functional, the units are produced, calibrated, and tested in factories. The process of HW testing generates HW test log files that capture the output of such activities in the factory. During these activities, manufacturing information data is stored to ensure the reliability of HW products across the different stages.

RAN SW comprises several large components referred to as SW modules. Each module is developed and maintained by one or several cross-functional teams. SW modules capture logs, events, and internal counters to enable instant feedback and alerts for any aberrant behavior in the system. Log files are a text-based function-related history of events that describe software state during its execution. Each line of log files indicates a different event and may hold various types of information such as function name, timestamp, and log message. Other types of logs, such as SW logs, HW logs, and traces, are typically used for debugging and capturing low-level events. A trace can span multiple functions and be tracked using unique references. SW internal counters, on the other hand, are time-dependent data points to monitor SW quality. Unlike standardized data, no common formatting schemes exist for different vendors to convey similar information for SW logs and counters. When a SW component is deployed in the cloud, cloud resource utilization data, referred to as telemetry data, are stored for elastic and dynamic resource management.

Before rolling out any SW release, each SW typically undergoes rigorous testing and review in a simulator. The process involves creating many test cases managed in test case specification documents, each crafted to validate dedicated system features. Testing SW produces SW test log files.

After rolling out SW/HW products in the field, ensuring the product's reliability is crucial. The reporting, analyzing, and resolving HW and SW faults are essential to providing stable, high-quality products. When a radio node experiences a problem/incident in an operational network, the operator raises a troubleshooting report or maintenance request as a textual document. The trouble report contains problem observations, system logs, and an answer section for resolved reports [9].

The diversity of data generated from Telecom ecosystem presents several challenges, including but not limited to data multimodality with large diversity in size, granularity, etc;

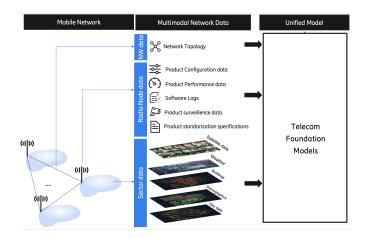


Fig. 2. Conceptual Telecom Foundation Model Architecture.

quickly outdated data quality; data scarcity for rare but impactful events; dynamics of data feedback loops; etc.

IV. TELECOM FOUNDATION MODELS

A. Developing Telecom Foundation Model

Telecom Foundation Models (TFMs) are domain-specific FMs trained on extensive data, spanning multiple sub-domains and modalities, and tailored for telecom applications. Here, we present the necessary components and pipelines for developing a TFM in telecom, which starts by assembling a dataset using multiple and multi-modal data sources.

Telecom data comes in different temporal, visual, and spatial modalities to build a holistic overview of how telecom networks function jointly. The next step is data fusion, which refers to integrating multiple data sources from various data sets to generate more accurate, comprehensive, and useful information. Telecom data may have multiple time granularity, asynchronous reporting times, and multi-modality.

The training phase of the model requires designing an architecture capable of capturing a comprehensive representation of telecom networks. This architecture conceptualizes telecom data as *context-aware multimodal heterogeneous graphs*. The model incorporates three key components:

- Radio Node Component: it is conceptualized as a multilevel graph that characterizes software execution, building upon previous work [10], [11], using (i) software performance time-series data, (ii) textual log messages, and (iii) configuration parameters. Each radio node is further characterized by manufacturing information to understand how sensors and products are jointly affected. In Fig. 2, we show different data types that can be extracted and used in training the model from each radio node.
- Network Component: it is conceptualized as a multimodal graph that characterizes interactions between different RAN nodes, building upon previous efforts [11], [10], using (i) RAN nodes relationships which indicate the relation between each pair of network nodes. The relations may be any one or a combination of a (1) mobility relation, and/or a (2) transport relation between pair of network nodes, and (ii) service area data. This

data captures the dynamics of the surrounding area, including geographical (e.g., satellite imagery), socio-demographic (e.g., population density), and economic factors in coverage areas. In Fig. 2, we use the network topology to extract the adjacency matrix that is crucial to train GNN-based architectures.

• Network Development Journey Component: it is conceptualized as a time-dependent set that can be collected through various network stages: (i) product development (e.g., product guidelines, test logs), (ii) network optimization (e.g., expert rules), (iii) network evolution (e.g., upgrade actions), and (iv) network healing (e.g., maintenance tickets). Such data is typically used when fine-tuning models for specialized tasks, building upon previous efforts [9], [11], [12].

A global model is initially trained on general telecom knowledge but is not fine-tuned for specific applications. In the downstream fine-tuning phase, this global model is leveraged in various telecom applications such as network optimization. During downstream tasks, transfer learning or fine-tuning becomes crucial to adapt the model to specific tasks while inheriting the general telecom knowledge from the global model.

B. Specialized Telecom Foundation Models

The general TFM, trained with telecom knowledge, can be tailored to different domains. The TFM architecture is illustrated in Figure 2.

Telecom applications are diverse, presenting unique limitations, tasks, and complexities. However, they all rely on core telecom domain knowledge. As a result, a general TFM can be adapted for specific tasks through different techniques.

To build a specialized model, one can consider the data modality and the use case's nature. These techniques can be broadly grouped into two classes:

- Domain Adaptation of Models, which involves training models.
 - Pre-training new models from scratch.
 - Continued pre-training of existing models.
 - Domain-specific fine-tuning/Instruction tuning of existing models, which involves adjusting the TFM's parameters to generate and optimize outputs for particular applications.
- In-context Learning and Knowledge Augmentation of existing models, which does not involve training models. The approaches are applicable to both general and domain-specific models.
 - Prompt engineering, which uses techniques such as Zero-Shot, One-Shot, and Few-Shot Learning.
 - Retrieval Augmented Generation (RAG) and Graph Retrieval Augmented Generation (Graph-RAG).

These techniques are mostly used in tandem with the FMs. Other techniques help reduce the cost of model fine-tuning and management. Several techniques have emerged recently. For instance, Low-rank adaptation (LoRA) significantly reduces

large models' trainable parameters and GPU memory requirements. LoRA enhances speed, reduces compute resource demands and cost, and improves memory efficiency by allowing parameters to be cached in memory instead of relying on slower disk reads.

Telecom has many applications, and each specialized TFM is tailored to specific domains that encompass similar tasks and require similar inputs. However, fine-tuning a TFM is a resource-intensive and costly process. Therefore, specialized TFMs are designed to cover categorized tasks efficiently. Domain-specific datasets are used to fine-tune a general model, resulting in multiple specialized TFMs customized for applications within specific subdomains. Data types, time granularity, and task similarities determine the categorization of these subdomains. This approach ensures that the specialized TFMs are well-suited for their respective applications' specific requirements and nuances. Figure 3 demonstrates fine-tuning a general TFM.

Specialized TFMs can be deployed in the central nodes or at the edge. Specialized TFMs can have different deployment requirements due to task description, HW, architectural limitations, required resources, latency requirement, etc. Due to impediments such as privacy, storage, and resource limitation, having one central model is always challenging. One possible solution to mitigate such challenges is to use recent distributed machine learning techniques such as Split Learning. The overall architecture of TFM deployment, including data sources, application areas, and the orchestrator, is depicted in Figure 4. Since there are multiple specialized TFMs for different tasks and they might be dependent on each other outputs, there is a need to add Orchestration Layer to the TFMs architecture design to ensure seamless network operation, provide communication flexibility, and mitigate conflicts. The orchestrator layer interconnects the TFMs and arranges the TFMs' tasks in a well-ordered, syncronized, and optimized pattern so that they function through accurate and automated repeated processes.

V. TFM FOR TELECOM APPLICATIONS AND STANDARDIZATION FRAMEWORKS

This section explores the potential of using TFMs in recent telecom standardization efforts.

A. Intent-Based Networking

The expanding scope of 5G and its diverse applications and requirements pose new challenges to telecom operations, especially in RAN. In this context, business intent and automated operations guarantee enterprise resilience. In telecom, "intent" signifies the predefined objectives or desired behaviors expected from customers. An intent describes "what" needs to be achieved without identifying "how" to achieve them. Intent must be quantifiable from network data to measure and evaluate the fulfillment result. Intent-based networks revolutionize network operations by autonomously interpreting these business intents—such as improving network quality—into actions. By automating the process of integrating customer/provider/operator intent, translating it,

and activating network features, these networks continuously align with business goals without human intervention. Using TFM, business intents are processed into service-level intents and translated into network KPIs. The FMs use these translated KPIs to propose and trigger actions in the network. These actions primarily affect the RAN, which bridges end users and business owners. Technological advancements in RAN to support diverse QoS requirements, such as high throughput and low latency, have increased the network's complexity.

Any action in the network may have an instantaneous or long-term impact on the network performance. Therefore, there is a need to assess the action and avoid the risk of performance degradation. Digital Twin (DT) [13] emerges as a potential solution to bridge this gap. DT serves as a virtual representation of the physical telecom network that enables operators to experiment with new techniques and configurations without risking the physical network infrastructure. Creating DTs accurately models network functions and ensures efficient synchronization between physical and digital entities [14].

The proposed actions by TFMs to adjust the network for satisfying intents can be assessed in DTs and then applied in the network. Moreover, TFMs can facilitate remote collaboration among distributed DTs and synchronize them with networks. Figure 5 illustrates the interaction between DT and intent-based networking.

B. Network Optimization

The continuous evolution of the network, as well as technological advancements like O-RAN, necessitates end-to-end network monitoring. Network monitoring allows interaction between network segments and real-time assessment of intent metrics like latency and throughput. TFMs enhance this monitoring by proactively identifying congestion points and inefficient resource usage, enabling optimized resource allocation for improved network efficiency. This information can be used to demonstrate intent satisfaction and network optimization applications such as slicing, energy saving, traffic steering, etc. Network optimization is a complex process that requires interactions among network segments. End-to-end optimization depends on models outlining interactions among multiple features. These models require skilled engineers possessing extensive domain expertise to identify pertinent aspects. Operators can benefit from using TFM-assisted DT to enhance this process and interact between network segments and parameter configurations [11].

C. Network Slicing

Network slicing empowers operators to optimize resource utilization and deliver diverse, scalable, differentiated services. A slice refers to dedicated end-to-end network resources. Each slice represents a virtualized and independent logical network tailored to specific intents, operating on the same physical network infrastructure [15]. According to Figure 5, the business intent is broken down into different intents. This splitting can be vertical in network architecture, meaning that each network segment has its intent to satisfy. However, the intent can be split into horizontal domains, meaning each service has its

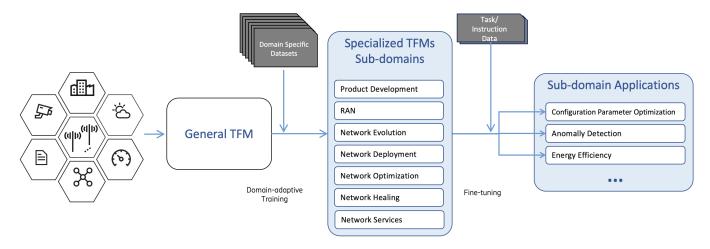


Fig. 3. Building specialized Telecom foundation models for each area

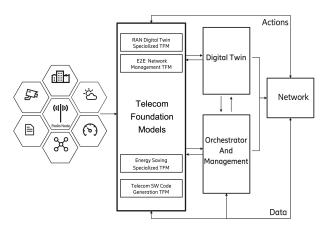


Fig. 4. Specialized TFMs Architecture and Orchestration

intent. Network slicing should consider both intents. Moreover, it needs to consider activated features in the network while facilitating slice resource allocation. In this context, TFM can collect the intents of all slices, activate network features along with network details, and recommend the necessary end-to-end resources for each slice. TFM passes the dedicated resources information to the DT to assess whether dedicated resources are insufficient, under-utilized, or adequate. If the dedicated resources are inadequate or under-utilized, TFM can submit a slice reconfiguration request. Consequently, TFMs offer slicing strategies that ensure intent feasibility and satisfaction per slice, prolonging the slice life cycle, enhancing resource utilization, and reducing energy consumption.

D. Network Healing

The complexities of deploying heterogeneous and dense telecom networks present a challenge to ensuring reliability, effectively finding optimal configuration parameters for resilient networks, and easily finding faults. TFMs promise to achieve self-healing networks capable of autonomously local-

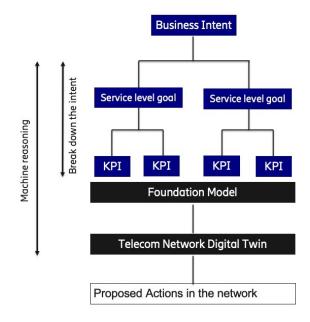


Fig. 5. DT-assisted intent-based networking with TFMs

izing faults, performing self-maintenance, and self-organizing without human intervention.

Troubleshooting efforts [10] can benefit from TFM-assisted DT to run tests [12] in realistic setup. With the diverse deployment scenarios in Telecom networks and the large number of deployed products that exhibit a wide diversity, TFMs can help to overcome potential incidents, and autonomously perform remedial actions have the ability. The advntage of using TFMs originates from the ability to combine different data modalities to attain a holistic and context-aware understanding of network failures, enable better fault management, and achieve trustworthy and resilient networks. Moreover, troubleshooting efforts are often conducted in a distributed manner by multiple experts and multiple spoken languages to document issues around the

globe. Language models have shown promising results [9] to remove language barriers and democratize expert knowledge through better knowledge search and troubleshooting recommendations from solved incidents.

E. AI-powered Network APIs

Telecom infrastructure generates vast temporal and spatial digital traces through device interactions, offering a comprehensive insight into digital interactions. The design focuses on context awareness in telecom networks and aims to understand environmental, temporal, and situational factors impacting network behaviors. Leveraging this ubiquitous telecom data allows operators to explore new revenue streams. TFMs promise to unlock new ways to monetize Telecom Data through APIs. For instance, in smart city applications, historical realtime network performance data collected from transportation devices can provide invaluable insights into decision-making processes for autonomous vehicles. TFMs can utilize combined UE measurements and device characteristics to optimize route selection based on signal quality, further aiding in tasks like constructing maps, updating transportation routes, estimating traffic, and predicting travel times influenced by traffic conditions. Enabling other applications in other verticals (e.g., smart manufacturing, transportation sector, etc.) necessitates rethinking traditional telecom approaches, propelling us into a new era fueled by AI-as-services.

VI. FUTURE TRENDS AND OPEN ISSUES

Despite TFMs' potential to achieve AI-centric decisionmaking and network operations, several critical challenges remain unsolved. These challenges necessitate further investigation to realize the benefits of TFMs in the telecom industry fully.

A. Scalability and Efficiency

Training FMs is a time-intensive process that demands significant computational resources, high power consumption, and specialized HW/SW infrastructure for model training. Moreover, deploying FMs can pose substantial computational demands, particularly regarding inference speed and model size. Consequently, the deployment of such models in realtime scenarios is still limited. Solutions such as model compression techniques (e.g., pruning and quantization) and distributed machine learning techniques (e.g., split learning, federated learning) can address these challenges. Model compression can lead to a trade-off between speed and accuracy, and distributed techniques can introduce communication overhead and synchronization issues. Therefore, ongoing research and development are crucial to enhance these methods and develop new approaches to make FMs more efficient and practical for real-time deployment in telecom networks. Despite these strategies, achieving high-speed inference with comparable performance remains a significant challenge.

B. Transparency and Interpretability

The complexity and scale of FMs present a significant challenge when applying FMs to the telecom domain. FMs still struggle with randomness when generating output. Hence, the trustworthiness and explainability of TMFs remain a challenge that hinders adopting the long-term vision of TFMs-native framework for telecom and fostering trust in the operation. It is aligned with current standarization. For instance, 3GPP has an ambition to create AI/ML processes with full interoperability among all the components of a 5G RAN system.

C. Time-critical Applications

Reducing latency in TFMs is essential for real-time applications such as dynamic spectrum resource sharing, resource allocation, user associations, handover management, and carrier aggregation. These use cases require control loops with low latency inference time. This can be challenging for specialized TFMs and may limit their applicability. Hence, latency must be minimized to ensure efficient and timely actions. A key strategy is to keep processing close to the data and TFM deployment location, reducing the need for extensive data transfers between compute resources and storage. This approach enhances performance, lowers costs, and mitigates data security risks, making it critical for effectively deploying TFMs in RAN.

VII. CONCLUSION

This article explored the untapped potential of how TFMs can shape the future of mobile networks. In particular, we explored how TFMs can be effectively used to develop, upgrade, operate, and manage networks. Moreover, we identified key ingredients to design and train FMs given the wide telecom ecosystem to rethink our traditional engineering approaches.

TFMs present an immense opportunity to significantly shape telecom networks on their journey towards 6G and beyond, simplify network operations, enhance development productivity, promote sustainability, expand the value chain, and ultimately boost business profitability.

Through our discussions, it has become apparent that TFMs face challenges in terms of scalability in resource-constrained environments. Despite these hurdles, TFMs offer to build efficient, resilient, and trustworthy networks.

REFERENCES

- [1] M. Chafii *et al.*, "Twelve scientific challenges for 6G: Rethinking the foundations of communications theory," *IEEE Communications Surveys and Tutorials*, vol. 25, no. 2, pp. 868–904, 2023, publisher Copyright: © 1998-2012 IEEE.
- [2] J. Du et al., "Distributed foundation models for multi-modal learning in 6G wireless networks," *IEEE Wireless Communications*, vol. 31, no. 3, pp. 20–30, 2024.
- [3] R. Bommasani et al., "On the opportunities and risks of foundation models," ArXiv, 2021. [Online]. Available: https://crfm.stanford.edu/ assets/report.pdf
- [4] A. Mekrache et al., "Intent-based management of next-generation networks: an LLM-centric approach," IEEE Network, pp. 1–1, 2024.
- [5] L. Bariah et al., "Large generative AI models for telecom: The next big thing?" IEEE Communications Magazine, pp. 1–7, 2024.

- [6] A. Fan et al., "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). IEEE, 2023, pp. 31–53.
- [7] C. Chaccour *et al.*, "Telecom's artificial general intelligence (agi) vision: Beyond the genai frontier," *IEEE Network*, 2024.
- [8] X. Lin et al., "Embracing AI in 5G-advanced toward 6G: A joint 3GPP and O-RAN perspective," IEEE Communications Standards Magazine, vol. 7, no. 4, pp. 76–83, 2023.
- [9] N. Bosch et al., "Fine-tuning BERT-based language models for duplicate trouble report retrieval," in 2022 IEEE International Conference on Big Data (Big Data), 2022, pp. 4737–4745.
- [10] R. Bourgerie and T. Zanouda, "Fault detection in telecom networks using bi-level federated graph neural networks," in 2023 IEEE International Conference on Data Mining Workshops (ICDMW), 2023.
- [11] S. Piroti et al., "Mobile network configuration recommendation using deep generative graph neural network," IEEE Networking Letters, 2024.
- [12] M. Nabeel et al., "Test code generation for telecom software systems using two-stage generative model," in 2024 IEEE International Conference on Communications Workshops (ICC Workshops). IEEE, 2024.
- [13] M. Masoudi et al., "Digital twin assisted risk-aware sleep mode management using deep Q-networks," IEEE Transactions on Vehicular Technology, vol. 72, no. 1, pp. 1224–1239, 2023.
- [14] L. U. Khan et al., "Digital twin of wireless systems: Overview, taxonomy, challenges, and opportunities," IEEE Communications Surveys & Tutorials, 2022.
- [15] M. Masoudi et al., "Energy-optimal end-to-end network slicing in cloud-based architecture," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 574–592, 2022.