Large Generative AI Models for Telecom: The Next Big Thing?

Lina Bariah, Qiyang Zhao, Hang Zou, Yu Tian, Faouzi Bader, and Merouane Debbah

***Abstract*—The evolution of generative artificial intelligence (GenAI) constitutes a turning point in reshaping the future of technology in different aspects. Wireless networks in particular, with the blooming of self-evolving networks, represent a rich field for exploiting GenAI and reaping several benefits that can fundamentally change the way how wireless networks are designed and operated nowadays. To be specific, large GenAI models are envisioned to open up a new era of autonomous wireless networks, in which multi-modal GenAI models trained over various Telecom data, can be fine-tuned to perform several downstream tasks, eliminating the need for building and training dedicated AI models for each specific task and paving the way for the realization of artificial general intelligence (AGI)-empowered wireless networks. In this article, we aim to unfold the opportunities that can be reaped from integrating large GenAI models into the Telecom domain. In particular, we first highlight the applications of large GenAI models in future wireless networks, defining potential use-cases and revealing insights on the associated theoretical and practical challenges. Furthermore, we unveil how 6G can open up new opportunities through connecting multiple on-device large GenAI models, and hence, paves the way to the collective intelligence paradigm. Finally, we put a forward-looking vision on how large GenAI models will be the key to realize self-evolving networks.**

I. Introduction

While recent wireless networks have witnessed several revolutions in terms of technological trends, it is apparent that future wireless generations are converging towards solidifying the principle of self-built, self-evolving networks, particularly with the maturing of the self-organizing networks (SON) paradigm and the remarkable advancements in artificial intelligence (AI) technologies [1]. The core principle of SON is based on the concept of enabling wireless networks to have the capability to adjust, reconfigure, and optimize their functions and parameters according to particular network conditions and user demands. However, despite the fact that SON is aimed for realizing automation in wireless networks, its performance is dependent on predefined network conditions and their corresponding configurations. The ultimate vision of self-driven networks is to realize a retained network performance, maintain sustainable resilience to network variations, and design versatile networks that are capable of handling new network conditions and scenarios.

As a cornerstone, generative AI (GenAI) can be a key player in realizing the vision of self-evolving networks. GenAI refers to a class of AI that is designed to generate new content, including text, images, videos, etc., based on patterns and information learned from large datasets. Accordingly, GenAI holds immense potentials in several directions, e.g., data augmentation, text and image generation, question-answering, sentiment analysis, conversational agents, human-machine interactions, automation, and many more. Within this context, Large Language Models (LLMs), a foundation model built on natural language understanding and a subfield of GenAI, have recently attracted a considerable attention from the research community as a revolutionary technology in the field of AI [2]. Subsequently, multi-modal LLMs have expanded the capabilities of LLMs from natural language to 2D/3D vision, sound signal, X-ray, etc., rendering them a promising candidate for applications in different domains including telecommunications [3].

Among several LLMs, Falcon LLM [4], generative pre-trained transformer (GPT)-2/3/4, Bidirectional Encoder Representation from Transformer (BERT), large language model Meta AI (LLaMA), as well as visual language models, e.g., DALL-E3 based on Contrastive Language-Image Pre-Training (CLIP) and diffusion model, have strongly impacted how AI is employed for inference and decision-making purposes, and laid down a new base for novel applications that can exploit the potentials of GenAI models. This is rooted to the generative and predictability capabilities of these Large-GenAI-Models (Large-GenAI-Models), in which large models (mainly based on the transformer architecture) are trained over a vast amount of unlabeled multimodal data (primarily textual and/or visual data), and therefore, are enabled to understand and generate human-like languages. Through the self-attention mechanism of transformers and the large amount of training data, the developed large models are able to capture the statistical patterns and relationships in the provided data, and hence, to predict and generate the required data. Similar approaches apply to visual generative models, where variational autoencoders (VAEs), generative adversarial networks (GANs) and diffusion models can be leveraged to map textual data with images and vice versa.

Being a promising candidate to revolutionize several technologies in various fields, we believe that GenAI models can introduce a radical change in wireless network design and operation. In particular, we anticipate that Large-GenAI-Models will introduce a tangible enhancement in the performance of different schemes within the Telecom domain. This can be achieved through exploiting the generative capabilities of Large-GenAI-Models in addition to the multimodality nature of the data acquired in wireless networks, including radio frequency (RF) signals, and 2D and 3D visual representations of wireless environments, to attain improved contextual, situational, and temporal awareness, and therefore, enhanced wireless communication. Furthermore, it is noted that Large-GenAI-Models can enable wireless networks to enjoy predictability features, and hence, realize improved and proactive localization, beamforming, power allocation, handover, as well as, spectrum management, even for unseen network scenarios. On the other hand, wireless network can connect multiple GenAI models with novel communication mechanisms at semantic and effectiveness level, to provide

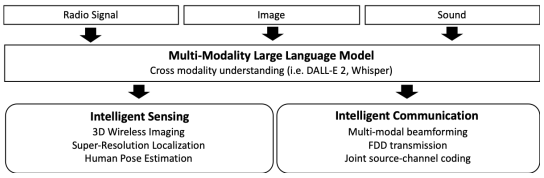

Fig. 1: Example applications of Large-GenAI-Models for wireless sensing and communication

faster sensing, inference, action with reduced resources consumption.

Recently, AI has been playing a key role in a wide range of applications in Telecom networks. However, most AI solutions in wireless networks are developed for solving dedicated problems, which is inefficient for general use in wireless network with emerging applications. Large-GenAI-Models on the other hand represent a pre-trained model that can solve many downstream and upstream tasks with efficiently designed prompts or through fine-tuning, rendering such models a promising candidate for AI native networks, in which massive data pertinent to system design (architecture, protocol) and operation (signaling, procedure) can be used with Large-GenAI-Models to perform many tasks cross different layers. We foresee that Large-GenAI-Model-like solutions could generalize over unseen data, tasks, and scenarios, and reduce the complexity, cost, and power in network design, deployment and operation. Large GenAI models in Telecom are envisioned to have a pivotal role in enhancing wireless communications by enabling automatic network control and optimization, as well as facilitating complex tasks like channel estimation and environment reconstruction. Additionally, the integration of cloud-based GenAI models into network edge devices should prompt a shift in wireless communication towards a knowledge-based paradigm, promoting collaborative interactions among on-device GenAI models to improve resource efficiency and reduce energy consumption.

A. Contribution

While there are several articles that discussed foundation models in general, and LLMs in particular, e.g., [5]–[8] - and the references therein, all of these articles have focused on the application of LLMs in the general domain, with specific interest in textual data with minor discussions on visual data. In this article, *for the first time in the literature*, we explore the integral role that large generative models will play in future wireless networks. From the one hand, we aim to unveil the potentials of leveraging large generative models to pre-train a single foundation model that can subsequently be employed across a range of different downstream tasks in the Telecom domain, for improved wireless sensing and transmission. This approach has several significant advantages, including improved efficiency, reduced training requirements, and enhanced adaptability to varying network conditions, as well as the generalizabilty of these models in different propagation environments. For this purpose, we introduce a new term, namely, *Large-GenAI-Models*, which refers to large generative models that are designed and trained to fit various applications and use-cases of Telecom networks, from network design and configuration, to optimization and operation. On the other hand, we draw a roadmap to how wireless networks can potentially contribute to developing efficient large models, that are capable of accommodating the needs of future 6G networks. We finally reveal how the two paradigms (Large-GenAI-Model for wireless and wireless for Large-GenAI-Model) pave the way to the conceptualization of the artificial general intelligence (AGI)-empowered wireless networks, the seed to fully self-evolving networks.

II. Large-GenAI-Model for Wireless

A. Large Language Models for Sensing

1) 3D Wireless Imaging: Deep Learning (DL) models have fundamentally contributed to the development of improved wireless sensing schemes, in which RF data can be acquired and mapped to 2D images for sensing applications, including localization, remote sensing, and resource allocation. While DL approaches have demonstrated an acceptable performance in several scenarios, they lack the generalizability to new network requirements and scenarios. Furthermore, they suffer from the complexity that arises from the reliance on supervised training data. For the latter, to achieve high-resolution sensing for mission-critical applications, large labeled data-sets are required, rendering the labeling process a huge burden. Within the same concept, neural radiance field (NeRF) framework [9] was proposed to understand the interaction of light signals with objects in a scene to create realistic and detailed 3D representations, relying on the DL architecture. Also, NeRF2 was proposed to understand the physical wave transmission which can be leveraged for accurate CSI acquisition and hence, improved performance in localization applications, however, it suffers from high computational complexity and lacks the ability to scale. The recent evolution in visual GenAI models has opened the horizon to a new era of sensing capabilities, in which machines now have the potential to generate high-quality images and understand visual content, i.e., generating 2D and 3D images from textual description or mapping images with corresponding text.

Among others, DALL-E3 (capable of generating high-quality images from textual descriptions), CLIP (capable of matching images with text captions) , Vector Quantized GAN (generate diverse and high-quality images), and GPT-4V (focuses on generating textual descriptions of images), have been identified as core models for 2D and 3D image generation from text, or vice versa. These models rely on the several advanced techniques including conditional GANs, discrete VAEs, and diffusion models to capture the visual patterns and their contextual relationship with textual content.

The promising potentials of visual GenAI models introduce a plethora of benefits that can be exploited for wireless communications, for enhanced system design and optimization. This vision is built upon the principle of enabling GenAI models to understanding the cross features between different data modalities, including images of wireless propagation environments and RF signals. Specifically, these visual models will be developed to generate super-resolution 3D images of

the surrounding environment from measured wireless data. The constructed 3D image will be then leveraged to enable improved communication schemes, in which enhanced contextual and situational awareness can be realized, and therefore, achieving improved beamforming, handover, resource allocation, etc. This is due to the fact that Telecom visual GenAI models are anticipated to capture the relationships between RF data and different information that can be extracted from 3D images (static system topology, non-moving objects, dynamic objects information), and the impact of the latter on the former, enabling accurate multi-dimensional reconstruction of the wireless environment from the acquired RF signals. Note that it has been shown that visual GenAI model (i.e. diffusion model) can effectively perform image construction from brain signals (via functional magnetic resonance imaging (fMRI)) [10]. Moreover, meta-transformer has demonstrated a unified transformer framework to incorporate sound, X-ray, and hyper-spectrum signals to perform many downstream tasks such as segmentation, detection, provided that it can align different modalities on a semantic latent space [3]. Unlike conventional visual GenAI models that map images with texts using text/image encoders/decoders, the architecture of Telecom visual GenAI models are aimed to comprise encoders that are capable of extracting features and patterns from wireless signals and map these features into 3D images. Telecom-oriented Large-GenAI-Model framework can be developed to extract and concatenate images and signals embeddings, and allow the machine to understand the relationship between the two data modalities through self-attention mechanism.

2) Super-Resolution Localization: Super-resolution localization is an essential element in current wireless networks, since determining the accurate locations of different wireless devices or nodes in the network readily facilitates optimizing network performance, efficiently allocating resources, and ensuring maximized quality-of-service (QoS) by dynamically adjusting network parameters, to accommodate the needs of different users according to their current status and positions. Current localization techniques follow mainly two directions, RF-based localization, which rely on RF signals from anchor nodes for position detection, or vision-aided localization, which exploits visual data and image processing for feature extraction and object detection. While the latter can provide more contextual information that render higher localization accuracy, the limited field-of-view and calibration and precise alignment of cameras or sensors constitute limiting factors in such approaches. These challenges are further pronounced in multimodal localization, in which the fusion and alignment of different data modalities results represent a bottleneck in exploiting the full potential of multimodal localization.

Within this context, Large-GenAI-Model can be a game changer in enabling efficient multimodal localization schemes, in which the general and self-attention nature of these large models can be the key to detect the contextual and situational information of network users and nodes, capture the mutual displacement between multiple images, and cross-correlate these images and their variations with the corresponding electromagnetic behaviour of wireless signals. The benefits of utilizing Large-GenAI-Models for high-precision localization are two-fold. First, Large-GenAI-Models can introduce improved perception skills of the surrounding environment through integrating multimodal data that allows the understanding of the environmental, temporal, and situational aspects of the network events and behaviours, and their impact on the network performance. With the generative capabilities of Large-GenAI-Models, it is envisioned that a pretrained large model will be capable of accurately locating multiple users within the network. Second, the exploitation of Large-GenAI-Models in localization applications is stemmed from the fact that pretrained large models allow wireless networks to enjoy improved predictability skills. In more details, the integration of 3D images with RF data enables the large models to identify both idle and active users, as well as predict their future activities through their generative and prediction capabilities. This offers a comprehensive understanding of the wireless environment and its variations, and accordingly enable enhanced network configuration and optimization. This yields more proactive networks, paving the way to self-built and self-healing networks.

B. Large Language Models for Transmission

1) Multi-modal beamforming: The evolution of high-frequency communications, e.g., mmWave and THz communication, has given a rise to several issues pertinent to the transceiver design (including efficient high-power antenna design and beamforming/MIMO schemes), as well as the reliability of signals transmitted over high frequencies. This is due to the lossy and highly directional nature of high-frequency communication, which yields severe signals degradation even with minor blockage or beam misalignment. Accordingly, efficient beamforming and beam alignment mechanisms are essential to be developed to achieve reliable high-frequency communication. Typically, beam selection is performed through the reliance on a predefined beam codebook, in which a wide-range of beam sweeping is performed between the transmitter and the receiver in order to select the optimum beam to establish a reliable link between the two nodes. Such a procedure is usually associated with high beam training overhead, introducing a performance limiting factor in high-mobility and latency-sensitive applications.

Large-GenAI-Models, pretrained over a large data-set of beamforming scenarios, have the capability to predict the optimum beam that maximizes the signal strength and minimizes interference. This can be achieved through exploiting multimodality for providing additional information in regard to blockage probability and users status and activities, and hence, an Large-GenAI-Model can be used to capture the various features and patterns within the network dynamicity, and predict the optimum beam according to current and future network scenarios.

2) Frequency Division Duplexing (FDD) Transmission: Within the context of FDD, Large-GenAI-Models can be utilized for channel state information (CSI) estimation purposes between the uplink and downlink transmissions. In conventional FDD systems, where separate frequency bands are allocated for uplink and downlink transmissions, CSI acquisition is

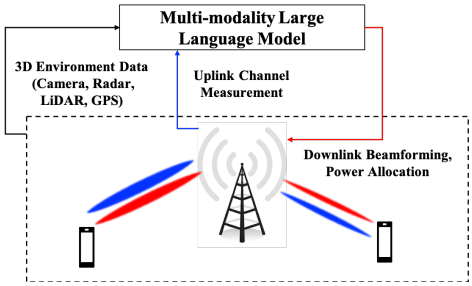

Fig. 2: Large-GenAI-Models for Beamforming in FDD Systems

typically performed separately for the two directions, consuming the network resources and introducing high latency. This issue is more pronounced in massive multiple-input multiple-output (MIMO) scenarios, where it is very challenging to acquire all CSI over the uplink [11]. Alternatively, partial uplink CSI knowledge can be exploited to extrapolate the full downlink CSI. Through the self-attention mechanism and the generative capabilities of Large-GenAI-Models, we envisage that Large-GenAI-Models will be able to capture the inherent relationship between the uplink and downlink transmission and exploit 3D multimodal environment data (including camera, radar, LiDAR, and GPS) in order to select the optimum uplink and downlink beam pair that yields perfect alignment between the angle-of-arrival and angle-of-departure, at a particular user position (Fig. 2). It should be emphasized that super-resolution localization achieved by Large-GenAI-Models (as discussed in Sec. II-A) can directly impact the beamforming performance in FDD systems, in which accurate localization allow Large-GenAI-Model-enabled beamforming schemes to predict future user position, and therefore, enable improved planning for the spectrum resource over the uplink and downlink transmissions.

3) Joint Source-Channel Coding (JSCC): Conventional wireless networks are designed in a block-based architecture, where each process has a dedicated block, including channel estimation, coding, decoding, equalization, to name a few. With the identified requirements of future 6G networks, such architecture cannot achieve the envisioned key performance indicators of latency, reliability, spectral and energy efficiency, and connectivity. It is further emphasized that block-based architectures are difficult to scale to large, more generalized scenarios [12]. In this regards, current research initiatives have been conducted to explore the appealing benefits of integrating channel and source coding into semantic-aware JSCC, in order to serve use-cases with extreme latency and bandwidth requirements, that are computationally demanding in long block-length techniques. When incorporating different data modality, e.g., images and videos, to be transmitted, DL approaches have demonstrated outstanding performance in terms of latency and quality-of-compression. However, DL models rely heavily on a huge amount of labeled training data for the joint optimization of source and channel coding, which, in most scenarios, is not available. This readily impact the generalizability of the trained models to unseen or varying (channel conditions, interference, noise, etc.) network scenarios.

In this regard, Large-GenAI-Models, can facilitate the realization of efficient JSCC schemes for improved wireless communication. In addition to the labeled data independency, Large-GenAI-Models can be exploited to understand the statistical behavior of the source data and to allow the reliable extraction of the needed information, and hence, realize efficient data compression [13]. Furthermore, through the self-attention mechanism, Large-GenAI-Model introduce improved robustness to channel errors, through enabling efficient error correction mechanisms. This can be achieved by using Large-GenAI-Models to capture the inherent features of wireless channels behavior, error statistics, and source data, and then learn the relation between the source data and the corresponding channel coding requirements, and therefore, realize improved prediction for errors and promotes the development of robust error correction codes. Moreover, by understanding the long interdependencies between the source data and channel coding, adaptive mechanisms can be designed with reliance on Large-GenAI-Models, where real-time adjustment of the code rate, modulation, and code selection can be performed according to the current channel condition, striking a balance between source coding and channel coding performance.

III. Wireless for Large-GenAI-Model

A. 6G with Collective Intelligence

An evolution of GenAI is to empower a massive number of wireless devices to deliver collective intelligence. To achieve this, communication systems should be transferred from data and model-based to knowledge and reasoning-based paradigms. Conventional communication networks aim to transfer data from one network node to another under targeted key performance indicators (KPIs). This is inefficient for a network connecting large-scale devices powered by Large-GenAI-Models. The goal of 6G is to develop a computing fabric that moves knowledge within the network. This can push nowadays cloud Large-GenAI-Model trained on world knowledge towards distributed collective intelligence. To achieve this, Large-GenAI-Models should be grounded to the real-world context, communicate with knowledge, to perform multi-agent planning, decision making, and reasoning. In this section, we summarize three key research perspectives to realize a multi-agent GenAI wireless network.

1) Semantic communication: Large-GenAI-Models built on raw data could cause large computation overhead and communication redundancy. Learning an effective semantic abstraction of raw data is essential to deploy and connect large GenAI models on resource constrained wireless devices. This requires wireless physical layer to be redesigned in a semantic and effective manner for the GenAI models. Large-GenAI-Models can learn from the semantic compression and abstraction of raw data represented as knowledge [13]. In this context, information should be characterized with a minimal structure, which is robust against changes in distribution, domain, and context. This does not only reduce the size of data and model in the device memory, but also can represent different data modalities with a common latent (concept) space. Knowledge

can be potentially modeled on a topological space, such as graphs, simplices, cells, complexes, which can largely reduce the data transmitted between agents while being effective for Large-GenAI-Models to accomplish the targeted tasks. On the other hand, the higher-order interactions between concepts on topological space allows Large-GenAI-Models to perform logical reasoning from one domain to another, which reduces the needed raw data collection and exchange between agents.

2) Emergent protocol learning: The GenAI powered wireless devices will have many emerging applications in different domains. Conventional 5G system protocols developed for specific applications (i.e. XR, MTC, IoT) are inflexible to meet the diverse goals of GenAI use cases. An emergent, adaptable, and autonomous communication protocol on a goal-oriented basis is essential to allow multiple wireless GenAI agents interact or collaborate effectively.

Multi-agent reinforcement learning (MARL) has the potential to learn a semantic language between agents through collaboration, such as when, where, and what information should be delivered. This permits to design a goal oriented, compositional and grounded wireless MAC and network layer protocols between devices. Moreover, multi-agent game can be an alternative technique to improve agent's action in a competition scenario. This can be performed jointly at the network level to optimize resource allocation, and at the device level to optimize vehicle or robotic control. Furthermore, Large-GenAI-Models can guide MARL and games using its expert model for planning, for improved convergence and reduce data collection.

3) Distributed Large-GenAI-Models powered AI agent: To achieve full autonomy for a self-evolved network, an autonomous agent with the capability of grounding, planning, criticising, and reasoning is essential. With autonomous agents, wireless devices not only perform inference with Large-GenAI-Models, but also perceive environment, plan tasks, memorize experiences and evaluate actions. This could lead to a fully autonomous control in both network and device without human intervention. To achieve this, the Large-GenAI-Model knowledge should be grounded to the real-world context. Reinforcement learning has recently been applied on Large-GenAI-Models, leveraging environment perception and rewards to improve the decisions made by the Large-GenAI-Models [14]. With distributed planning, Large-GenAI-Models can a create sequences of actions to accomplish a task, and distribute sub-tasks to different wireless devices so as to improve the execution efficiency via cooperation. As a generative model, an LLM can make wrong decisions and answers, where the criticism between multiple LLM instances is proved to improve the performance [15]. Moreover, Large-GenAI-Models can perform reasoning via planning, such as chain or tree of thoughts. By learning the logical relations between concepts abstracted from data, Large-GenAI-Models can have better generalization capabilities, with less training data, computation cost, and faster inference, which is important on energy constrained wireless devices. Furthermore, semantic communication and emergent protocol can be used to connect multiple autonomous agents effectively for cooperative planning and reasoning. The cloud-based Large-GenAI-Models can act as a world model guide wireless agents to optimize in a real-world context from Large-GenAI-Model's common-sense knowledge.

B. Use Cases of Collective Intelligence

GenAI has the potential to realize automation in intent-driven networking. With Telecom Large-GenAI-Model-empowered autonomous agents, network can self-evolve in a distributed manner without human intervention. With the capability of distributed planning, Large-GenAI-Models can break down a higher-level intents into lower level actionable tasks in wireless network, sense radio environment, take action to configure system, and memorize experience for future. Furthermore, effectiveness communication between multiple GenAI agents in network devices can eliminate the need for a centralized control, and reduce the signaling load on control plane.

Wireless GenAI network can also bring collective intelligence to autonomous vehicles. With autonomous agents deployed on vehicles and road side units, semantic communication and emergent protocols on the V2X data plane can largely improve the transmission efficiency, latency and reliability in both communication and control, compared to 5G delivering raw perception via eMBB and control commands via URLLC. Moreover, cooperative perception and control through Large-GenAI-Model-guided MARL or games can improve the traffic flow and safety, bringing full autonomy to vehicle network.

IV. AGI-Empowered Wireless Networks: A Vision Forward

The evolution of several AI paradigms is approaching the concept of AGI, in which machines will enjoy a level of intelligence that is equivalent to or surpassing humans intelligence. The notion of the AGI represents a paradigm for building machines that are capable of mastering a wide range of tasks without being particularly trained on them, through possessing a broad spectrum of human-like cognition capabilities, including understanding, learning, applying knowledge, and performing reasoning and inferring in various domains, in an autonomous fashion. It is envisioned that Large-GenAI-Models, with their generalized and generative capabilities, will be the keystone towards the successful deployment of AGI in wireless networks, in which AGI, empowered by Large-GenAI-Models, can be the orchestrator to the efficient planning, design, deployment, configuration, and operation of future wireless networks. In what follows, we put our forward-looking vision of how Large-GenAI-Models, or more generally foundation models, will be integrated into wireless networks, with the aim to realize the true vision of network automation.

A. Task-Agnostic Large Telecom Model

The main purpose of adapting the large generative AI models into the Telecom domain is to develop foundation models that are trained using multi-modal data, in order to perform general Telecom tasks. These foundation models are aimed

to enable improved communication and control in wireless networks. In particular, these foundation models, or large Telecom models, are anticipated to act as a general-purpose backbone to the network, with high scalability and flexibility to be deployed on edge devices. This can be achieved by designing large GenAI models to perform several general Telecom-oriented tasks, in order to reduce the cost of training multiple AI models to perform specific tasks. After that, the pretrained model can be tailored to fit within a particular downstream task of interest, including modulation, coding, power allocation, beamforming, etc., through fine-tuning with the relevant data, through fine-tuning the foundation model using the relevant data. This represents a stepping stone to implementing Large-GenAI-Models at edge devices, where the fine-tuning process is much less complex and smaller data-set is required.

B. Self-Evolving Networks

The concept of self-evolving networks refers to the networks that are capable to adapt, change, and evolve with the variations experiences in the network, and the surrounding environment. It can autonomously plan long-term tasks, make decisions, reason, and memorize logic, to achieve particular goals without human intervention. By leveraging AI, these networks will have the capability to optimize their configurations in real-time fashion, according to the users demands and the network dynamics. This will result in a sustainable performance without human intervention. With the generalization capability, multi-modality, and planning and reasoning, Large-GenAI-Models can realize the vision of self-evolving networks which will go beyond the self-adaptation and self-optimization principles. In specific, we anticipate that Large-GenAI-Models will be leveraged to contribute to the initial steps of designing, planning, deploying, configuring, and operating wireless networks. The use of textual documents from the standards and research reports will allow pretrained Large-GenAI-Models to generate the required software codes and hardware design specification, which will be then go to the deployment stage. In the later stages, Large-GenAI-Models can be employed to identify the optimum configuration of the network according to the initial design, requirements, and users needs. It can be further exploited to build new communication schemes, that are not necessarily compatible with a particular standard, according to a particular network condition, for example, novel waveform to cope with high mobility scenario, modulation and coding design to accommodate multiple high data-rate users, etc. This vision of *AGI-empowered wireless networks* (Fig. 3) can be achieved with the utilization of the diverse Telecom data modality and the predictability and generative nature of Large-GenAI-Models.

V. Challenges and Open Research Directions

A. Telecom Large-GenAI-Models Architectures

Despite the high number of architectures designed for the Large-GenAI-Model, that are capable of handling multimodality, adapting the paradigm of Large-GenAI-Models to the Telecom domain will require sophisticated architectures that are designed and trained from scratch, in order to fit within the Telecom domain. This is stemmed from the fact that Telecom data enjoy unique characteristics, compared to textual data and images/videos. In particular, it is expected that the available LLM architectures are incapable of handling RF data, and integrating it with other data modalities.

B. Distributed GPT

With the vast amount of data generated at different nodes in the network, distributed GPT, in which the large generative AI model is trained in a distributed fashion, offers several benefits that make it appealing particularly for wireless networks. From the one hand, distributed GPT allows high horizontal scalability in which a larger number of users can contribute to the training process, and hence, reduced latency can be achieved. The latter can be a consequence of processing the data and training the model closer to the data source. Furthermore, distributed GPT can be more efficient in adapting to the dynamicity of wireless networks, in which decisions can be made faster according to locally-processed data. However, distributed GPT faces several challenges in terms of models coordination and aggregation, and data synchronization.

C. Explainable Large GenAI Models in Telecom

The explainability of large generative models is a key factor to be thoroughly studied in the Telecom domain. As being integrated into the network design, management, and control, and being granted authoritative and decision-making roles, it is essential to understand how these complex generative models arrive at their conclusions, and rationale behind their decisions. The explainability of Large-GenAI-Models is essential for two main reasons. First, in many scenarios, human operators work in collaboration with AI agents, and therefore, it is important for these operators to understand why a particular decision is made by the AI agent, enabling them to make more informed choices. Second, understanding the logic behind opens up many opportunities for optimization and enhancement, as well as realize trustworthy behavior for Large-GenAI-Model agents.

D. GenAI On-Device

The distributed nature and the emergence of edge-computing paradigm push for the deployment of large GenAI models on edge devices, offering several benefits in terms of reduced latency, offline operation, as well as the reduce reliance on cloud services. However, taking into consideration the extremely large sizes of Large-GenAI-Models and the huge amount of data to be processed, computing and energy resources are two main factors to be optimized when implementing GenAI models on edge devices. Several techniques can be adopted to optimize the model efficiency in order to reduce the computational requirements, including model compression and quantization, and knowledge distillation. Furthermore, such a resource optimization problem can be achieved through exploiting schemes such as tasks offloading, dynamic resource allocation, and optimized communication protocols.

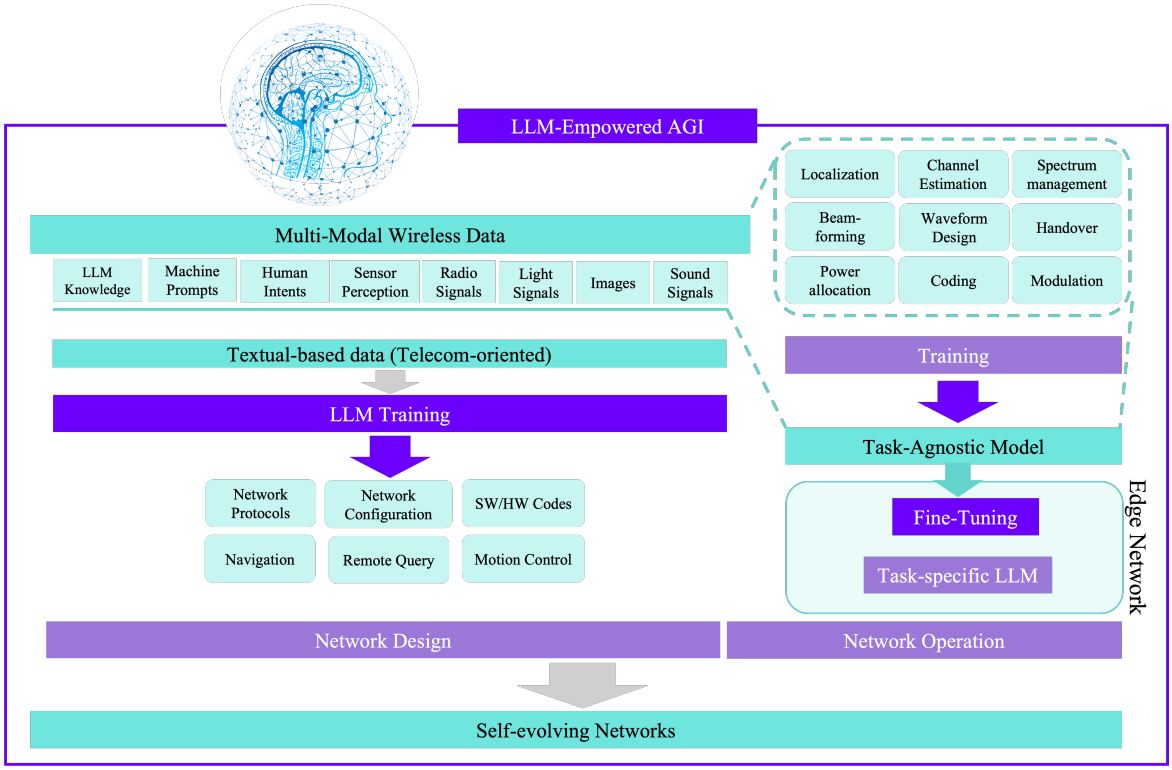

Fig. 3: AGI-empowered wireless networks.

VI. Conclusion

In this article, we explored how Large-GenAI-Models can be an essential tool in designing, configuring, and operating future wireless networks. In particular, we identified the key opportunities, with respect to sensing and communication, that can be acquired when employing Large-GenAI-Models in wireless networks, and we overviewed the role of wireless networks in enabling machines to communicate using Large-GenAI-Models. Moreover, we laid down the foundation for the development of the AGI-empowered wireless networks through Large-GenAI-Models, which paves the way to the successful implementation of self-evolving networks.

References

[1] M. Chafii et al., "Twelve scientific challenges for 6G: Rethinking the foundations of communications theory," IEEE Commun. Surveys Tuts., 2023.
[2] W. X. Zhao et al., "A survey of large language models," arXiv preprint arXiv:2303.18223, Sep. 2023.
[3] Y. Zhang et al., "Meta-transformer: A unified framework for multimodal learning," Jul. 2023.
[4] Falcon LLM. [Online]. Available: https://huggingface.co/tiiuae
[5] B. Min et al., "Recent advances in natural language processing via large pre-trained language models: A survey," ACM Computing Surveys.
[6] C. Zhou et al., "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt," arXiv preprint arXiv:2302.09419, May 2023.
[7] M. U. Hadi et al., "A survey on large language models: Applications, challenges, limitations, and practical usage," TechRxiv, Sep. 2023.
[8] J. Y. Koh et al., "Grounding language models to images for multimodal generation," arXiv preprint arXiv:2301.13823, Jun. 2023.
[9] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," Communications of the ACM, vol. 65, no. 1, pp. 99–106, 2021.
[10] Y. Takagi and S. Nishimoto, "High-resolution image reconstruction with latent diffusion models from human brain activity," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14 453–14 463.
[11] Y. Yang et al., "Deep learning based antenna selection for channel extrapolation in FDD massive MIMO," in International conference on wireless communications and signal processing (WCSP). IEEE, 2020, pp. 182–187.
[12] Y. Wang et al., "Transformer-empowered 6G intelligent networks: From massive MIMO processing to semantic communication," IEEE Wireless Commun., 2022.
[13] H. Gilbert et al., "Semantic compression with large language models," 2023.
[14] T. Carta et al., "Grounding large language models in interactive environments with online reinforcement learning," Sep. 2023.
[15] Y. Du et al., "Improving factuality and reasoning in language models through multiagent debate," May 2023.

Biographies

Lina Bariah (lina.bariah@ieee.org) is a Senior Researcher at Technology Innovation Institute, UAE.

Qiyang Zhao (Qiyang.Zhao@tii.ae) is a Lead Researcher at Technology Innovation Institute, UAE.

Hang Zou (Hang.Zou@tii.ae) is a Researcher at Technology Innovation Institute, UAE.

Yu Tian (Yu.Tian@tii.ae) is a Researcher at Technology Innovation Institute, UAE.

Faouzi Bader (Carlos-Faouzi.Bader@tii.ae) is the Director of Telecom Unit at Technology Innovation Institute.

Mérouane Debbah (Merouane.Debbah@ku.ac.ae) is a Professor at Khalifa University, UAE.