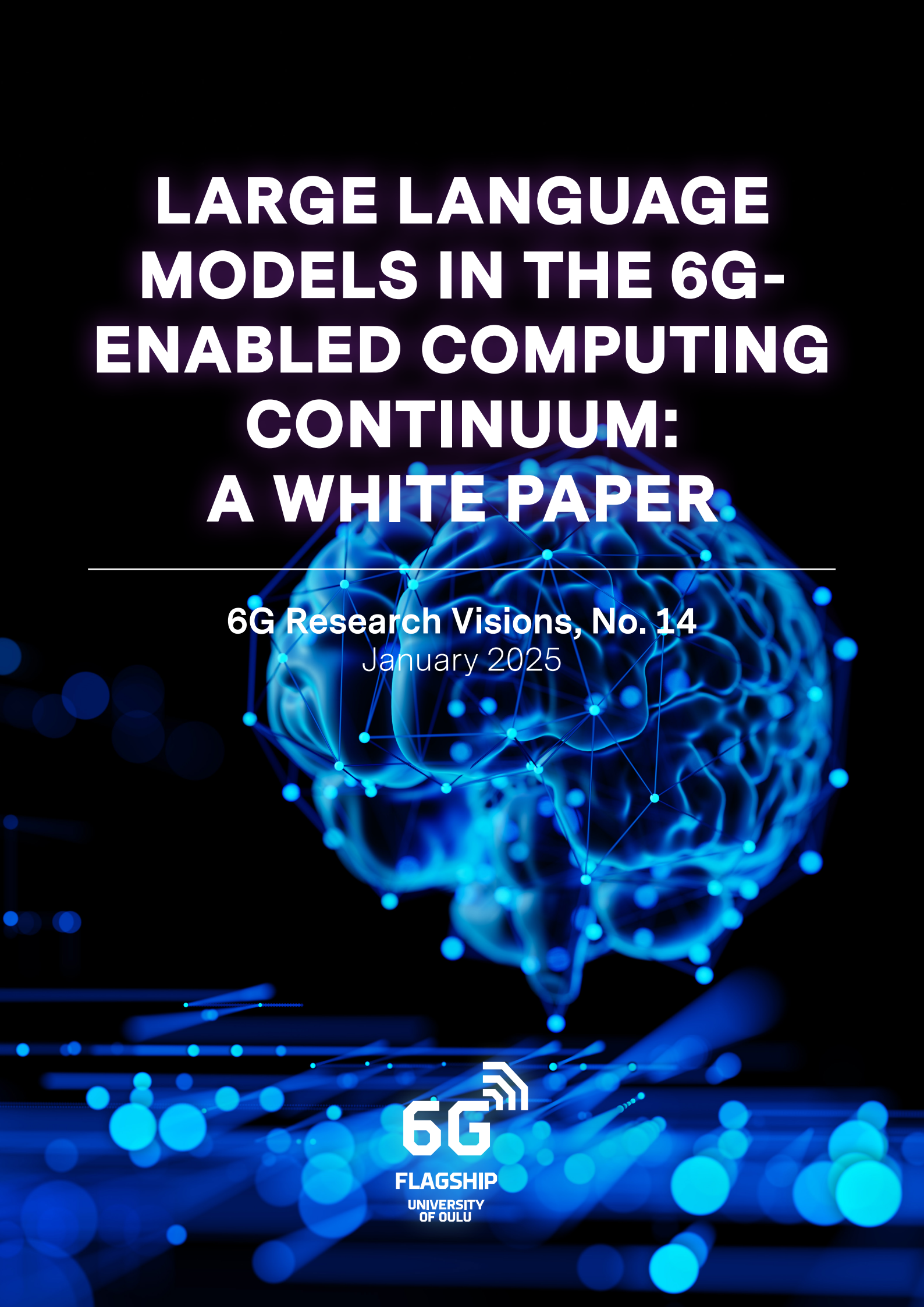
Large Language Models in the 6G-Enabled Computing Continuum: a White Paper

Markus Abel, Ijaz Ahmad, Constantino Alvarez Casado, Rico Berner, Mickaël Bettinelli, Kaj Mikael Björk, Michele Capobianco, James Gross, Tri Hong Nguyen, Pan Hui, et al.

LARGE LANGUAGE MODELS IN THE 6G-ENABLED COMPUTING CONTINUUM: A WHITE PAPER

6G Research Visions, No. 14
January 2025



6G FLAGSHIP
UNIVERSITY OF OULU

Table of Contents

Large Language Models in the 6G-Enabled Computing Continuum: a White Paper

6G Research Visions, No. 14, 2024
ISSN 2669-9621 (print)
ISSN 2669-963X (online)
ISBN 978-952-62-4375-7 (print)
ISBN 978-952-62-4376-4 (online)

List of authors

- Markus Abel, Ambrosys GmbH, Germany
- Ijaz Ahmad, VTT Technical Research Centre of Finland, Finland
- Constantino Alvarez Casado, Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland
- Rico Berner, Ambrosys GmbH, Germany
- Mickaël Bettinelli, LISTIC, University Savoie Mont Blanc, France
- Kaj Mikael Björk, Centre For Intelligent Computing (CIC), University of Turku, Finland
- Michele Capobianco, Capobianco, Italy
- James Gross, KTH Royal Institute of Technology, Sweden
- Tri Hong Nguyen, Department of Computer Science, Aalto University, Finland
- Pan Hui, Hong Kong University of Science and Technology (GuangZhou), China
- Panos Kostakos, European Public Prosecutors Office (EPPO), Luxembourg
- Abhishek Kumar, Center for Ubiquitous Computing (UBICOMP), University of Oulu, Finland
- Mika-Petri Laakkonen, Oulu University of Applied Sciences, Finland
- Xiaoli Liu, University of Helsinki, Finland
- Zhi Liu, The University of Electro-Communications (UEC), Japan
- Le Nguyen, Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland
- Huong Nguyen, Center for Ubiquitous Computing (UBICOMP), University of Oulu, Finland
- Basak Ozan Ozparlak, Ozyegin University, Türkiye
- Ville Pietiläinen, University of Lapland, Finland
- Susanna Pirttikangas, Center for Ubiquitous Computing (UBICOMP), University of Oulu, Finland
- Stéphan Plassart, LISTIC, University Savoie Mont Blanc, France
- Sampo Pyysalo, University of Turku, Finland
- Soheyb Ribouh, LITIS, Université de Rouen Normandie, France
- Jari Rinne, University of Lapland, Finland
- Mehdi Safanpour, Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland
- Alaa Saleh, Center for Ubiquitous Computing (UBICOMP), University of Oulu, Finland
- Saeid Sheikhi, Center for Ubiquitous Computing (UBICOMP), University of Oulu, Finland
- Olli Silvén, Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland
- Harry Souris, Silo AI, Finland
- Xiang Su, University of Helsinki, Finland
- Roope Suomalainen, Silo AI, Finland
- Athanasios V. Vasilakos, University of Agder, Norway
- Aleksandr Zavodovski, Ericsson, Finland
- Qi Zhang, Aarhus University, Denmark
- Peng Yuan Zhou, Aarhus University, Denmark
- Alireza Zourmand, Häme University of Applied Sciences (HAMK), Finland

Editors
- Lauri Lovén, Center for Ubiquitous Computing (UBICOMP), University of Oulu, Finland
- Miguel Bordallo López, Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland
- Roberto Morabito, EURECOM, France
- Jaakko Sauvola, Empirical Software Engineering in Software, Systems, and Services Research Unit (M3S), University of Oulu, Finland
- Sasu Tarkoma, University of Helsinki, Finland

Reviewers
- Marja Matinmikko-Blue, Infotech Oulu and Centre for Wireless Communications (CWC), University of Oulu, Finland
- Marcos Katz, Centre for Wireless Communications (CWC), University of Oulu, Finland
- Shahriar Shahabuddin, Oklahoma State University, US
- Mehdi Bennis, Centre for Wireless Communications (CWC), University of Oulu, Finland

Please cite
Lovén, L., Bordallo López, M., Morabito, R., Sauvola, J. & Tarkoma, S. (Eds.) (2025). *Large Language Models in the 6G-Enabled Computing Continuum: a White Paper* [White paper]. (6G Research Visions, No. 14). University of Oulu. ***https://urn.fi/URN:NBN:fi:oulu-202501211268***

6G Flagship, University of Oulu, Finland
January 2025

Acknowledgement
This white paper has been written by an international expert group, led by the Finnish 6G Flagship program (6gflagship.com) at the University of Oulu, within a series of 6G white papers.

Abstract

The evolution towards 6G architecture will shift communication networks, with artificial intelligence (AI) playing a key role. This white paper examines the integration of Large Language Models (LLMs) within 6G systems. Their ability to grasp intent, reason, and plan, and execute commands will redefine network functionalities and interactions. An essential component is the AI Interconnect framework, designed to facilitate AI operations within the network. Building on the evolving state-of-the-art, we present a new architectural perspective for the next generation of mobile networks. Here, LLMs will work together with pre-generative AI and machine learning (ML) algorithms. This union combines old and new methods, merging established approaches with AI technologies. We provide an overview of this evolution and explore the applications arising from such an integration. We envisage an integration where AI becomes central to future communication networks, offering insight into the structure and function of a 6G network centered on AI.

Index Terms
6G, Generative AI, Large Language Model (LLM), Generative Pre-trained Transformers (GPT), AI Interconnect, Edge Intelligence, Open RAN (O-RAN), AI RAN

1 Introduction

AI's potential is evident across sectors, especially in telecommunications. As AI's role in telecom has grown, the term "AI-native telco" has become popular. Recent contributions from industry and academia outline the path for mobile networks and 6G architecture [1], [2], [3], describing an AI-native system with trustworthy AI capabilities. These systems integrate AI into their design, deployment, operation, and maintenance. Their hallmark is a data-centric, knowledge-driven environment where data is produced and used to develop AI functionalities, supporting applications across various use cases and domains [4], [5]. This shifts from static, rule-based systems to adaptive, learning-oriented AI models. Large Language Models (LLMs), especially Generalized Pretrained Transformers (GPT) [6], [7], [8], are tools for understanding intent, creating plans, and executing instructions [7]. This capability is enhanced by prompt engineering, programming LLMs via prompts [9]. Thus, they are essential components of future networks and applications.

In this white paper, we advocate for an AI-native 6G network that seamlessly integrates a diverse range of LLMs, allowing for their dynamic selection, provisioning, updating, and creation. Central to this vision is the AI Interconnect, which streamlines AI-centric operations within the network. By incorporating AI directly into the network, we anticipate marked improvements in areas such as radio and network optimization, privacy and security via tailored AI and resource selection, enhanced accountability through meticulous AI operation monitoring and audit trails, and meeting stringent latency and other network-specific criteria. Such advancements are poised to be invaluable for domains like the Internet of Things (IoT), robotics, smart cities, and autonomous systems, to name a few.

This white paper was developed through a collaborative process, starting with an open call for contributions. Proposed contributions were reviewed by the editorial team and incorporated into the document, ensuring diverse perspectives and expertise. Additionally, a set of reviewers provided detailed comments to enhance its quality and depth. Adopting the form of a review, this white paper provides a comprehensive exploration of the topic, synthesizing existing knowledge while proposing new directions to advance the field. The primary goal of this white paper is provide a vision grounded in ongoing innovations, that could guide research and development efforts, and outline actionable insights that align with the future needs of 6G technologies and their applications.

Editors:
Lauri Lovén, Miguel Bordallo López, Roberto Morabito, Jaakko Sauvola, Sasu Tarkoma

Reviewers:
Marja Matinmikko-Blue, Marcos Katz, Shahriar Shahabuddin, Mehdi Bennis

Contributors:
Markus Abel, Constantino Alvarez Casado, Rico Berner, Mickaël Bettinelli, Kaj Mikael Björk, Michele Capobianco, James Gross, Tri Hong Nguyen, Pan Hui, Ijaz Ahmad, Panos Kostakos, Abhishek Kumar, Mika-Petri Laakkonen, Xiaoli Liu, Zhi Liu, Le Nguyen, Huong Nguyen, Basak Ozan Ozparlak, Ville Pietiläinen, Susanna Pirttikangas, Stéphan Plassart, Sampo Pyysalo, Soheyb Ribouh, Jari Rinne, Mehdi Safarpour, Alaa Saleh, Olli Silvén, Harry Souris, Xiang Su, Roope Suomalainen, Athanasios V. Vasilakos, Aleksandr Zavodovski, Qi Zhang, Peng Yuan Zhou, Alireza Zourmand

Motivation

In June 2023, the International Telecommunication Union's Radio Communication Sector Working Party 5D (ITU-R WP 5D) approved the Framework Recommendation for IMT-2030 [10], marking a significant step towards 6G. This framework, shown in a wheel-shaped diagram, outlines six usage scenarios supported by four design principles: (i) sustainability, (ii) security, privacy, and resilience, (iii) connecting the unconnected, and (iv) ubiquitous intelligence.

Including AI and Communication among the six key usage scenarios, along with the focus on ubiquitous intelligence, indicates that 6G's promise extends beyond improvements in speed and reliability. It envisions a paradigm where AI-driven autonomous systems, dynamic network configurations, and edge intelligence become standard. This direction is exemplified by the concept of "AI-native Telecom," where AI becomes central to hardware and software network components [11]. Such integration is expected to enhance network efficiency and adaptability. Moreover, combining AI with 6G is set to enable new applications and services – from ultra-reliable low-latency communication to immersive augmented realities – highlighting AI's essential role in future communication [6], [7], [8], [12], [13], [14], [15], [16].

However, moving towards this AI-enhanced networked future presents challenges. The complex nature of telecom and network ecosystems introduces difficulties, especially regarding cross-layer and seamless interoperability among AI-enabled components. Despite these hurdles, the pursuit of innovative AI solutions, including foundational models like GPTs [6], continues. This transformation requires rethinking traditional engineering approaches, leading to data-centric, knowledge-driven ecosystems [17]. In this evolving landscape, LLMs and GPTs can have a significant impact, particularly as we address the requirements and demands of 6G – a domain where meeting requirements for connectivity, capacity, latency, mobility, and reliability becomes increasingly challenging without AI [17].

AI Interconnect: A Glimpse

The AI Interconnect is a framework designed to enable seamless integration, orchestration, and optimization of AI operations within the 6G network, facilitating dynamic interaction between AI models and network components. Building on this AI-centric paradigm shift in telecommunications, our proposed AI Interconnect framework aligns seamlessly with the vision of AI-native networks. It anticipates and addresses the requirements pertinent to distributed AI operations, such as prompt processing and the effective selection and execution of LLMs. This aligns with the broader objective of fostering trust in the operation of AI frameworks, particularly in the context of the 6G architecture. Moreover, as we move forward in this direction, it is essential to holistically consider the myriad engineering implications introduced by this evolution, ensuring that we are not just technologically equipped but also strategically positioned to leverage the transformative power of AI in telecommunications.

Contextual information, network and computing elements, state machines, and feedback loops are indispensable in ensuring optimal network performance, resilience, and adaptability. Feedback loops, such as the MAPE-K (Monitor-Analyze-Plan-Execute over a shared Knowledge) model [18], allow these network elements and components to continuously monitor their performance, analyze the collected data, plan for enhancements, and execute these plans, thereby facilitating self-regulation and adaptation. Specifically, the MAPE-K model serves as the backbone of the AI Interconnect managing system, guiding it through adaptive cycles responsive to the 6G network's dynamic environment. This seamless integration fosters a symbiotic relationship where the MAPE-K framework supports and is enhanced by LLM and GPT technologies. The MAPE-K and similar models are envisioned to leverage LLMs for advanced capabilities. The ReAct ("Reason+Act") model represents a transformative approach in leveraging LLMs for agent-based actions within a specified environment [19]. At its essence, ReAct utilizes the LLM as a dynamic planner by prompting it to "think out loud". This is achieved by presenting the LLM with a comprehensive textual overview encompassing the current environment, a defined objective, an array of potential actions, and a chronological record of preceding actions and observations. Upon processing this data, the LLM generates a sequence of contemplative thoughts before arriving at a specific action. Once determined, this action is seamlessly executed within the stipulated environment, thereby establishing the LLM as a passive analytical tool and an active participant capable of decision-making and subsequent action.

Fig. 1 provides an overview of the AI Interconnect's cross-layer design, covering the cloud-to-edge continuum from devices to the core network, including fronthaul, midhaul, and backhaul—collectively called x-haul [20]. Fronthaul connects radio heads to BaseBand Units (BBUs). Midhaul links the Digital Unit (DU) and Centralized Unit (CU), while backhaul connects the CU to the core network [21].

The AI Interconnect can support learning and inference capabilities; however, this paper focuses on high-level management and orchestration of LLM-style models without detailing specific distributed learning and inference solutions. LLMs can understand, manage, and coordinate network and application behaviors based on raw data (e.g., KPIs and radio parameters), prior network knowledge (expert models), and by subscribing

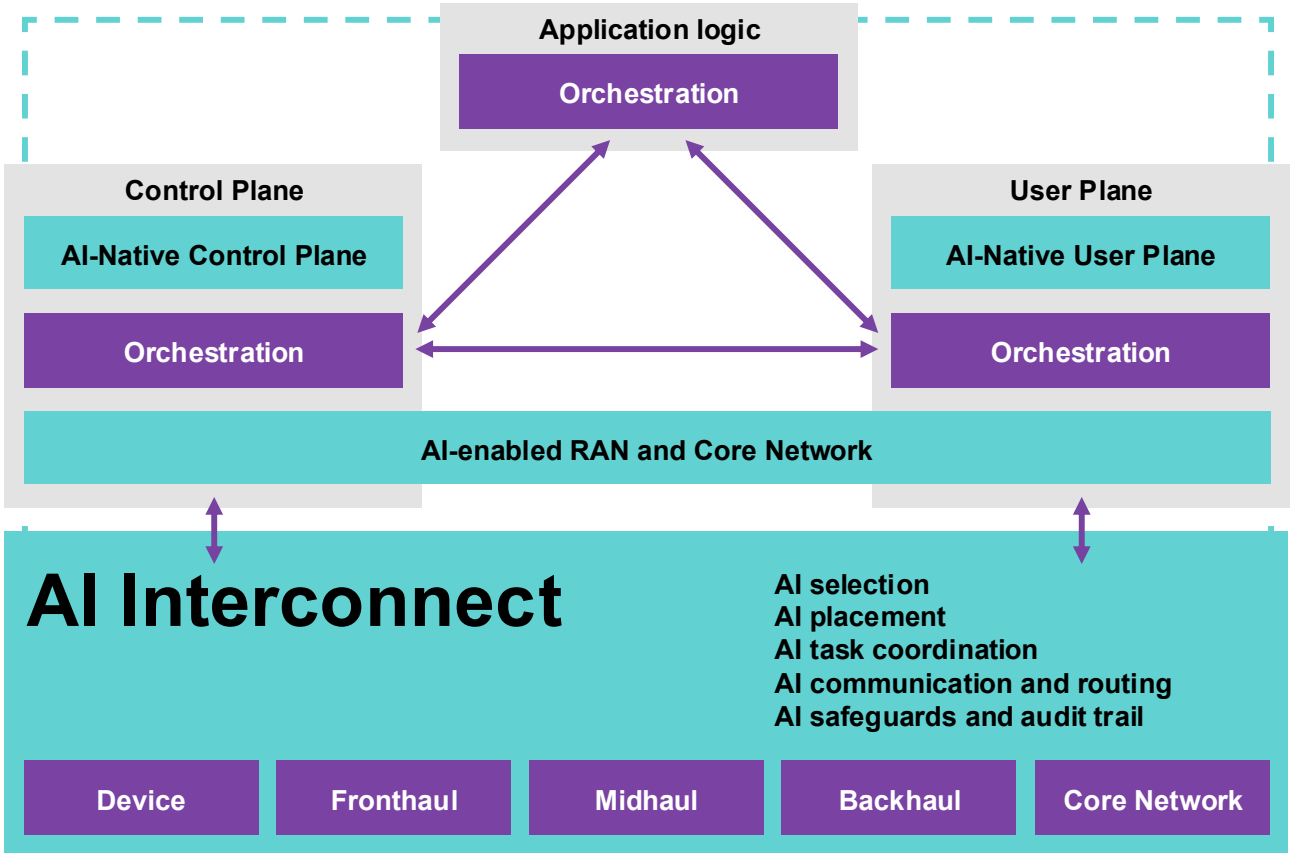


Fig. 1: AI Interconnect's cross-layer design across control, user, and application planes.

and publishing inference results through an interconnect. Expert models can support the orchestration and coordination of information and computing in the edge-cloud continuum, considering communication, energy, privacy, security, sustainability, and computing constraints. Resource efficiency is an important design consideration for the AI Interconnect.

To bridge high-level capabilities with practical implementation, the AI Interconnect provides LLM messaging and brokering for the control plane, user plane, and application logic. We envision the AI Interconnect as message-oriented, featuring both request/reply and publish/subscribe (pub/sub) APIs for requesting AI inference and other learning capabilities supported by the network. Its asynchronous, message-based nature enables reactive tasks and offers an auditing capability for LLM usage. The AI Interconnect's functionalities include AI selection, AI placement, AI task coordination, AI communication and routing, and AI safeguards with an audit trail. For a given task, the AI Interconnect selects appropriate LLMs, devises a strategy to achieve the task's objectives, and ensures seamless communication between the LLMs and network components. Its message-centric design enables effective monitoring of AI pipeline operations, which is essential for understanding system behaviors, evaluating outcomes, and conducting audits.

For 6G applications, the AI Interconnect aims to facilitate the use of LLMs for various application behaviors, including real-time content generation. Edge servers can host AI control logic (e.g., prompt engineering modules) and LLMs of different complexities. For applications, the AI Interconnect addresses the need for local and private placement of AI components, generates reports of AI operations for auditing, and accommodates various use cases. This includes general-purpose applications, network deployment and management, service-level agreements (SLAs) management, as well as generative AI (GenAI) applications and behaviors driven by user and application intent.

While LLMs hold promise for mobile and 6G networks, they will function alongside traditional ML/AI models. Some tasks in telecom may require the granularity, transparency, or specificity that conventional ML/AI algorithms provide. LLMs may not be suitable for all tasks, especially those requiring detailed analytical insights or real-time responsiveness. The AI Interconnect, with its architecture and cross-layer design, accommodates this hybrid setup, facilitating communication and cooperation between LLMs and traditional ML/AI models. Therefore, a hybrid approach that leverages both within the AI Interconnect framework may be the most practical solution.

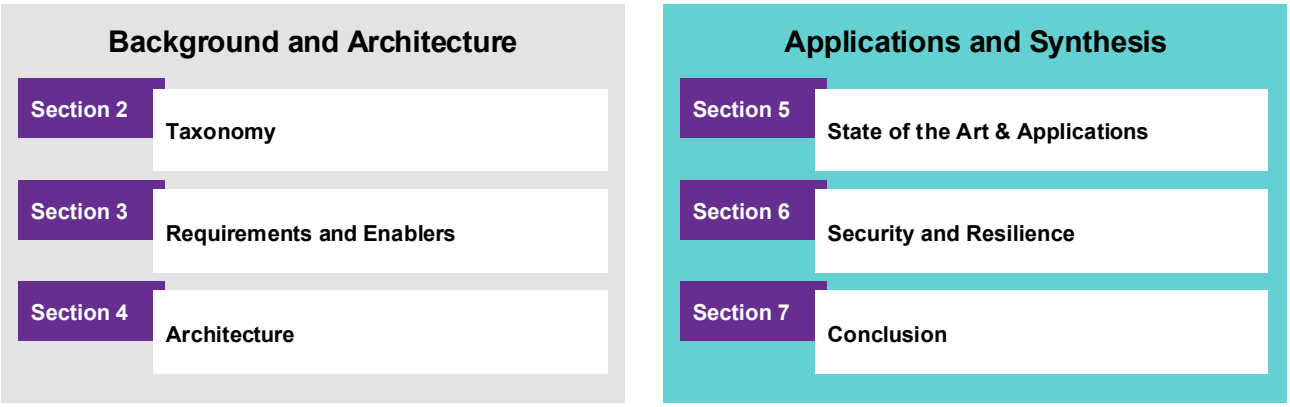


Fig. 2: Graphical representation of the paper's organization, highlighting the distinction between the 'Background and Architecture' sections (2-4) and the 'Applications and Synthesis' sections (5-7).

Paper Structure

Initially, we start with a **Taxonomy**, providing a conceptual overview of LLMs and GPTs for 6G. This leads into section **Requirements and Enablers**, which offers a strategic view of 6G integration with LLMs and GPT technologies.

Section **Architecture** lays the groundwork for understanding the architectural foundations and the progressive integration of intelligence in 5G and beyond, and then introduces the 6G AI Interconnect framework. Further, the section looks at LLMOps (the operationalization and life-cycle management of Large Language Models) in the 6G compute continuum, and further drills down into model architectures and hardware considerations.

Section **State of the Art & Applications** looks at recent studies on the topic, and further details certain application areas. Finally, before concluding, the white paper offers a view on the **Security and Resilience** of LLM usage in the 6G compute continuum.

For a visual representation of how this paper is structured, Fig. 2 depicts the delineation between the background and architecture as well as applications and synthesis within the paper.

2 Taxonomy

LLMs are expected to enhance and support 6G networks in multiple use cases and on multiple sites of its architecture [22], [23], [24].

Figure 3 illustrates a taxonomy of LLMs mapped to their potential roles in a 6G environment. This taxonomy builds on the fundamental 6G use cases and societal values introduced in [5], as well as the business perspectives for 6G described in [25].

The upper part of the figure underscores the various 6G use cases LLMs can support. Ranging from "Robot to cobots" that might involve automating processes and coordination, to the "Telepresence" that can potentially enhance real-time communication experiences. These use cases touch on hyper-connected resilient networks, sustainability, and other evolving use cases that 6G aims to address. Contrastingly, the lower part of the figure sheds light on the network architecture enablers where LLMs might find an application. With components like "Security and Privacy" emphasizing the importance of safe communications, to "Cognitive Network Management" which may automate and optimize network configurations, it is evident that LLMs can play an important role in several key network operations.

Central to this taxonomy is the "LLM Controller" It serves as a nexus, connecting both the use cases and the network enablers. This suggests that while there may be models explicitly designed for particular domains (e.g., "Green-GPT" for sustainability), there is also envisioned a role for a central controller GPT—possibly supported by additional AI components. Such a component would be instrumental in managing these expert models, fostering cooperation among them, and ensuring seamless integration.

Furthermore, these LLMs are not just restricted to high-level operations. We envision their applicability extending to finer-grained models and specific downstream tasks across different ISO layers, such as the physical (PHY) layer for tasks like beamforming, the network (NET) layer for power management, and other layers for handovers, as discussed in [23]. As such, the potential for LLMs in a 6G ecosystem is vast, bridging the gap between abstract use cases and the concrete architectural components that enable them.

To exemplify this, consider the increasing possibilities given by Multimodal Large Language Models (MLLMs). Tasked with processing diverse data types, MLLMs exemplify the versatility and adaptability of LLMs [26]. These models can help addressing challenges such as data heterogeneity, semantic ambiguity, and signal fading, cementing their role as robust task-solvers [27]. Moreover, MLLMs underscore the notion of seamless integration and cooperation among models, as they often function in tandem with unimodal LLMs and other AI components. The duality of their operation, handling both high-level tasks and delving into the intricacies of finer-grained operations, is evident. While traditional LLMs primarily cater to natural language processing (NLP) tasks, MLLMs embrace a wider operational spectrum, enhancing user interaction and communication flexibility with machines [27]. Yet, it is important to note that the journey of MLLMs is still in its initial stages. Despite their transformative capabilities, they often deal with the constraint of input-side multimodal understanding, limiting their ability to produce content across diverse modalities. Nevertheless, emerging systems like NExT-GPT aim to redress these limitations, introducing a new set of comprehensive MM-LLM systems suitable for both multimodal understanding and generation [28].

While much of the current practical efforts concerning MLLMs are directed towards multimodal signals, including text, audio, image, and video, these are not directly aligned with telco-specific needs. This high-

lights a pressing need to amplify research and development in this area, as highlighted in [23]. Nonetheless, Figure 4 provides a glimpse into how such integration could potentially operate within the telecommunications context, illustrating potential applications of how MLLMs might process multimodal data [29].

While Figure 3 offers a conceptual overview of the interplay between LLMs and the broader 6G ecosystem, a more granular perspective is necessary to fully appreciate the potential of these models in this space. The subsequent table, presented in Figure 5, lists scenarios where LLMs can be transformative compared to the current state-of-the-art. In fact, while traditional ML/AI methodologies have made significant strides in these domains [14], [30], LLMs, given their vast training data and swift adaptability, can potentially offer more nuanced solutions. These models possess the capability to quickly discern patterns from massive datasets and adapt to new information at an unprecedented scale [31]. This makes them particularly suited for dynamic scenarios, like those associated with 6G, where real-time adaptability and extensive knowledge bases are paramount. For instance, in scenarios like 'Near real-time Analytics' and 'Network Security', the advantage is not merely about processing data quickly – it is about the depth of insight and foresight that models like GPTs can provide because of their comprehensive training. The table delineates how LLMs can be leveraged across these 6G scenarios, underlining their potential transformative impacts.

However, it is important to understand that this table touches upon just a fraction of the myriad applications LLMs can assist with in a 6G environment. As we advance further into this technological era, it is likely that the unique capabilities of LLMs will find relevance in even more areas (as highlighted in [22]), some of which we might not have even envisioned yet.

Building on this, it is essential to categorize the different types of LLM models available, as they can vary based on their scope, application, and accessibility. We delineate four principal LLM categories, with models potentially being open source or reliant on closed APIs. LLMs are anticipated to be accessible via standardized embedding formats and open APIs, such as those offered by OpenAI [32] or LangChain [33]. The four categories are:

1. **Foundation:** A universal LLM model that is not tailored to a specific domain. Although foundational, this general model can be enhanced through prompt programming and fine-tuned for specific use cases.

2. **Specialized:** An LLM model tailored for a particular user or application, available in either closed or open-source fashion. Examples of its use include chat applications and instruction generation tools. It can also be adapted to environments with limited resources.

3. **Hybrid:** This LLM model combines the broad, general knowledge of the foundation model with the specialized expertise of the specialized model.

4. **Controller:** Driven by LLM-based autonomous agents, this model functions as a controller, linking other models and system components. It operates either autonomously or under human oversight.

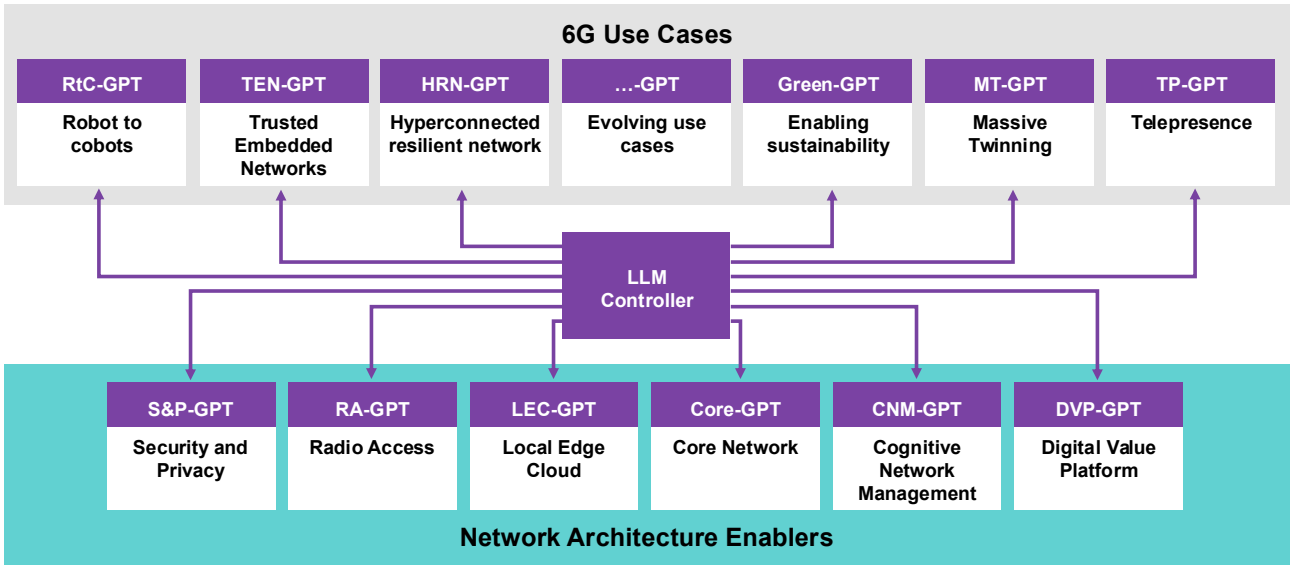


Fig. 3: Taxonomy of LLMs for 6G, mapping the potential applications of LLMs to both 6G use cases and underlying network architecture enablers. The LLM controller serves as a hub, suggesting the cooperative and management roles LLMs can play in a 6G ecosystem.

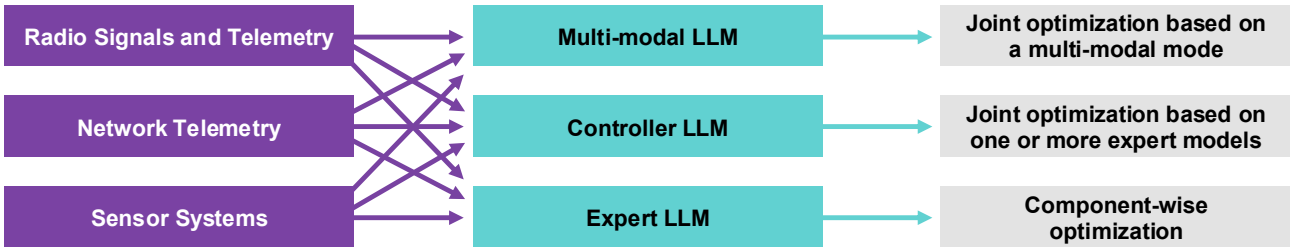


Fig. 4: Illustration of three LLM processing approaches for multimodal data. The scenarios include inputs from radio signals and telemetry, network telemetry, and sensor systems. The depicted approaches are: a multimodal capable LLM, a controller LLM for processing and result aggregation, and component-wise LLMs without aggregation.

Scenario	Description	LLM/GPT Support	Potential Impact
Near real-time Analytics	Immediate processing and analysis of data at its source	Rapid, real-time adaptation to incoming data streams for more accurate and immediate pattern recognition than traditional models	Enhanced, more accurate decisionmaking processes and a user experience tailored to real-time data trends
Interoperability	Connecting systems across formats, protocols and APIs	Natural language capabilities to interpret and generate diverse API calls, facilitating better protocol intercommunication	Seamless, frictionless communication between diverse systems, reducing system integration challenges and improving efficiency
Distributed AI	AI operations spread across multiple nodes	Swift multi-node communication and distributed reasoning and intent, leveraging vast training to infer across varied data sources	Superior resource allocation due to dynamic node coordination, resulting in faster, more efficient AI operations
IoT Device Management	Managing vast arrays of interconnected devices	Real-time data processing to handle device configurations and adapt to device behavior on-the-fly	Improved device performance and health, leading to prolonged device life and better user experience
Network Security	Protecting network from threats and anomalies	Advanced pattern recognition from a vast range of data, making anomaly detection more accurate and early	Proactive threat neutralization, minimizing vulnerabilities and ensuring more robust network integrity
Content Caching	Storing content closer to the user to reduce latency	Predictive modeling based on vast internet content understanding, forecasting user content needs more accurately	Personalized user experience with instant content delivery, reducing wait times and enhancing user satisfaction
Dynamic Spectrum Allocation	Allocating bandwidth dynamically based on the need	Deep traffic pattern understanding, predicting bandwidth needs based on diverse internet use-cases	Optimized bandwidth distribution, minimizing wastage and ensuring highquality communication experiences
Augmented Reality (AR)	Enhancing AR experiences with real-time data processing	Rich content understanding allows for generating and adapting AR data in diverse real-world scenarios	More realistic and responsive AR environments, offering users immersive and dynamic experiences tailored to their context
Autonomous Vehicle Coordination	Coordinating self-driving vehicles in real-time	Advanced predictive capabilities based on understanding a wide range of driving scenarios and conditions	Safer, more efficient driving routes and strategies, potentially reducing accidents and improving traffic flow
Smart Grid Management	Managing electricity distribution in real-time	Advanced predictive capabilities based on understanding a wide range of driving scenarios and conditions	Enhanced grid reliability, fewer outages, and optimized energy distribution catered to real-time needs

Fig. 5: Overview of potential LLM use cases in 6G. The table highlights the capabilities and potential impacts of leveraging advanced language-based models, distinguishing them from traditional ML/AI approaches.

3 Requirements and enablers

The emergence of 6G with the integration of LLMs has the potential to establish new standards for efficiency, flexibility, sustainability, and adaptability, even creating new service categories for different stakeholders in the entire 6G framework. According to our holistic view, the 6G architecture can be sectioned into four layers: *(1) strategic*, *(2) logical*, *(3) operational*, and *(4) implementation*. Within this framework, we need to deploy proper requirements for each layer, with attributes dealing with service quality, operating performance, development capability, consistent lifecycle management, and hardware configuration & software optimization, each with their contributions to requirements of implementing the different variants of LLM technology models into 6G architecture, as depicted in Figure 6. Such requirements can be clustered via the level of abstraction, purpose, and stakeholder role (e.g., management, architects, operators, developers, and users).

In this framework, each layer serves specific functions in the 6G ecosystem, culminating in the implementation layer, which stands as the execution platform. In fact, the implementation layer stands out due to its tangible, practical nature and requirements, whereas the others provide a more high-level, theoretical framework. Being the most practical of the lot, this layer serves as the execution platform for the strategies outlined in the preceding layers. In the upper layers, to understand the potential of adopting and adapting suitable LLM technology variants in 6G, we propose three key requirement clusters for each purpose: A) *Strategic requirements* that serve the highest-level guidance and regulation to the entire 6G architecture, mainly discussing service quality and operating performance issues; B) *Functional and Non-Functional requirements* that provide directions for the operational behaviour of LLM. Functional requirements dictate the "perceived experience of the users of utilities", while Non-functional requirements encompass over 20 attributes related to "non-visible experience, maintenance, and quality guarantees". These requirements support over-the-life-time end-to-end operations, configuration, and maintenance. The third cluster offers C) *Sustainability requirements* addressing the standard 6G implementation while ensuring a consistent life cycle and continuous development capability for the entire 6G fabric. A holistic view of sustainability in 6G emphasizes energy efficiency, sustainable usage (e.g., maintainability, reconfigurability), resource-efficient implementation, eco-friendly disposal, and adaptable design. These principles ensure sustainability is embedded across the lifecycle, addressing both environmental and operational goals.

Highlighting their interplay, all the requirement sections contribute to the integration of LLM into 6G. Each offers unique roles and implementation levels with a functional purpose that adds significant value to operational performance targets in 6G. This supports the evolution of 6G architecture to be flexible and adaptive to situational contexts, proximity requirements, and end-user needs, among others. To harness the potential of LLM technology role, implementation, value, and suitability across the four distinct layers of the 6G architecture, it is imperative to analyze each layer's cumulative effect within its respective level of abstraction (whether *strategic*, *logical*, *operational*, or *implementation*).

1. **Strategic Layer:** Conceptual, focusing on integrating LLM capabilities into high-level decision-making processes and predictive analytics tools. At this layer, LLMs provide high-level direction and support to the entire 6G architecture. Their predictive analytics capabilities can, for example, forecast network structure and growth, adapt the network to user behavior, and adaptively configure the network, ensuring it remains agile and obeys e.g., regulatory requirements and SLA guarantees.

2. **Logical Layer:** Focused on being integrated into network design tools, this layer ensures a seamless data flow and desired network configuration. Upholding integrity and control over the entire network behavior presents intricate, context-sensitive challenges that continually adapt with user evolution. Here, LLMs can be calibrated to refine data flows, ensuring efficient communication, security, and effective distribution of applications, tasks, and resources. This meets both "Functional" (driven by users) and "Non-functional" (driven by architectural quality) objectives. Concurrently, LLMs can be tasked to fine-tune diverse network structures and configurations, perpetually evaluating their security and efficacy. This balances 6G's capabilities between users and their operational demands. Merging these dual capabilities guarantees an ever-efficient, adaptable, and safeguarded network.

3. **Operational Layer:** This layer is dedicated to overseeing the development and lifecycle of applications, AI, maintenance protocols, security frameworks, and threat detection mechanisms, all while maintaining a real-time monitoring focus. It provides continuous and adaptive support to the real-time capabilities of 6G. Guided by the insights from the Logical layer, LLM operational frameworks and runtimes can, for example, monitor, control, and prioritize network traffic. They can pre-emptively pinpoint bottlenecks, define protocols for different users based on their application and service profiles, and adapt contextually to 6G SLA performance standards. This is achieved by channeling resources, decentralizing applications, interfacing with different AI capabilities, dynamically sharing capabilities at different levels of networking, and orchestrating timely reactions to disruptions or emergent requirements.

4. **Implementation Layer:** Integration into software development and optimization tools, as well as inclusion in hardware analysis and configuration tools. LLMs can be tailored to bolster software enhancements across 6G network elements, ranging from edge devices to assorted network gear, including base stations and Open Radio Access Network (O-RAN) [34] modules. The aim here is to ensure a "continuous seamless integration" of diverse applications and services. Within this layer, the role of LLMs emerges as a vital asset, prepared to be

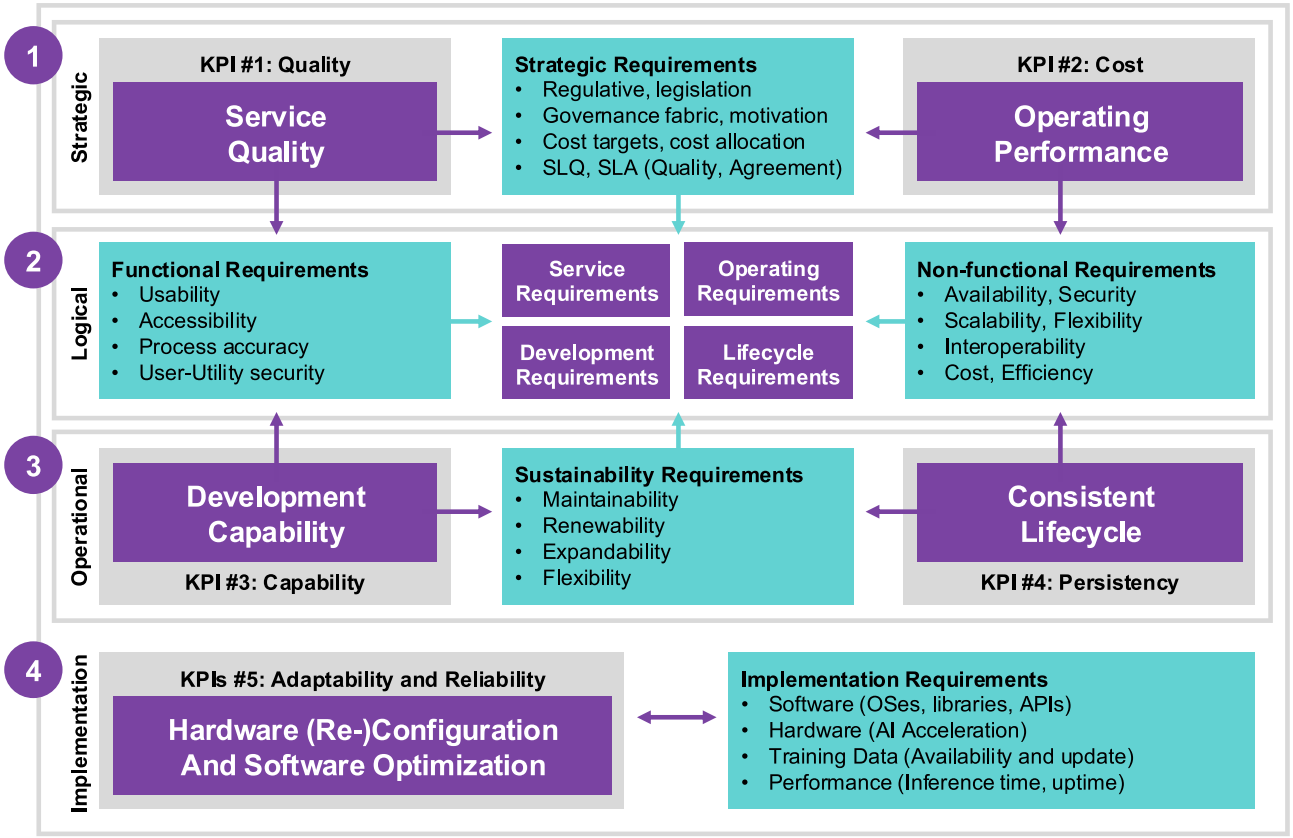


Fig. 6: A holistic representation of the 6G architectural framework, illustrating the layered approach from high-level strategic objectives to practical implementation considerations. This framework integrates LLM and GPT technologies across strategic, logical, operational, and implementation layers, with associated key performance indicators (KPIs) and requirements.

"trained for purpose". LLMs can perpetually assess hardware components, guarantee optimal setups rooted in behavioral insights, and ensuring security and compatibility, provisioning reliable and adaptable operations.

The central convergence in the Logical layer highlights the vision of translating the strategic vision into logical steps that can be operationalized. It serves as a bridge, ensuring that strategic objectives are aligned with operational capabilities. This requires a thorough understanding of the service's functional, service, development, operating, and lifecycle requirements and their interrelationships. By focusing on this convergence within the Logical layer, one can ensure that strategic aims are accurately reflected in the system's operational design, leading to a service that is not only strategically aligned but also operationally viable and efficient.

While LLMs can enhance the AI-native capabilities of 6G systems, it is important to recognize and address their inherent limitations to harness their capabilities fully.

Limitations of LLMs: Need for safeguards and human-in-the-loop

As LLMs increasingly influence various sectors of modern society, including mobile networks, understanding their limitations becomes fundamental. This section delves into a subset of pressing concerns associated with the deployment of LLMs, including trustworthiness, resource constraints, and the extent of automation. While these represent key areas of focus, it is worth noting that there exist additional nuances and risks that further underline the importance of a balanced approach in utilizing such technologies.

Ensuring the trustworthiness of AI has become a paramount concern in the progression of technology, especially as AI systems are increasingly integrated into critical sectors of society [35]. Ethical and transparent AI functionalities are essential for societal acceptance and the practical functionality and reliability of AI applications. A leading reference, the European Commission's "Ethics Guidelines for Trustworthy AI", emphasizes the significance of ensuring AI systems are lawful, ethical, and robust from both a technical and social perspective [36]. Furthermore, the EU's proposed AI Act [37] aims to establish a risk assessment-based regulatory framework that ensures AI practices in the EU adhere to high safety standards and respect fundamental rights. As AI continues its trajectory of profound influence on global societies and economies, establishing its trustworthiness through rigorous standards and ethical considerations becomes indispensable.

Resource constraints can often hamper the efficiency and capability of deploying ML models. This is particularly true when dealing with large-scale data sets and complex computations, which demand significant memory and processing power. However, advances in hardware acceleration technology, such as graphics processing units (GPUs) and tensor processing units (TPUs), can significantly reduce computational time and allow for more complex modeling [38]. Moreover, ML methods, including dimensionality reduction and model optimization, can help mitigate some operational concerns. These techniques can make models more efficient and less resource-intensive, enabling sophisticated computations even on resource-constrained systems [8], [39], [40].

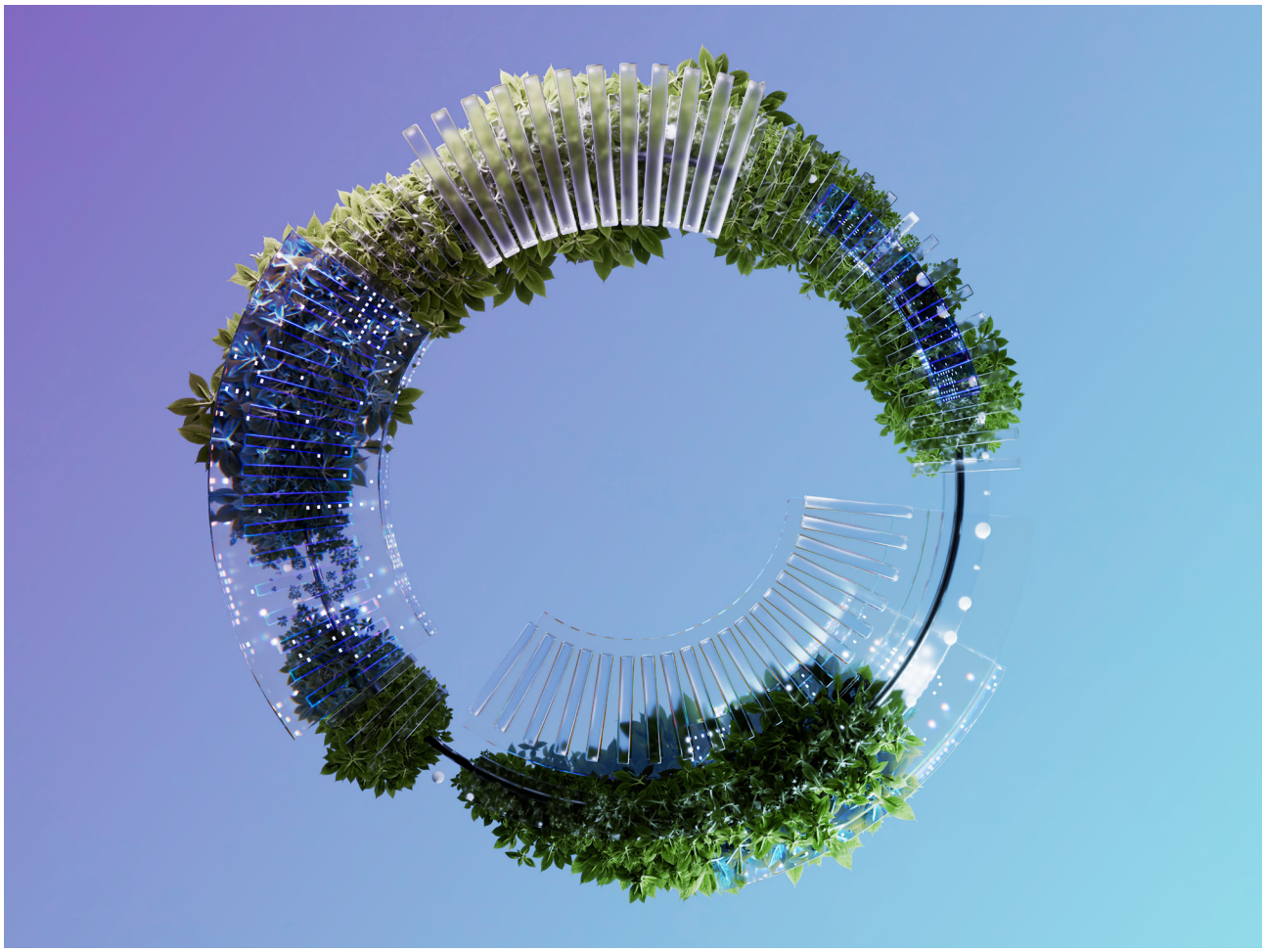
While automation can significantly improve the efficiency of network operations, it is essential to recognize that there may be inherent limits to the extent to which these processes can be fully automated. Complex tasks often involve variables and considerations beyond the capacity of current AI and ML technologies. Moreover, unforeseen anomalies, exceptions, or crises might demand human judgment and decision-making. This is where the concept of "human-in-the-loop" solutions comes into play. This approach ensures that while most routine operations can be automated, there remains a human element for oversight, management, and control. The human operator can provide the nuanced understanding, context awareness, and problem-solving abilities necessary to handle complex or unexpected situations. This balance between automation and human intervention can optimize operational efficiency while ensuring the network's robustness and reliability.

Beyond these limitations, it is also essential to touch upon the energy and environmental implications of deploying such advanced models. While this is a critical topic, we will provide an overview rather than an in-depth analysis, as this paper's primary emphasis is on the high-level architectural, management, and orchestrating aspects of LLM-enabled systems.

Energy and Environmental Implications

The upcoming era of 6G communication and computing systems promises not only a exceptional influence on global growth, productivity, and societal functions but also intersects notably with global sustainability objectives [41]. The United Nations' Sustainable Development Goals (UN SDGs) chart a path for a future that seeks to address pressing challenges ranging from poverty alleviation and gender equality to climate change action and urban development [42]. 6G, with its impending commercial launch targeted for the 2030s, aligns closely with the timeline set for the realization of these global goals [41].

The vision of 6G, as proposed in [41], looks beyond merely offering communication services. It envisions 6G as a multi-faceted entity: a service provider aligned with UN SDGs, a granular data collection tool for indica-



tor reporting, and a foundational block for future ecosystems that align with these goals. Such ecosystems will harness the capabilities of 6G, targeting goals like smart cities, gender equality, and climate change mitigation. Simultaneously, these advancements in 6G will facilitate breakthroughs in various fields such as virtual learning and smart traveling, contributing further to carbon footprint reduction.

Two complementary approaches need to be considered when introducing sustainability in 6G. First, 6G for sustainability, which focuses on how 6G can support the UN SDGs across economic, social, and environmental dimensions. Second, sustainable 6G, which addresses making 6G systems themselves sustainable. This holistic view extends beyond energy efficiency to include sustainable design, implementation, usage, and disposal. These dimensions collectively ensure 6G systems contribute meaningfully to sustainability goals while being inherently sustainable in their lifecycle [43].

Within this framework, according to [44], energy efficiency stands as a paramount design criterion for the 6G framework. The network's performance is intrinsically tied to the energy availability across its architectural domains. This focus on energy efficiency is further echoed in the Hexa-X (concluded) and Hexa-X II (ongoing) European 6G flagship projects, which target both energy efficiency and the CO^2 footprint of network infrastructure as core challenges to be addressed [45]. The importance of this endeavor lies in the fact that ICT technologies, which 6G aims to revolutionize, have a significant carbon footprint on communication networks and wireless terminals. Addressing this will not only reduce the environmental impact but also foster a wider adoption of these technologies in everyday life. In turn, this adoption can lead to optimized operations in sectors like agriculture, transport, and environmental monitoring [45].

Given the stringent sustainability and efficiency prerequisites of 6G systems, integrating resource-intensive technologies like LLMs demands careful attention. Therefore, it becomes paramount to scrutinize the energy and environmental footprint of the tools steering the 6G advancements. Central to this is the role of LLMs, which intriguingly position themselves as both potential contributors and mitigators within this intricate ecosystem.

In this respect, with growing consciousness regarding the environmental impact of technological advance-

ments and the requisite sustainability of AI [46], there has been a noticeable shift towards acknowledging and addressing the resource consumption and carbon footprint intrinsic to the lifecycle management of LLMs. A notable step in this direction has been taken by Meta, who disclosed the electricity consumption and carbon footprint of their LLaMA models [47]. This action aligns with a broader trend, where other technological giants have also unveiled detailed analyses concerning the energy and carbon footprint of prominent models such as Pathways Language Model (PaLM), GPT-3, and Evolved Transformer [48], [49], [50].

Analyzing the figures released for the LLaMA model training process, a massive computational undertaking involved utilizing *two thousand forty eight* 80GB GPUs for an estimated five months [47]. This extensive operation consumed around 2,638,000 KWh of electricity, analogous to the yearly consumption of 1,648 average households in Denmark. The process emitted about 1,015 tonnes of carbon dioxide equivalent (tCO2e), comparable to the annual carbon footprint of 92 Danish citizens.

What sets apart the reporting approach adopted in [47] is its encompassing methodology. Instead of limiting the reporting to the final stages of model training, the entire computational journey, including experimental and unsuccessful runs, has been accounted for, reflecting a comprehensive view of the environmental cost of ML and LLM development. This methodology aligns with the Operational Lifecycle Analysis (OLCA) for ML presented in [51]. By tracking emissions from the nascent exploratory stages to the deployment of the final model, a holistic understanding of a model's environmental footprint emerges. This view is particularly salient as it captures emissions often overlooked by standard metrics.

Many optimization strategies geared towards energy-efficient LLMs predominantly focus on refining the model's architecture or pivoting towards sustainable power supplies such as renewable energy sources. While these strategies are vital, given our focus, our interest leans towards solutions intertwined directly with mobile systems architecture at the intersection of 6G and LLMs.

The wave before generative AI saw a distinct trend of training and deploying AI models on energy-efficient platforms such as low-power CPUs, GPUs, or specialized hardware like TPUs. This direction also resonated with efforts to enhance various components of the 5G network [44]. Such hardware specializations not only lead to significant reductions in electricity consumption during the training and inference phases but also contribute to reduced carbon emissions. The development of telco-specific LLMs is likely to benefit immensely from such optimized hardware. Furthermore, much like the emergence of hardware specifically tailored for certain AI tasks, like vision processing units (VPUs) and TPUs for computer vision [52], we anticipate the evolution of hardware specialized for LLM execution in the coming years.

From another perspective, the dynamic and heterogeneous nature of mobile networks necessitates a flexible approach to computational resource allocation. This becomes particularly relevant when considering LLMs instances. Guided by real-time workloads and accuracy benchmarks, LLMs should efficiently scale their resource demands to ensure optimized energy usage and minimized carbon footprints. This dynamic resource allocation, coupled with strategies such as caching, memorization, and incremental training, can effectively minimize redundant computations, enhancing overall operational efficiency, also within the 6G landscape [53].

Reflecting on all these approaches and emerging directions, it is necessary for the telecommunications sector to embrace comprehensive and transparent methodologies in computing and reporting the environmental footprints associated with the deployment of AI-native network infrastructures, particularly focusing on the integration of LLMs within the 6G landscape.

Decision-Making

A decision-making process is a structured approach to identifying and selecting the best course of action from a set of available options. A typical decision-making process involves several steps, such as gathering information and evaluating available options, that help ensure thoughtful and well-considered choices.

This generic decision making process applies similarly when the "decision agent" is human (or a group of humans) and when the decision agent is "artificial" (or a group of artificial decision agents). AI has shown already remarkable capabilities in supporting decision-making processes across various domains [54], [55]. In real-time applications of complex cyber physical systems, where responsiveness and accuracy are crucial, the integration of AI decision agents offers immense potential and represents a fundamental enabling technology for advanced applications.

Moreover, the evolution of decision-making has witnessed a gradual shift from human-centric approaches to the seamless integration of AI capabilities. This transition can be characterized along two axes: the expansion of AI agent capabilities and the evolution of AI's role in decision-making. Initially, an AI agent learns from past experiences, gradually refining its performance. As it expands its goals and timescale, the agent takes on new tasks, increasing its scope of responsibility. Concurrently, AI evolves from providing decision support, where it aids human decision-making, to decision augmentation, where it significantly enhances human

capabilities. Ultimately, this trajectory leads to autonomous decision-making, where AI independently manages long-term goals and complex tasks, minimizing human intervention [56]. Together, these axes illustrate the transition from supportive to fully autonomous AI systems, capable of handling increasingly complex and extended objectives.

LLMs seem particularly suited to play an active role especially while dealing with high level decision problems involving multiple criteria and potentially complex decision scenarios [57]. Such LLM-based AI decision agents may be integrated with Internet of Things (IoT) devices and edge computing infrastructure, enabling real-time decision-making at the edge of the network. This could lead to faster response times and greater efficiency in various applications.

Accordingly, we envisage dedicated 6G LLM based agents, continuously learning and adapting to the ever-evolving demands and complexities of the future 6G network environment and of the IoT-Edge-Cloud Continuum, ensuring optimal performance and efficient decision-making across a wide range of applications. The agents act as intelligent coordinators within the network (e.g. the orchestrators of Figure 1 and the LLM controller of Figure 3), interacting and bridging the gap between other specialized and dedicated Decision Agents [58] and AI solutions (e.g. dedicated ML/AI solutions across the network and/or at various levels of the network). This collaborative approach leverages the strengths of different AI models to create a robust and comprehensive decision-making ecosystem.

Controlling the quality of decision-making by such LLM-based agents starts by defining their goals. We need to establish guidelines and standards that specify the expected decision quality, constraints, and requirements, and define performance metrics aligned with desired outcomes. Finally, we need to implement solutions to regularly monitor these metrics to assess the agent's performance (see KPIs in Figure 7).

It is worth noting that the increasing applications of AI, and the increasing role of LLMs and GenAI, is prompting a reevaluation of decision-making processes, especially in scenarios requiring the management of multiple objectives with varying levels of criticality. This shift is driving a paradigm change in how complex problems are approached. Applications in 6G networks can benefit from insights gained in other sectors, which are increasingly highlighting both challenges and opportunities [59], [60], [61], [62], [63].

Regulation

Lewis Mumford envisioned an *"invisible city"* where physical presence in a city center would become unnecessary [64], a concept increasingly feasible with the advent of 6G. By the 2030s, 6G is expected to enable ultra-low latency communication, where even human senses could be transmitted, fundamentally altering the fabric of urban life and reducing population pressures in city centers [65]. Unlike its predecessors, 6G will integrate AI systems, including large language models (LLMs), as a core component of the network, enabling seamless human-network interaction through advanced telepresence, digital twinning, and smart cities . These developments necessitate that LLMs operate in alignment with human goals, a concept referred to as "beneficial AI" [66].

The introduction of 6G, particularly its use of AI and LLMs, presents significant legal challenges, especially concerning data privacy and cybersecurity. Effective regulation of LLMs within 6G networks must establish a framework for their responsible use, ensuring these technologies are trustworthy and aligned with societal needs [67]. The concept of "tactile regulation," which adapts in real-time to emerging risks and leverages AI tools for compliance, may be essential in this context. For example, the European Union's AI Act [68] introduces mechanisms for regulatory compliance in high-risk AI systems, a model that could be expanded to 6G.

Regulatory frameworks must also address the legal status of data within the 6G era. Data should not be treated merely as property, given its critical role in achieving sustainability and privacy goals. Trust in the 6G network will depend on individuals having control over their data, necessitating regulations that empower users and scrutinize data brokers effectively [69]. Furthermore, traditional data classifications (e.g., sensitive vs. non-sensitive) may become obsolete in 6G, where massive, complex datasets will include cognitive and behavioral information. As the Ada Lovelace Institute highlighted in their 2020 report "The Data Will See You Now," these novel data points require an updated approach to data protection regulations for a sustainable framework [70]. The regulatory oversight of LLMs and AI within the 6G compute continuum must thus be dynamic and adaptive, addressing both ethical standards and real-time challenges. Such a framework would not only ensure the protection of users and their data but also foster innovation and trust in this transformative technological landscape.

Scenario	Quality KPIs	Cost KPIs	Capability KPIs	Persistency KPIs	Adaptability and Reliability KPIs
Near real-time Analytics	• Accuracy of anomaly detection • Timeliness of insights generation • Completeness of data analysis • Real-time data processing rate • Capacity of Out-of-distribution (OOD) reasoning	• Resource efficiency for analytics processing • Reduction in data storage requirements • Time to process and analyse data for near real-time analytics	• Volume of data processed per unit time • Ability to handle diverse data sources • Precision and reliability of predictions and insights derived from real-time data analysis	• Uptime of analytics engine • Accuracy, completeness, and consistency of the data used for analytics, ensuring high-quality insights and decision making	• Ability to adapt to evolving network behaviour • Resilience to data anomalies and errors • Scalability, to handle increasing data volumes and analytics demands
Interoperability	• Compatibility with various protocols • Seamless data exchange success rate • Number of successfully integrated devices and platforms • Reduction in compatibility errors and frequency and severity of errors or failures during protocol conversion • Time taken to translate and convert protocols	• Minimization of network overhead due to interoperability protocols • Level of compatibility achieved between diverse formats, protocols, and APIs, ensuring smooth interoperability across the IoT-Edge-Cloud infrastructure.	• Range of supported communication protocols and standards • Ability to handle legacy and future technologies • Percentage of successful conversions and seamless communication between different protocols supported by the decision agent.	• Continuous operation despite device or platform updates • Maintainability of interoperability across network changes • The extent to which the decision agent adheres to established industry standards for interoperability, ensuring compatibility across systems.	• Flexibility in adapting to new protocols and standards • Self-healing capabilities for interoperability issues
Distributed AI	• Consistency of decision-making across edge devices • Accuracy of collective network-wide decisions • Minimization of latency in distributed AI communication • Ability to distribute AI reasoning and decision-making tasks efficiently	• Reduction in reliance on centralized network resources • Message passing latency, i.e. the time it takes for messages and data to be exchanged between distributed AI nodes, aiming for low-latency communication.	• Scalability of AI operations and of processing power across the network • Ability to handle varying levels of computational resources at the edge • Consistency of decisions across Edge nodes	• Availability of distributed AI agents throughout the network • Fault tolerance in case of individual agent failures • Fault tolerance in case of multiple agent failures • Resilience to local “weaknesses”	• Ability to self-optimize AI models based on real-time network data • Dynamic allocation of tasks and resources for distributed AI processing • Efficient utilization of computational resources across distributed AI nodes
IoT Device Management	• IoT Device health and security status accuracy • Timeliness of IoT device issue identification and remediation • Availability and reliability of device connectivity information provided by the decision agent	• Reduction in manual intervention for device management tasks • Decision agent's capability to optimize energy consumption and prolong the battery life of IoT devices without compromising functionality	• Optimization of device resource utilization (battery, processing power) • Speed and accuracy of device configuration and adaptation performed by the decision agent based on real-time data and requirements	• Continuous monitoring of device health and security • Decision agent's effectiveness in managing the entire lifecycle of IoT devices, including provisioning, predictive maintenance, and decommissioning.	• Range of supported device types and protocols • Ability to handle diverse device capabilities and configurations • Progressive optimization of network resources
Network Security	• Anomaly detection accuracy • Accuracy of threat detection and classification • Response time for threat recognition and mitigation • Effectiveness of security incident response	• Resource consumption, i.e. the efficient utilization of computational and network resources by the decision agent for security monitoring and protection purposes	• Reduction in reliance on manual security monitoring and intervention • Incident prevention coordination • Incident response coordination	• Range of security threats and vulnerabilities detected • Ability to adapt to emerging security threats • Compliance with security standards and practices to maintain a secure IoT-Edge-Cloud environment	• Continuous monitoring of network activity for suspicious behaviour • Capacity to assimilate cybersecurity best practices in the IoT-Edge-Cloud Continuum
Content Caching	• Caching efficiency for content delivery, i.e. the effectiveness in caching and delivering content to users, reducing latency. • Latency reduction achieved, i.e. the extent to which latency is reduced by caching content closer to the users or edge devices.	• Optimization of content placement based on user demand and network congestion • Load balancing across caches • Optimization of local storage use	• Cache hit rate, i.e. the percentage of requests served from the cache without accessing the origin server. • Cache eviction rate, i.e. the frequency and efficiency of cache eviction or replacement for the optimal use of cache storage.	• Range of content types and formats supported for caching • Content lifecycle management support • Content Delivery Network (CDN) management support	• Ability to predict user demand for content pre-positioning • Bandwidth optimization, i.e. the decision agent's capability to optimize bandwidth usage by strategically caching and delivering content based on demand patterns.

Scenario	Quality KPIs	Cost KPIs	Capability KPIs	Persistency KPIs	Adaptability and Reliability KPIs
Dynamic Spectrum Allocation	• Spectrum utilization efficiency • Minimization of co-channel interference and spectrum congestion • Bandwidth allocation accuracy to different users or services • Ability of the decision agent to maintain desired levels of service quality through effective spectrum allocation.	• Optimization of network capacity based on real-time traffic demands • The efficient utilization of available spectrum resources by the decision agent, maximizing network capacity and minimizing interference	• Reduction in wasted spectrum resources • Response time for dynamic spectrum allocation • Capacity to predict spectrum allocation and needs for spectrum resources • Spectrum related beamforming strategies	• Range of supported spectrum frequencies and technologies (i.e. including non terrestrial links) • Fairness in spectrum allocation and spectrum utilization	• Ability to predict future spectrum needs based on network usage patterns • Interference management, i.e. the decision agent's effectiveness in mitigating interference and optimizing spectrum usage to minimize signal degradation and maximize network performance.
Augmented Reality (AR)	• Quality of AR experience (latency, jitter, resolution) • Optimization of network resources for AR data delivery • Real-time data processing rate for AR experiences	• Minimization of network impact on AR user experience • Energy efficiency, i.e. the decision agent's ability to optimize energy consumption in AR devices or edge nodes during AR content generation and rendering	• Reduction in resource overhead for AR applications • User satisfaction, i.e. the level of user satisfaction with the AR experiences delivered by the decision agent, considering factors such as visual quality, responsiveness, and usability.	• Range of supported AR applications and protocols • Accuracy of AR data generation and adaptation, i.e. the precision and fidelity of the decision agent in generating and adapting AR content based on real-world scenarios and user interactions	• Ability to prioritize network resources for critical AR data (e.g., safety information) • Seamless integration with network infrastructures
Autonomous Vehicle Coordination	• Communication latency for critical vehicle safety data • Accuracy and reliability of vehicle location and sensor data • Precision and effectiveness of the decision agent in coordinating the behaviour of autonomous vehicles in real-time • Delivery of network quality information to autonomous vehicles	• Minimization of network congestion during peak autonomous vehicle traffic periods • Response time for predicting driving scenarios and conditions, i.e. the time taken to predict and anticipate driving scenarios and conditions, enabling proactive coordination.	• Reduction in accidents and traffic disruptions • Collision avoidance rate, i.e. the decision agent's ability to successfully avoid collisions and ensure safe interactions between autonomous vehicles • Spatial awareness and navigation skills	• Range of supported connected vehicle technologies and protocols • Traffic efficiency, i.e. the decision agent's effectiveness in optimizing traffic flow, reducing congestion, and minimizing travel time for autonomous vehicles.	• Ability to predict traffic patterns and optimize network resources for autonomous vehicles • Adaptability to dynamic environment, i.e. the decision agent's capability to adapt and respond to changing road conditions, traffic patterns, and unforeseen events in real-time
Smart Grid Management	• Understanding of smart grids standards and protocols • Accuracy of energy demand and distribution forecasting • Minimization of energy waste and distribution losses • Optimization of energy generation and consumption across the grid	• Energy consumption optimization, i.e. the ability to contribute to the optimization of energy consumption in the smart grid, balancing supply and demand for the operation of the IoT-Edge- • Cloud Continuum • Support to condition based monitoring	• Range of supported renewable energy sources and distributed generation technologies • Renewable energy integration, i.e. the capability to efficiently integrate and manage renewable energy sources within the smart grid, maximizing their utilization and minimizing reliance on non-renewable resources.	• Reduction in reliance on traditional grid management methods • Ability to predict and respond to fluctuations in energy supply and demand • Effectiveness in maintaining grid stability, managing voltage fluctuations, and preventing blackouts or disruptions.	• Continuous monitoring of grid health and real-time response for managing grid operations and demand in relation with changing grid conditions and need to optimize electricity distribution in real-time. • Capacity to adapt in real time to changing grid configurations

Fig. 7: Overview of LLM KPIs in 6G scenarios. The table highlights the KPIs in relation with the use cases of Figure 5 and of the KPIs categories illustrated in Figure 6.

4 Architecture

From 5G to AI Interconnect

The evolution of 5G since its first release has not only been characterized by enhanced connectivity performance but also by a strategic alignment towards a more flexible, responsive, and intelligent architecture spanning from the RAN to the Core network. This journey included multiple advancements ranging from evolved network functions, service provisioning and orchestrated operational patterns, to the integration of AI capabilities.

However, while 5G has successfully fostered a favorable environment for the deployment and functionality of AI applications, its native architecture cannot be considered as "AI-native" [71]. In fact, while 5G architecture is proficient in provisioning AI applications, it does not inherently possess built-in AI processing or decision-making faculties. Consequently, essential AI-driven tasks such as real-time data analysis and predictive maintenance are primarily handled by supporting AI systems and functionality, introducing additional complexities and potential latency in network operations. Furthermore, 5G has embraced enhanced operational agility through the definition of diverse communication patterns and a service-based architecture, fostering a flexible and scalable network ecosystem. These elements, along with the exploration of intelligence integration, comprise the core of this section, depicting a holistic picture of the 5G evolution and its trajectory towards a more intelligent and adaptable network infrastructure.

In the 5G system architecture [72], standardized by the 3GPP, a service bus plays a crucial role in interconnecting various network functions. This service bus employs two primary communication patterns: request-response and subscribe-notify [72]. The request-response pattern is a synchronous method that requires immediate responses, such as setting up a user's connection or modifying a session. This pattern allows for direct communication between network functions, facilitating efficient data exchange. Contrarily, the subscribe-notify pattern provides an asynchronous notification capability, which is particularly useful for reacting to network or application-related events. In this pattern, network functions can subscribe to specific topics and receive notifications when other functions publish messages to these topics [73]. The decoupling of network functions, enabled by the Service-Based Architecture (SBA) [74], has contributed to the increased flexibility of the 5G architecture. In fact, it enables the network functions to be provisioned and scaled independently, fostering adaptability and resilience in the network, crucial for efficiently supporting a plethora of services and applications [75].

5G orchestration also emerges as pivotal aspect of the 5G network architecture, being responsible for the automated management of network services and resources [72]. It orchestrates various network functions and services, such as setting up network slices, managing resources, and ensuring Quality of Service (QoS) for diverse applications [76]. Network slicing–a significant feature of 5G–enable the set up of multiple virtual networks on a single physical infrastructure. Each slice can be tailored to meet the specific needs of a particular service or application, and the orchestration layer manages these slices [77]. Additionally, 5G orchestration handles the real-time allocation of network resources (e.g., bandwidth and computational power), adjusting these allocations as needed to optimize network performance. It also ensures the QoS for different applications and services by prioritizing network traffic based on factors like the type of application, the user's SLA, and current network conditions. Within this orchestrated ecosystem, the 3GPP has devised key 5G Core components such as the Network Data Analytics Function (NWDAF) and the Management Data Analytics

Function (MDAF) to bolster the network's intelligence of 5G networks [78], [79]. The NWDAF collects and analyzes network-wide information, enabling an intelligent, real-time understanding of network conditions and performance. This aids in making predictive decisions to enhance overall network performance. The MDAF serves a complementary role by focusing on collecting and analyzing management data. This includes performance metrics, fault information, and configuration data from network entities, allowing for a holistic view of the network status and facilitating data-driven network management decisions. Both NWDAF and MDAF functions are central to the 3GPP's approach to applying AI and ML techniques for more efficient, responsive, and adaptable 5G networks.

Looking at the RAN, there has been a surge in R&D efforts aimed at fostering AI-driven radio networks in recent years [80], [81], [82], [83]. In line with this trend, the O-RAN Alliance [34] emerges as a central initiative, committed to revolutionizing the RAN through open standards and architectures. By advocating for open interfaces, disaggregated network components, and a heightened focus on AI-driven intelligence and automation, the O-RAN Alliance aspires to cultivate a more innovative, interoperable, and economically efficient wireless network ecosystem [84], [85]. O-RAN operates based on two Radio Intelligent Controller (RIC) designs [86]. The Non-real-time RIC is configured to manage operations that last several seconds, making it ideal for AI/ML training and aiding service provision. Conversely, the Near-real-time RIC is optimized for operations that range from tens of milliseconds to a second, aligning well with the management of RAN control primitives and the execution of inference tasks. The Near-RT RIC facilitates the deployment of xApps, applications interfacing with the RAN elements to perform specific control and optimization functions. The O-RAN E2 interface serves as the main channel for essential data for RICs, learning, and inference processes [87]. O-RAN incorporates an AI/ML framework that significantly transforms the RAN architecture, infusing it with intelligent functionalities [85], [87], including intelligent slicing and dynamic control [88]. This framework is anchored on the RIC, housing the xApps—applications employing AI/ML algorithms to automate and optimize various RAN functionalities. The RIC, interfacing with both non-real-time (Near-RT RIC) and near-real-time RAN domains, enables a dynamic information exchange, promoting a more adaptable and responsive RAN. It uses standardized, open interfaces to foster interoperability, minimize vendor lock-in, and encourage a diverse ecosystem of innovative RAN solutions. The internal messaging infrastructure of O-RAN connects xApps, platform services, and interface end-points. While there's no specific technology prescribed (for instance, the O-RAN Software Community has delineated the RIC Message Router or RMR [89]), the system is required to meet certain standards. It should enable registration, discovery, and removal of endpoints, like RIC components and xApps, and should provide APIs for direct messaging or through pub/sub methods, guaranteeing efficient routing and data protection [90].

The emerging **AI RAN** research further emphasizes the integration of AI/ML capabilities directly within the RAN to enhance operational efficiency, support new use cases, and drive innovation in network functionalities. It is possible to determine three types of AI RAN paradigms. AI-on-RAN enables the execution of AI-driven applications at the network edge, leveraging the RAN's proximity to end-users for low-latency, high-throughput processing. This architecture supports a wide range of applications, including intelligent video analytics, augmented and virtual reality (AR/VR) experiences, and advanced positioning services. By embedding AI capabilities closer to the data source, AI-on-RAN minimizes the need for data transfer to centralized cloud servers, thereby reducing latency, improving real-time responsiveness, and optimizing network resources. The **AI-on-RAN** framework is designed to be highly adaptable, accommodating diverse AI models and algorithms tailored to specific RAN functionalities, further contributing to a more intelligent and responsive network environment. **AI-for-RAN** specifically targets the enhancement of RAN performance by leveraging AI to optimize core network operations, such as resource allocation, spectral efficiency, and interference management. By integrating advanced AI-driven control mechanisms, AI-for-RAN enables dynamic adaptation to fluctuating network conditions, predictive maintenance, and real-time optimization of radio parameters, ultimately leading to improved user experiences and more efficient network utilization. This approach transforms RAN from a traditionally static system into a self-optimizing network layer capable of making intelligent adjustments on the fly, significantly enhancing the overall resilience and flexibility of the network infrastructure. To tie both paradigms **AI-and-RAN** studies the synergistic approach where AI and RAN share the same physical and virtual infrastructure, seamlessly co-locating AI processing capabilities with RAN operations. This co-location allows for efficient asset utilization, reducing overhead costs and streamlining network management by harnessing shared computational resources. AI-and-RAN enables a unified operational environment where AI tasks such as data analysis, inferencing, and decision-making are performed in close proximity to RAN functions, allowing for faster processing times and reduced latency.

The outlined architectural advancements in 5G and beyond, particularly the introduction of analytics-based orchestration and open, intelligent RAN concepts, depict a trajectory towards more integrated, flexible, and

intelligent network architectures. These improvements form a foundational basis that seems conducive for the native, seamless incorporation and functioning of advanced AI technologies such as LLMs and GPTs in the upcoming 6G networks.

The 6G AI Interconnect Framework

The envisaged AI Interconnect framework for 6G leverages the power of advanced AI/ML models, such as LLMs and GPTs, coupled with enhanced data analytics capabilities, to comprehend and interpret the vast volumes of data traversing the network. This analysis allows the AI components to discern patterns and trends, predict network congestion, and make insightful decisions about routing data and component placement to optimize the use of radio and network resources. Moreover, the AI Interconnect can diagnose, analyze, and manage network and application tasks in real-time across the edge-cloud continuum. Central to this approach is the MAPE-K feedback loop, a fundamental architectural model in the design of self-adaptive systems and autonomous computing [18].

The MAPE-K model outlines a cyclical process enabling systems to self-manage and adapt to evolving conditions for optimized performance. It encompasses four interdependent stages: *(i)* Monitoring, where the system's environment and performance are observed; *(ii)* Analysis, where data gathered is evaluated to understand the system's status; *(iii)* Planning, where strategies are formed based on analysis results; and *(iv)* Execution, where these strategies are implemented. The shared 'Knowledge' component supports all these stages, providing a centralized information base.

In the context of 6G networks (Figure 8), the AI Interconnect functions as the "Managing System", deploying the MAPE-K loop as a strategic tool to facilitate intelligent control and adaptability in the "Managed System". Specifically aimed at optimizing quality properties essential for 6G operations, the AI Interconnect uses the MAPE-K loop to address complex network challenges. LLMs and GPT technologies, within this setup, are instrumental in enhancing the functionality of the MAPE-K loop. They are particularly significant in addressing learning problems within the broader adaptation challenges, enriching the AI Interconnect's capability to execute timely and informed adjustments in the 6G network's operations. The sophisticated processing capabilities of LLMs allow them to manage extensive datasets efficiently, offering critical insights that enhance the *Analysis* and *Planning* stages of the MAPE-K loop, subsequently fine-tuning the adaptive responses of the 6G network.

The AI Interconnect aims to empower the MAPE-K stages to leverage LLMs, and even undertake MAPE-K iterations through a singular LLM or a set of LLMs. This AI-driven interconnect approach offers several significant advantages over traditional network interconnects. Primarily, it grants the network the dynamism required to adapt to evolving conditions, such as fluctuations in data traffic or modifications in network topology. Consequently, the network can sustain optimal performance even when navigating through challenging operational terrains. Moreover, the synergy between

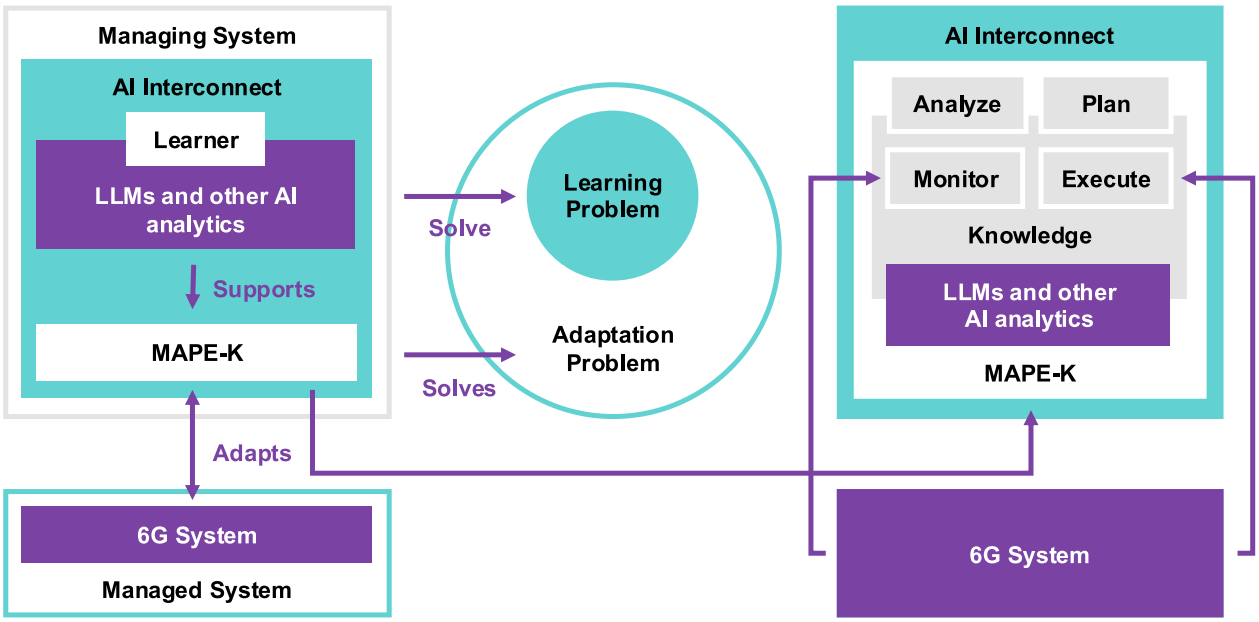


Fig. 8: MAPE-K as enabling paradigm of the AI Interconnect. On the left, relation of learning problem and adaptation problem with the components of the managing system. On the right, structure of an autonomic element. Elements interact with other elements and with human programmers via their autonomic managers.

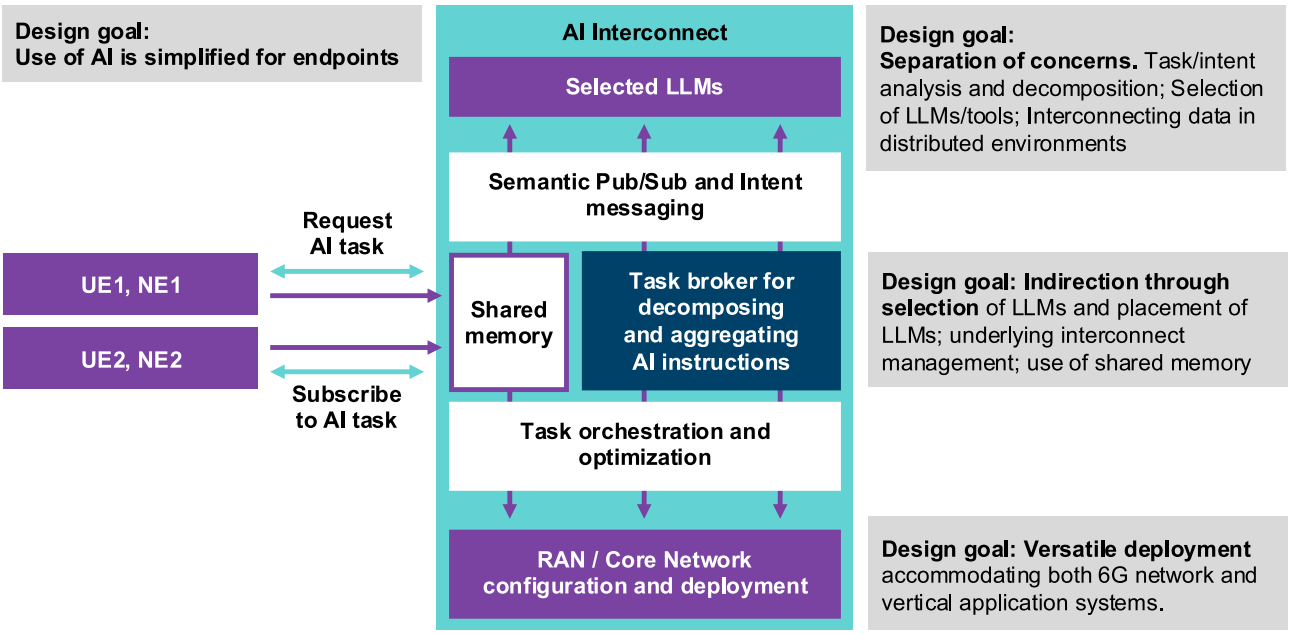


Fig. 9: GPT-Driven AI Interconnect Architecture. Our approach is structured around four pivotal design goals, from simplifying end-user interactions to ensuring versatile deployment across varied contexts, all rooted in foundational components for optimum efficiency and adaptability.

AI interconnect and orchestration emerges as a crucial catalyst in realizing Intent-Based Networking (IBN) [91]. IBNs have the capability to translate users' business objectives into strategies for network configuration, operation, and maintenance [92]. The integrated use of MAPE-K, AI Interconnect, and GPT/LLM technologies orchestrates a robust, forward-thinking paradigm for managing and optimizing 6G networks. Collectively, these elements create a dynamic, self-adapting network infrastructure, capable of effectively meeting the demands of future connectivity with unparalleled efficiency and intelligence.

Figure 9 delineates our envisioned AI Interconnect system architecture in detail. Building upon preceding concepts of on-device and in-network LLMs, our blueprint proposes an edge-cloud continuum orchestration, allowing LLMs to be strategically positioned across various network elements, ranging from the user equipment (UE) to the public cloud. In developing our system, we contemplate a foundational architecture driven by four design goals and three fundamental components. The first design goal, *simplicity at the endpoints*, advocates for a redistribution of complexity away from the network edge, resulting in a simplified end-user interface and device requirements. The second goal, *separation of concerns*, ensures that different system functionalities are divided into separate components, each having a distinct responsibility, thereby fostering modularity and scalability. Our architectural vision also embraces *indirection through selection* as third design goal. This introduces an intermediary layer optimized for system interactions, fostering a robust adaptability and flexibility in the face of varied operational conditions. *Versatile deployment* rounds off our design goals, emphasizing the architecture's inherent capability to proficiently navigate both network-level and application-level concerns, ensuring a comprehensive functionality across a spectrum of contexts and use cases.

The realization of these design goals is reflected in the mobile network, through the three components depicted in Figure 1: *(i)* the *control plane*, which is responsible for the orchestration and coordination of network-centric activities such as resource allocation and network configuration; *(ii)* the *user plane*, which handles the data transmission, ensuring the efficient and reliable management of user data packets; *(iii)* the higher-level *application logic*, which servers as the business layer that supports end-user services, analytics, and advanced applications. All these layers are meant to utilize 3GPP and O-RAN interfaces, promoting a more open and intelligent network architecture. Within this framework, the RIC, complemented by the insights derived from NWDAF and MDAF, plays a central role in orchestrating and controlling these interfaces, acting as the system's cognitive core. This enhanced "brain" navigates informed decisions about resource allocation, service provisioning management and orchestration, signal processing, and optimization pathways for network efficiency. This comprehensive design approach can ensure that our system is robust, versatile, and ready to meet the demands of future AI-native network infrastructure.

We base the AI Interconnect and orchestration design on three fundamental concepts:

- **Semantic Publish/Subscribe and Intent-based Messaging.** Our AI Interconnect is designed with a specialized focus on multi-layer semantic request/reply and pub/sub mechanisms. The interconnect is multi-layered to meet the diverse requirements of AI applications, such as the delivery of raw data, prompts and prompt fragments, inference results, and model updates. This approach ensures that endpoints are decoupled across spatial, temporal, and synchronization dimensions. At the core of this design lies a message-based interconnect system that carries more than just data; it encompasses essential AI-related metadata, providing a detailed account of AI operations and forming the basis for essential oversight mechanisms. This inherent attribute promotes accountability by keeping a detailed record of these operations, enabling a rigorous audit of AI-driven processes.

- **LLMs as Controllers.** In our architecture, LLMs are re-imagined as central orchestrators, strengthened with the expertise of specialized models and systems. They bring an element of dynamic flow control, introducing a layer of agility and adaptability to the system. Positioning LLMs centrally as controllers equips the system with the capability to expertly navigate and supervise complex AI landscapes.

- **LLMs as Dynamic Tool Builders and API Brokers.** Beyond their role as controllers, LLMs are also envisioned as dynamic tool creators. They have the capability to dynamically craft tools that are tailored to specific operational needs, ensuring a high degree of flexibility and precision in AI tasks. Furthermore, LLMs act as API brokers, facilitating the seamless interaction and integration between diverse AI tools and platforms, simplifying operations and enhancing the coherence of the AI ecosystem.

Another aspect to explore in the context of LLMs is the potential for semantic compression and communication [27], [93], [94], [95]. This innovative capability can lead to additional possibilities for efficient data transmission and interaction, paving the way for more advanced and intelligent communication paradigms within the AI Interconnect. It could potentially enhance the roles of LLMs as controllers, tool builders, and API brokers by introducing a new layer of semantic intelligence to these functions. However, a deep dive into this aspect of LLM capabilities lies beyond the scope of this paper, but it signifies an intriguing direction for future research and exploration.

Our AI Interconnect seamlessly integrates with open edge-cloud continuum APIs and execution environments, leveraging existing standards and platforms, such as O-RAN, to ensure a unified, interoperable execution framework that gracefully spans across the edge and the cloud.

LLMOps: Lifecycle Management of LLMs in 6G Architectures

Despite the undeniable potential of AI-native communication networks, the landscape of development and operation needs reevaluation to cater to the emerging technological requirements in such an environment. This section focuses on the challenges associated with the lifecycle management of LLMs (LLMOps) in environments supported by 6G. From the perspective of AI and DevOps practitioners, we offer a starting point (see Fig. 10) for discussions about where the "traditional" DevOps/MLOps lifecycle concludes and the point at which AI-human and AI-AI interaction becomes essential. Here, we will mainly discuss LLMs. We note that the consideration hereafter are nonetheless rather general and apply similarly to other ML architectures that can be considered large (billions of parameters) that rather serve a general purpose (e.g. GPTs).

Artificial intelligence permeates almost all areas of life and work. In his essay [96], Ryan Calo discusses the social challenges with respect to AI. The importance of reproducibility is highlighted in the sense of how AI was built, certification, privacy as well as AI-human interoperability. In the same context, the European Parliament has approved the world's first comprehensive framework for constraining the risks of AI (AI act [68]) by transparency and a regulated way of operating AI. In order to address the issues raised by this necessary regulation, the development, testing and operating of LLMs have been rethought. Additionally to the technological and social aspects also the business perspective on LLMs requires a reconsidered LLM lifecycle. A changing factor, for instance, is that costs for LLMs inference become a relevant factor and may be higher than costs for training (e.g. fine-tuning foundational models) and experimentation. This holds in particular when central high-performance infrastructure has to be operated and maintained for a robust delivery of LLM. In this regard training, retraining and inference can profit from distributing these tasks in the computing continuum to edge devices. Here, techniques such as federated learning [97], could help to address these new demands on a distributed computing environment. Ultimately, it is imperative to consider the legal implications that emerge from the outlined requirements at the very onset of any forthcoming development in LLMs. Reproducibility remains a pivotal factor in ensuring transparency throughout this process. Key areas where meticulous record-keeping is essential include the certification of datasets and their sources, documenting infrastructure to guarantee adequate precision, detailing the LLM algorithms employed, and establishing reproducible test

environments to evaluate the robustness of LLMs in terms of both safety and security.

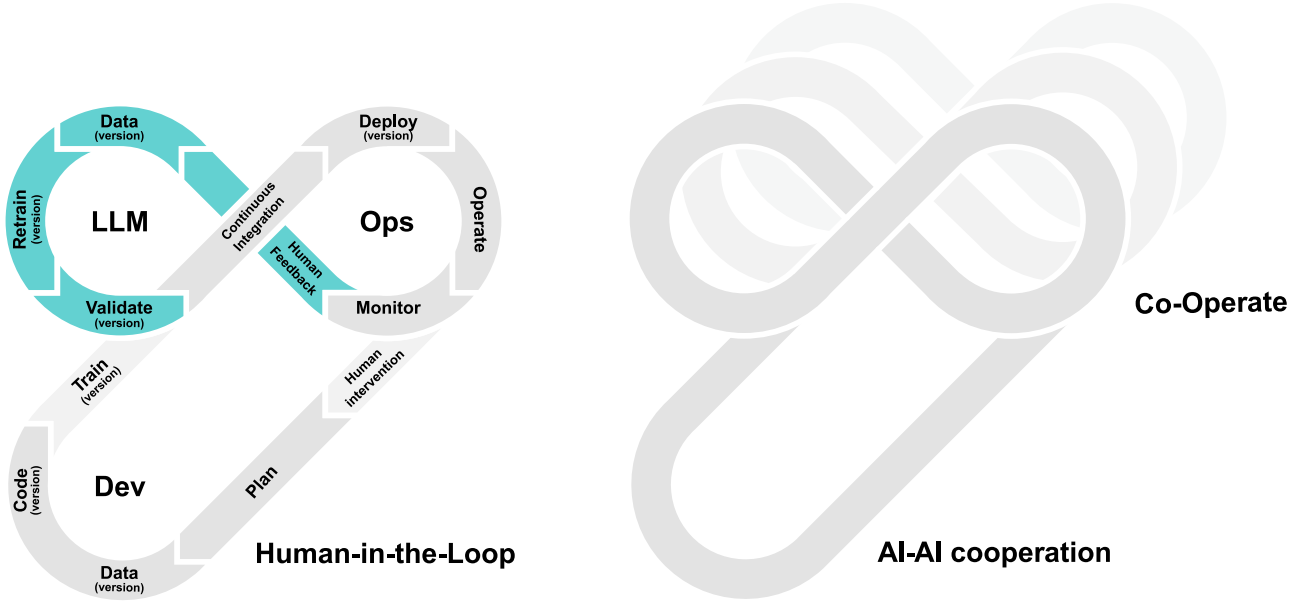
With these new challenges as described above, how could the lifecycle of LLMs look like in the future and how is it connected to the need of Reproducibility and Interoperability? To provide a viewpoint, we would like to start with the question: What is LLMOps? Simply speaking, LLMOps is MLOps (DevOps for machine learning [98]) extended to address the special needs of LLMs with respect to the deployment, operation, maintenance as well as the handling of the interaction of LLMs with other actors. What does it mean for the LLMOps life cycle in the 6G enabled computing continuum?

The regular MLOps life cycle usually starts with gathering, exploring and processing of data. In the next step, the code for the ML approach is developed and then sent together with the data to the central computing infrastructure where the training happens. After that the resulting ML model is validated and then deployed to infrastructure where it runs for the intended use case. During the operation phase the model performance is monitored and new data is acquired. Based on the monitoring and some time of operating, the ML software is reconsidered and the whole cycle begins from usually an adjusted set of the three key assets of ML models that are data, code, and infrastructure [99].

From the high-level perspective, the lifecycle of LLMs does not differ very much from the one of "classical" ML solutions. However, due to their inherent capability of direct interaction with humans (here by using language), particular things change during their life and also during their conception. These changes also concern the features of LLMs which are considered as "risky" as described above. The changes to the life cycle are sketches in Figure 10.

When we would start with the regular development of an LLM it would start again with data, code and training, see Fig. 10 (left, Dev). Once the model has been validated, it is integrated and deployed to production where it operates and is monitored. Different to the classical ML approaches, LLMs nowadays are developed to use human feedback to improve by e.g. reinforcement learning [100]. Consequently, the LLM model itself could be by construction capable of using human feedback, that extends the training data, to retrain and then operate in a fine-tuned version of itself. Hence, the human feedback lets the LLM escape the "classical" DevOps lifecycle and have its own developer-independent (Dev-idependent) life cycle which can continue independently, with a stop due to, e.g. security, or performance, or similar issues. In the latter case a human intervention is needed (the second part of the human-in-the loop) and the DevOps cycle starts again.

As described above, reproducibility and Interoperability are crucial for the trustworthiness and robustness of LLMs. Reproducibility can be only addressed by a proper tooling that allows to track the development and self-development of LLMs. In Figure 10 (left), we indicate the parts where versioning is needed. A very fine-granular and still manageable solution for versioning, particularly model versioning, was earlier discussed by Holzinger et. al. [101] to be necessary to foster reproducibility. We believe that versioning and proper track-



ig. 10: Illustration of the LLMOps life cycle in a 6G enabled computing continuum highlighting the AI-human (left) and AI-AI interaction (right) as well as places where proper versioning is needed to support reproducibility.

ing is needed everywhere where the foundations of ML/LL models, namely data, code, infrastructure, as well as the training and validation is touched. For this tooling is needed capable of bookkeeping all severe parts including those in the self-development loop. To guarantee human-AI interoperability, particularly the monitoring of LLM and the security of their interfaces are important. As described before the monitoring can either lead to human feedback or to rebuild the LLM by human intersection. KPIs (key performance indicators) and safety measures have to be implemented in order to decide when a human intervention is needed. Likewise, security and privacy measures have to be considered before using the human feedback in the self-deployment loop of LLMs. Furthermore, the human user needs to be informed when interacting with an artificial intelligence such as LLMs. Only this way, self-determined decisions on the provision of private data to the feedback loop can be done taking advantages and risks into account.

The environment allowing for cooperative AI is yet another feature offered by a 6G enabled computing continuum. Hence, many of the previously described LLMOps lifecycles can potentially co-exist and mutually benefit from each other. From a DevOps perspective, unlocking the full potential of this system necessitates AI-AI interoperability. Achieving this requires the establishment of standard interfaces between LLMs operating across diverse devices. Moreover, integration tests for these interfaces must be developed, in conjunction with traditional DevOps methodologies like test data generators, to ensure seamless AI integration. Additionally, effective version control for the models is crucial. This goes beyond mere reproducibility, facilitating a framework where different versions of LLMs are designated compatible or incompatible, thus ensuring coherent communication between them.

Distributed Inference

The success of LLMs is primarily attributed to their scalability. Distributed inference allows LLMs to scale to efficiently perform complex tasks with low latency, such as natural language understanding and text generation. LLMs are likely to have billions of parameters. The scalable inference of LLMs is accompanied by a substantial demand for computation resources (high end hardware), such as powerful CPUs, GPUs or TPUs and ample RAM. In the context of 6G connectivity, which involves devices with diverse capacities, the challenge of resource efficiency becomes crucial, especially for resource-constrained devices.

Utilizing pre-trained LMMs typically involves two stages, i.e., fine tuning and inference. We particular focus on distributed inference of LLMs on resource constrained and unstable environment, which faces significant challenges due to LLMs' new characteristics, such as huge parameters and autoregressive inference. The investigation sheds light on current approaches addressing the resource challenges posed by LLMs, with the aim of not only providing insights into existing methodologies but also potentially inspiring future scientific breakthroughs in 6G-enabled computing continuum.

Overview of distributed inference for Deep Neural Network services

Distributed inference is one crucial approach to accelerate inference for latency sensitive Deep Neural Network services. There exist many research works on how to partition DNN inference tasks, e.g., image classification, and assign them to devices with different computation capacities [102], [103]. In general, distributed inference can be realized through horizontal collaboration or vertical collaboration. In horizontal collaboration, an inference task is partitioned through data/tensor partition along one or more dimensions, e.g., grid-based spatial partition, segment-based spatial partition, and channel partition, and the partitioned tasks can be allocated to different computing devices [104]. Part of the partitioned computation outcomes shall be exchanged between distributed devices and be merged with the local computation results, then continue with the computation of next layers, which inevitably causes communication overhead. Such data exchange can take place at every layer (i.e., layer-wise parallization) or once every multiple layers (i.e. fused-layer parallization), depending on how a DNN model is sliced. There is tradeoff in communication and computation overhead in different partition and parallization approaches. Generally speaking, layer-wise parallization causes more communication overhead, while fused-layer parallization requires a bigger receptive field for input data and thus leads to higher computation overhead. An optimization problem can be formulated to find optimal data partition and model split points based on the communication bandwidth, and computation speed of different devices. To minimize communication overhead in distributed inference, [105] proposes a novel task collaboration scheme which maximizes the overlapping between communication and computation process, thereby minimizing the total inference time. In vertical collaboration, a DNN model is split among local device, Edge or Cloud, which can execute part of the inference pipeline. This is often referred to as split computing or device-edge collaborative inference [106]. The optimal model split depends on the size of intermediate feature maps, communication data rate and the computation power at device/edge/Cloud [107]. Due to time-varying wireless communication channel, to ensure stringent inference delay constraint, it is promising to leverage dynamic inference realized by dynamic neural network [108], for example, early-exits architecture [109]. While the existing methods for distributed inference for DNN could still be relevant for LLM distributed inference, it is worth mentioning that once the partition is settled, for given communication data rate and computation

resource, DNN inference time is rather deterministic, which is different from the case of LLM inference, as LLM inference follows an autoregressive patterns.

Distributed inference for LLMs

In LLM inference, it generates one token by running the model and the generated token is used as input to generate the following token using the model, and this process continues until the model generates the end-of-sequence token, which is referred to as autoregressive. The LLM inference process requires to track the states previously generated, demanding high memory capability and hinders the inference efficiency. However, in many situations device memory resources are scarce, which causes memory boundary to make it impossible to directly employ the LLMs on a single GPU or a single node with multiple GPUs. For example, Imagebind requires about 6G bytes memory, which usually exceeds the memory limit of a device. Therefore, it is crucial to investigate how to overcome the challenges to perform LLMs inference on resource limited devices especially 6G networks are expected to integrate edge computing more extensively. We present several promising techniques as follows.

Parallelism approach: Parallelism can accelerate inference process and facilitate the deployment of LLMs by making the partition of LLM and parallel running on on diverse hardware devices without accuracy loss. In current research work, parallelism techniques, such as data parallelism, tensor parallelism, and pipeline parallelism, are employed for LLM distributed inference. One challenge of implementing parallelism for LLMs is ensuring efficient coordination and synchronization among parallel process while maintaining the model's accuracy. Therefore, it is essential to take into account the diversity in hardware characteristics and transformer architectures during the parallelism process. The model structure affects the optimal parallelism strategies, such as tensor and pipeline parallelism are effective only for dense transformers, whereas expert parallelism is specifically designed for sparse transformers [110]. Meanwhile, the place where parallelism is carried on (a single node with multiple GPUs or on a multi-node) affects which parallelism methodology should be adopted.

Borzunov *et al.* [110] propose a novel distributed pipeline-parallelism inference algorithm, which can quickly recover from the failed server and transfer the task to replacement servers by using the dual attention caches. Meanwhile, they also design a decentralized load-balancing protocol where the overall system throughput is maximized by distributing transformer blocks to each server. Furthermore, participants can seamlessly include or exclude their devices during the inference process due to the decentralized feature protocol which greatly enhance the efficiency of utilizing idle GPU.

DeepSpeed-Inference [111] uses a combination of parallelism strategies to deal with Mixture-of-Experts (MOE) transformer models including both the dense and sparse transformer components. The proposed approach includes two parts: 1) a three-layer DeepSpeed Transformer system and each layer with the purpose to reduce the latency, specifically, single GPU transformer kernels are used for optimizing the memory bandwidth usage, tensor slicing and pipeline parallelism are utilized for scaling dense MOE among GPUs, and tensor slicing together with expert parallelism are employed to distribute model parameters to hundreds of GPUs to fasten the process, and 2) ZeRo-Inference utilizing the memory resources of CPU, NVMe, and GPU alongside of GPU to enable extensive model inference with limited resources. There are also other parallelism approaches such as parameter offloading and data parallelism. Data parallelism is one of the primary choices if the model can fit into a single GPU. However, that is not the common case and hence data parallelism usually is combined with other parallelism methods.

ORCA [112] implemented a distributed inference for transformer-based generative models, with scalability to models with hundreds of billions of parameters. It accelerates inference by utilizing intra-layer parallelism and inter-layer parallelism. Intra-layer parallelism, similar to tensor parallelism, basically splits matrix multiplications and their associated parameters over multiple GPUs, and inter-layer parallelism, similar to pipeline parallelism, splits Transformer layers over multiple GPUs. While ORCA significantly improves inference throughput, its iteration-level first-come-first-served processing has head-of-line blocking issue. FastServe [113] addresses this issue by using preemptive scheduling to minimize job completion time.

Model collaboration and Hierarchical Inference: In the context of deep neural networks (DNNs), collaborative or hierarchical concepts have been proposed where a smaller, local DNN interacts with a cloud-based larger DNN to handle inference [114] [115]. Smaller models can typically be placed on constrained devices closer to the edge, while they suffer from lower accuracy. Through offloading or other collaboration schemes with the cloud, this deficiency can be overcome.

In the context of LLMs, these trade-offs still exists. Small models for LLMs require fewer computing resources, which makes them a feasible approach for deployment on device with limited hardware resources. There exist challenges of using small models for inference, for example, the accuracy of using the small model on the edge to perform the inference is low if the small model do not see the testing set during the training process. Meanwhile, even that the small model see the testing set during the training process, the accuracy of the small model is greatly affected by the data

size and data quality that used to training, susceptible to overfitting or underfitting, and the accuracy is much lower than the LLM on the cloud [116]. To tackle this issue, one approach is through edge cloud cooperation to customize smaller task-specific model deployed on the device/edge by leveraging the expert LLMs located on the cloud where LLMs' predictions as pesudo labels are used to supervise the customizing process with objective to minimize the loss to pull the embedding of inputs of small model close to LLMs to enhance open set capability of the small model. The proposed system [116] has a model selection module to choose the appropriate architectures for the small models considering the profiles of the edge devices, such as memory constrains. Meanwhile, the system includes an inference engine that can control using the LLMs on the cloud or the customized small model on the edge to perform the inference task taking the data uncertainty and network conditions into consideration.

The above process of generating dynamical small models with the help of LLM could be regarded as a kind of model collaboration. The idea of model collaboration for performing LLM inference on resource constrained device is using a cost-efficient small LLM that can be accommodated within the device's memory with the help of larger LLMs and assigning majority of tokens to the small LLM for the inference tasks. A common used strategy for model collaboration is "generate-then-verify" where the small LLM is a generator and the expert LLM is served as a verifier. The strategy is also known as speculative decoding, which guarantees the accuracy by using the expert LLM as a verifier with fast verification. However, it also bring other challenges, such as overlooking correct tokens and adding verification time [117]. SpecTr [118] utilizes speculative decoding to accelerate the sampling by using a small model to sample a block or sequences of tokens instead of one token at a time to compensate the increased time caused by the verification process. Xu *et al.* [117] propose a speculative decoding based model collaboration inference engine LLMCad, which employs three modules to enhance the efficiency by designing token tree for token generating and verification instead of using a linear token sequences and adopt a self-adaptive fallback strategy. Another kind of model collaboration is offering various sizes of models for acceleration to decrease resource consumption and enhance efficiency without generating new models [119] [120].

Model compression: Model compression includes a series of techniques, such as pruning and quantization techniques, to reduce the model size while preserve the model's performance. There is a potential trade-off as the size of the model is reduced, which can greatly reduce the memory footprint and energy consumption, speed up inference, and enhance the scalability, while it can cause accuracy loss, introduce distortions in the model, and increase additional computation overhead. It is crucial to balance those aspects. Compression techniques for LLMs/GPT have primarily concentrated on quantization. The work presented in [121] introduces "Zeroquant", an end-to-end quantization and inference pipeline leveraging post-training quantization and layer-wiser knowledge distillations to reduce the precision of LLMs' numerical representations while minimize the accuracy loss. SparseGPT [122] utilizes the pruning method to reduce the model size and shows that pretrained LLMs can be pruned at least to 50% sparsity and up to 60% sparsity in one shot without any retraining while achieving the minor accuracy loss. SparseGPT enables the execution of GPT models with 175 billion parameters to be completed within a few hours on a single GPU. Besides model compression, the work in [123] presents a prompt compression method to deal with the resource challenges where LLMs are accessible only through APIs.

Open questions and research opportunities

LLM distributed or hierarchical inference in 6G: 6G networked infrastructures have in general machine learning workloads in sight as one of the main application domains. To facilitate such workloads novel architectures could be introduces thatsignificantly improve distributed inference for LLM, particularly by 1) reducing communication overhead among distributed nodes for efficient coordination of the inference process; 2) improving scalability with efficient networking infrastructure which supports the dynamic allocation of networking resources and seamless communication between nodes; and 3) addressing fault tolerance, for example for ensuring the reliability and consistency of distributed inference for LLMs. Likewise arguments can be made for hierarchical inference. Multiple layers of inference models could be placed in future networks, offering different trade-offs between accuracy and latency. How to devise and activate these different layes, and how to support them optimally through architectures is a open challenge. A substantial challenge for 6G networks will furthermore be the management of mobility under either distributed or hierarchical inference workloads. In tendency, smaller models closer to the end devices will be subject to relocation, while larger models will be placed such that relocation is less likely to be necessary. However, the trade-offs are task-dependent as well as context-dependent, requiring entirely new mobility management architectures as well as algorithms.

On-device LLM inference: To allow LLM-empowered ubiquitous, privacy-preserving, and highly available GenAI, LLM inference will ultimately sink to near-user devices. Preliminary efforts have been already made to bring LLaMA-7B to smartphones and PCs. In the future, how to support on-device LLM inference, considering the trade-off of accuracy, consumption of resources, and scalability, will become a key competitive force

in the business model of hardware vendors. As mentioned above, hierarchical inference adapted to LLMs can play a substantial role in supporting the migration process to smaller models running on end devices.

Optimization of inference across heterogeneous de-centralized environment: LLM partitions can be more effectively distributed to achieve higher inference throughput in a setup with homogeneous computing nodes. The challenges lies in dynamically optimizing model partition for parallelism by leveraging heterogeneous edge devices, taking into account factors such as GPU memory bandwidth, memory constrains, peak FLOPS, diverse edge connections, and model architecture.

Benchmark for evaluating LLM inference: There is not a single benchmark for evaluating distributed LLM inference, as it highly depends on the specific tasks and evaluation goals. Therefore, key challenges pertain to development of benchmark datasets and evaluation metrics to assess the performance of distributed LLM inference. Performance evaluation metrics include accuracy, efficiency, scalability, resource usage, robustness, etc. Investigating trade-off of these metrics in different LLM systems require continuous evaluation in real-world setups.

Privacy, secure, and trustworthy LLM inference: Ensuring privacy, security, and trustworthiness are important considerations in distributed LLM inference. Privacy-preserving techniques, such as differential privacy, could be implemented to protect user data and ensure compliance with privacy regulations during the inference process. Development platform that employs access control and other security mechanisms will defend against adversarial attacks aimed at manipulating inference output or extracting sensitive information. Ensuring trustworthiness is another critical challenge, as users require assurance that LLMs deliver reliable and unbiased inference results. Addressing these challenges necessitates the development of trustworthy LLMs suitable for real-world applications.

Graphical Approaches and Spatial Reasoning with LLMs for 6G

LLMs, using pre-trained parameters, can answer questions; however, their internal knowledge can sometimes be incomplete or inaccurate, leading to factually incorrect responses [124]. Moreover, LLMs often lack domain-specific expertise, which can result in unfounded assertions that are difficult to verify due to the lack of transparency [125]. This limitation is particularly concerning for high-stakes applications such as medical diagnosis [126]. LLMs also struggle with understanding complex semantic relationships between multiple entities, which is essential for grasping analogous concepts across varied textual scenarios. Fine-tuning LLMs to update their knowledge base is also a time-consuming process [127] [128].

Furthermore, LLMs frequently encounter challenges with complex multi-step logical reasoning [129] [130], making it difficult for them to solve problems requiring detailed, sequential thought processes. Their limitations in understanding spatial and topological relationships restrict their utility in fields that rely heavily on geometric and structural data. Additionally, their lack of temporal awareness hampers their ability to process and predict time-dependent data sequences effectively.

Addressing these challenges is essential for expanding the application range of LLMs and enhancing their utility. Recent research efforts have focused on improving their reasoning abilities, precision in calculations, and understanding of spatial and temporal relationships [131]. Innovations in model architecture, training methodologies, and data representation are key areas of development. For instance, the introduction of new training datasets that include temporal and spatial dimensions, and the exploration of models that can integrate external knowledge bases for better context understanding, are among the strategies being pursued to overcome the limitations of current LLMs.

The goal of such improvements is not only to refine the performance of LLMs on traditional NLP tasks but also to enable their application in more complex problem-solving scenarios. This includes tasks that require a deep understanding of the physical world, intricate decision-making based on dynamic data, and the ability to reason over long time horizons. By evolving LLMs in to Graph-LLMs [132] [133], researchers aim to make it possible for these models to contribute beyond text processing.

Graph Knowledge and Spatial Reasoning in LLMs

Knowledge graphs (KGs) offer a structured, interpretable, transparent, and dynamic layer of knowledge for LLMs, enhancing their trustworthiness in real-world applications. This layer organizes knowledge through entities and their relationships, acting as a pivotal reference point. It ensures consistency of responses by enabling LLMs to understand the context and spatial and temporal interconnections between data points. It also allows LLMs to deduce new insights from established relationships, resulting in more accurate and relevant outputs [134].

Through this dynamic layer, LLMs are able to fill the knowledge update gap by staying in the loop with the latest domain-specific insights. Keeping up with changes through KGs, which can be continually updated with new information, is crucial for helping LLMs deliver accurate and relevant answers in rapidly evolving fields. With this

layer, LLMs gain semantic search capabilities, transparency, and explainability, enabling them to understand queries more deeply. Consequently, LLMs can more accurately discern the underlying intent of a query, leading to responses that more closely match user expectations. Additionally, it allows users to gain a deeper understanding of the model's conclusions, which improves their confidence in its outputs. In this section, our aim is to highlight studies that have contributed to the advancement of this layer within LLMs. We provide an overview of techniques to incorporate graph knowledge and spatial reasoning into LLMs via training graph foundation models [135] [136], fine-tuning, prompting [137], evaluation, and interaction among multi-LLMs agents. We discuss how to integrate Graph-LLMs with their spatial reasoning capabilities into the strategic layer as high-level decision-making processes and predictive analytics tools and into the logical layer as network design tools.

KG-Enhanced LLM Inference: As a result of changing the inputs to the model for inference, knowledge can be updated within the model. LLMs capture both textual semantic meanings and the latest real-world information to effectively implement KGs. It involves retrieving relevant knowledge from large corpora and integrating it into LLMs. In an alternative approach, KGs prompting is used to transform structured data from KGs into textual sequences. In turn, LLMs can use these sequences as context, enhancing their reasoning capabilities by leveraging the KG structure. However, this process typically involves manual prompt design, which requires significant human effort.

To avoid generation of factually inaccurate answers and high costs of updating knowledge, Baek *et al.* [138] proposes augmenting the knowledge directly in LLMs inputs. A semantic similarity between the input question and its associated facts is used to retrieve the relevant facts from the KGs. A prompt containing the retrieved facts is then appended to the input question, which is forwarded to LLMs for generation of an answer. Ye *et al*. [139] designed a series of rule-based instruction prompts for general graph structure representation and graph ML. Buehler *et al.* [140] suggest adding relevant information to the prompt. Thus, the model has access to expanded context, including details, measurements, and new data. This greatly expands the capabilities of a LLMs during generation.

Jiang *et al.* [141] propose a method that involves gathering relevant evidence from structured data sources and allowing LLMs to focus primarily on the reasoning process using the gathered information. Through this approach, LLMs will be able to systematically enhance their reasoning capabilities when dealing with structured data, enabling them to derive accurate answers to their specific questions. Kang *et al*. [142] address the limitations of fine-tuning small Language Models (LMs) for knowledge-intensive reasoning tasks, which require extensive knowledge and reasoning capabilities. Their innovative approach enhances small LMs by fine-tuning them to produce rationales derived from LLMs equipped with external knowledge. Additionally, they suggest implementing a neural reranker to select documents relevant to the generation of rationales, further improving its efficiency.

Furthermore, Baldazzi *et al.* [143] utilizes a reasoning verbalization approach to create prompt-response pairs coupled with a lifting method that leverages patterns in reasoning to improve LLMs for task- and domain-specific applications. Using an ontological reasoning task applied to Enterprise Knowledge Graphs, this method will guide the fine-tuning process.

Sun *et al.* [144] integrate frequently updated external KGs into LLM reasoning to address the issue of hallucinations in LLMs. They propose that LLMs are dynamic agents with dynamic interactions with KGs, exploring related entities and relationships as the basis of reasoning. With this approach, LLMs can generate reasoning outcomes for tasks that require extensive knowledge based on retrieved knowledge.

According to Wen *et al.*'s study [145], KGs are used to prompt LLMs, integrating up-to-date knowledge and extracting reasoning processes from LLMs. This approach enables LLMs to understand inputs from KGs and to deduce by combining implicit internal knowledge with external information retrieved from KGs. Consequently, LLMs are capable of conducting reasoning tasks and producing answers accompanied by visual representations of the rationales for these answers.

Tian *et al.* [146] introduce Graph Neural Prompting (GNP), a groundbreaking method designed to enhance the capability of pre-trained LLMs by facilitating the learning of valuable knowledge from KGs.

Luo *et al.* [147] introduce Reasoning on Graphs (RoG) method, which combines LLMs with KGs to enable reliable and interpretable reasoning. They propose a planning-retrieval-reasoning framework to overcome challenges such as hallucinations and knowledge deficits. Using this framework, LLMs can take advantage of the latest knowledge by reasoning using faithful plans structured around graphs.

Guan *et al. [*148] introduce Knowledge Graph-based Retrofitting (KGR), a strategy that integrates KGs into large models' reasoning process. By retrofitting LLM initial draft responses with factual knowledge from KGs, this approach aims to reduce factual hallucinations. The process involves identifying draft responses that need verification, retrieving relevant facts from the KGs, and using these facts to improve the drafts.

KG-Enhanced LLM Evaluation: LLMs are criticized for their inability to interpret their internal mechanisms and decisions, despite their success in various NLP tasks. By enhancing interpretability, LLMs can be more suitable for critical areas such as medical diagnosis and legal decisions. The use of KGs facilitates clearer reasoning and interpretation because they provide a structured representation of knowledge. Many users find that they need to conduct additional searches to verify the accuracy of the information provided by LLMs. Consequently, the criteria used to evaluate the output quality of LLMs when prompted can be significantly improved.

LLMs probing is essential to uncover the vast knowledge they possess based on extensive training on large datasets. While LLMs contain vast amounts of knowledge, this information is embedded in a way that is not readily accessible, making it difficult to discern the specific type of knowledge they hold. Additionally, LLMs can produce factually incorrect statements due to hallucination, lowering their reliability. To ensure their accuracy and reliability, it is crucial to probe and validate the knowledge within LLMs.

Additionally, the analysis of LLMs using KGs seeks to unravel key questions regarding the operational mechanisms behind LLMs, specifically how they produce results and the functional and structural dynamics at play within these models.

Zhao *et al.* [149] propose a methodology to assess the effectiveness of retrieval-augmented LLMs through symbolic language reconstruction and text passage retrieval.

Wang *et al.* [150] outline a four-stage methodology that significantly enhances LLMs responses' reliability. This method includes Question Decomposition, which organizes questions according to predefined templates; Knowledge Retrieval, which gathers relevant information; Candidate Reasoning, which evaluates potential answers; and Response Generation, which generates answers along with their reasoning paths.

Using RDF-KGs, Mountantonakis *et al.* [151] introduce a framework for enhancing ChatGPT responses with comprehensive information. The system identifies entities within a response, annotates them with statistics and hyperlinks, including URIs, facts, and connections to other KGs. As a result of this enhancement, the content associated with entities can be enriched, facilitating fact-checking and validation in real time. Thus, ChatGPT's responses can be verified more efficiently and additional information can be accessed.

Gao *et al.* [152] use KGs as auxiliary tools to help interpret and summarize complex medical concepts. Using a new graph model, they enhance the ability of LLMs to generate automated diagnoses by selecting the top N diagnoses through multihop paths, thus surpassing traditional concept extraction methods.

Luo *et al.* [153] introduce a framework designed to methodically evaluate LLM factual knowledge capabilities using KGs. The method involves creating a set of questions and expected answers based on the information contained within a specific KG. Following that, the framework assesses whether LLMs are accurate in answering these questions.

Wang *et al.* [154] propose an automated testing framework for detecting factual errors in LLMs. This system constructs a structured KG on a topic chosen by the user, sourcing fact triplets from a large-scale knowledge database. To test LLMs' understanding of relationships across various topics and entities, it formulates questions across simple (one-hop) and complex (multi-hop), providing correct answers. LLMs' answers are evaluated using matching strategies customized for each type of question.

KG-Enhanced Multi-LLMs Agents Interrelationships: Aside from the integration of KG structures throughout the entire lifecycle of LLM agents, from training and fine-tuning to reasoning and evaluation, there is also the possibility of applying KG structures to the relationship networks between LLM agents, in order to model the interactions and information flow between LLM agents effectively. The interrelationship of LLMs agents is crucial since they govern the mechanisms of interaction within these structures, whether they are collaborative, competitive, or hierarchical in nature. Allocating roles and tasks, as well as managing the distributed constraints on resources effectively, are crucial in cooperative multi-LLM agents engaging in collaborative decision-making. This collaboration can be used within the planning phase to select the most effective agents according to specific metrics [155], [156] or to distribute tasks and correct each agent in order to improve the overall team-work performance [157], [158]. Additionally, it can be used during execution phase through the collaboration and feedback of multiple agents, allowing the agent to create a final outcome by learning from the interactions of multiple agents [159], [160], as well as during reasoning phase [161].

With their structured factual information through entities and their relationships, knowledge graphs can provide sophisticated methods for interconnecting data and understanding tasks semantically for cooperative multi-LLM agents. This architecture provides a dynamic, structured, and semantically rich knowledge repository. This enables more efficient allocation of roles based on agents' knowledge, skills, and current workload. As well, knowledge graphs can be dynamically updated with information gathered by LLM

agents. This facilitates continuous learning and adaptation as well as empowering agents to refine their collaboration strategies.

Additionally, KGs can enhance strategic depth in competitive interactions among LLM agents by providing insights into their strengths, weaknesses, strategies, and historical performance. The LLM agent can then develop more informed and nuanced strategies, anticipate competitors' moves, infer the partner's intention and reason actions, and identify strategic opportunities or threats [162]. The hierarchical relationship among multi-LLM agents facilitates effective collaboration across multiple levels, resembling a tree-like structure where parent node agents are responsible for breaking down complex tasks and distributing these tasks to child-node agents. Knowledge graphs can improve the efficiency of these hierarchical control systems by optimizing the flow of information, as well as enhance the methods used for task decomposition. With knowledge graphs, parent nodes can efficiently break complex tasks down into smaller, more manageable chunks, and then assign these chunks to child nodes based on dependencies and capabilities needed.

In order to achieve seamless integration between multi-LLM agents, it is important to orchestrate communication and information exchange among a potentially vast network of agents [163] [164]. Decentralized Planning Decentralized Execution systems one of the planning methodologies for the he orchestration of multiple agents. It is characterized by individual agents that independently plan and execute tasks with minimal coordination. In spite of this, the advent of multi-LLM systems has ushered in a paradigm that involves more sophisticated information flow management and coordination across agents. This poses a number of challenges such as communication overhead, latency, bandwidth utilization, and ensuring agents are updated with the latest updates. In each of these cases, the system's performance and responsiveness are adversely affected, resulting in suboptimal efficiency and effectiveness [165].

Additionally, the traditional architecture of multi-LLM agent systems has relied on shared memory, a centralized data structure, to facilitate the exchange of information between agents. In this approach, agents can store and retrieve data within a common memory space, enabling them to collaborate and share information efficiently. By using shared memory, agents can easily access and update information necessary for their operations, simplifying communication processes among them [165]. Although this centralized model for information sharing offers many advantages, it poses some challenges as well. Due to the increasing number of agents requesting access to the shared memory, issues related to contention and synchronization arise. Furthermore, scalable, distributed, and mobile agent systems make ensuring consistency across agents and data increasingly challenging.

Graph Neural Networks: Graph Neural Networks (GNNs) have emerged as a promising approach to tackle various challenges in wireless communication networks. In this research direction, network devices and communication channels are represented as nodes or edges in a graph, device information as node features, and channel-specific parameters as edge features.

Shen *et al.* [166] formulated wireless networks as wireless channel graphs, where transceiver pairs were considered as nodes and communication channels were directed edges. This approach leveraged permutation equivariance of GNNs to achieve robustness and efficiency in power control and beamforming. Chen *et al.* [167] modeled a wireless network as a directed graph whose nodes are the desirable communication links and edges are the harmful interference links. Their approach could be generalized to different network settings and were robust to corrupted input features.

In a vehicle-to-everything (V2X) network, each vehicle-to-vehicle (V2V) link is a node in a graph, a GNN learns the low-dimensional feature of each node and a deep Q-network learns to optimize the sum capacity of the V2X network for spectrum allocation [168]. In order to automate network management tasks, LLMs were utilized to generate graph manipulation code from natural language queries on various network topologies and communication graphs [169]. In distributed machine learning, Wang *et al.* [170] utilized GNNs for over-the-air federated learning by mapping channel coefficients to optimized model parameters. Their approach represented edge devices, edge servers, and Reflective Intelligent Surfaces (RIS) as nodes in a GNN.

These results have showed that GNNs could improve various aspects of wireless communication networks, ranging from traditional tasks such as power control and resource allocation to emerging applications such as federated learning and network management automation. Hence, integrating graph knowledge and spatial reasoning ability of GNNs into LLMs promises to enhance the efficiency, reliability, and scalability of wireless communication systems, especially in the context of AI-native 6G networks.

Challenges in 6G Systems

We highlight challenges emerged from the characteristics of 6G systems which can affect Graph-LLMs, such as: local connectivity [171], graph dynamicity [172], and data velocity [173]. These challenges are analyzed in two use cases: Graph-LLMs for interactive 6G signal

modeling in 3D environments [174] and Multi-agent autonomous driving systems [175] [176] [165], each with its own LLM.

Fatemi *et al.* [171] showed that pure text representations of structured data were not sufficient for graph reasoning with LLMs, in comparison to traditional graph algorithms. They used prompting methods to measure the performance of pre-trained LLMs on graph reasoning tasks, including: edge existence, node degree, node count, edge count, connected nodes, and cycle check. The evaluation indicated that graph encoding methods (such as adjacency and incidence matrices) significantly impacted LLM reasoning capability on graph problems.

In addition, the results illustrated that LLMs performed worse on disconnectivity-related tasks since the encoding methods could not express the absence of connections. Zhang [172] introduced a benchmark to evaluate the spatial-temporal understanding abilities of LLMs on dynamic graphs, which can appear in vehicular networks of autonomous driving systems. Another challenge of such 6G systems as autonomous driving vehicles is the amount of data collected by built-in sensors. Processing the information to ensure prompt decision-making is crucial [173].

6G Signal Modeling in 3D Environments: LLMs, as well as Vision-Language Models (VLMs), have proven their capabilities in multiple tasks. By equipping them with the knowledge of our 3D physical world, they can be enriched with such concepts as spatial relationships, affordances, physics, and layouts [174].

To effectively model the 3D physical world in an interactive manner, 3D-GPT, a procedural generation method utilizing LLMs, has emerged as a promising approach [177]. In order to reduce the effort to produce a Radio Environment Map (REM) of an Ultra Dense Network (UDN), a GNN was trained on sparse power spectral density measurements to perform REM prediction [178]. Building on these three pillars, we envisage an interactive radio propagation modeling framework, which allow users to explain their requirements in natural languages (see Figure 11). A Graph-LLM processes the requests and generate 3D scenes and REMs procedurally, as well as providing explanation. The system can deal with graph dynamicity [172] [179] due to diverse and refining requirements of the users.

Multi-agent Autonomous Driving Systems: Recently, Cheng *et al.* [165] have explored the generalization capabilities of LLM-based agents across diverse applications from general-purpose assistants to various scientific and engineering domains. LLMs offer an enabler to innovate the autonomous driving domain, in which each vehicle has its own LLM (see Figure 12 for an overview). Following this direction, Wen *et al.* [176] integrated a reasoning module to allow the driving agent to perform decision-making based on common-sense knowledge.

In autonomous driving, Graph-LLMs facilitate multimodal perception of the environment from various sources such as sensors, weather conditions, real-time traffic information, passenger requests in natural languages, and road rules. For example, a physics-based GNN [180] can classify light detection and ranging (LiDAR) data, where the LiDAR point cloud was transformed into an undirected graph by connecting each point to its k-nearest neighbors. Furthermore, multiple LLM-based driving agents can communicate to each other for advanced planning [181], cooperative decision making [182], and knowledge sharing [155] [163]. In this use case, there are two important challenges: graph dynamicity [172] [179] and data velocity [173]. Driving agents empowered by Graph-LLMs should be

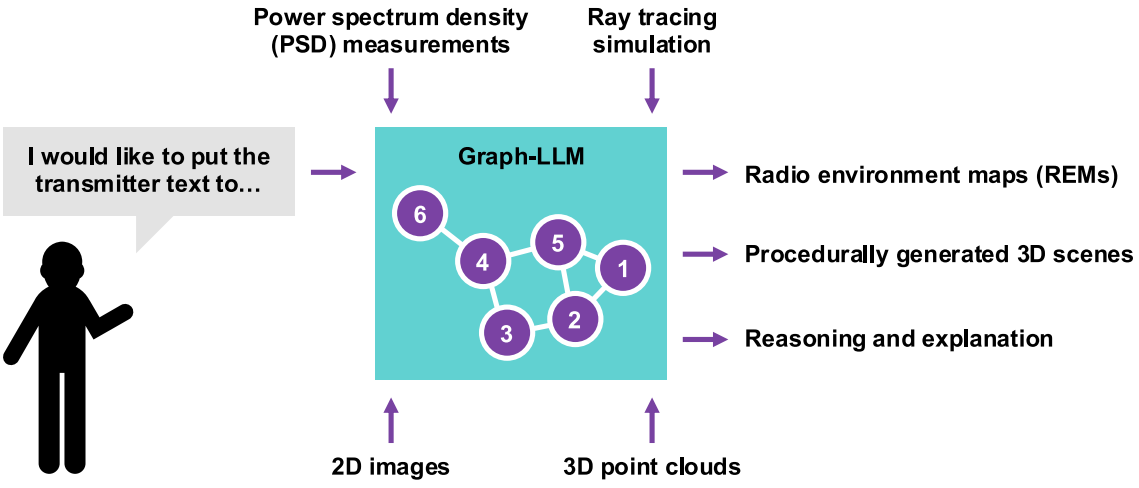


Fig. 11: Graph-LLMs for interactive 3D modeling

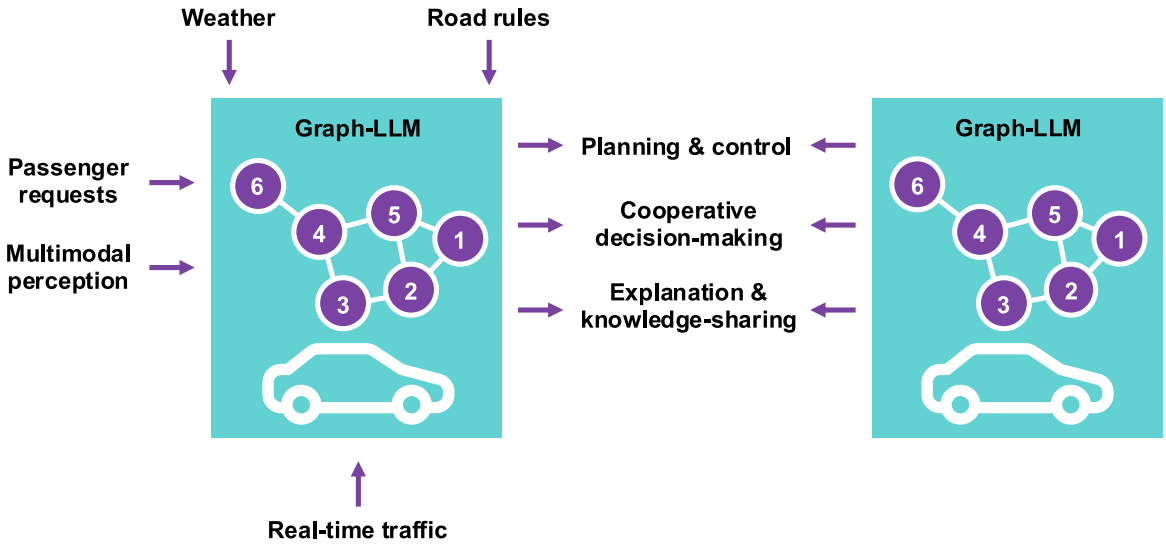


Fig. 12: Graph-LLMs for autonomous multi-agent systems

able to handle constantly-changing graph structures and environmental conditions to ensure safety [173].

Hardware Acceleration for Large Language Models

The deployment of Large Language Models (LLMs) in 6G networks presents significant challenges in terms of the hardware required to support their computational needs. In particular, for a distributed deployment across the computing continuum, several innovations are necessary to enhance the efficiency and feasibility of running LLMs in 6G environments. Given the high power consumption and thermal issues associated with current Neural Processing Units (NPUs), there is a pressing need for new hardware solutions that are both power-efficient and cost-effective. Here, we discuss some the existing limitations and potential advancements in hardware technology that can facilitate the integration of transformer-based models into commercial wireless systems.

Transformers in wireless SoC

Due to the inherent complexity and size of Large Language Models (LLMs), which involve billions of parameters working together in transformer architectures, integrating them into wireless systems presents unique challenges compared to smaller neural networks commonly used in transceiver algorithms like detection and channel estimation. While these smaller networks rely on supervised learning and are computationally manageable, LLMs demand significantly higher computational and energy resources.

Given the costs of energy and manufacturing, power efficient and low cost realization of hardware for acceleration of transformers is desirable for telecom vendors and operators. Considering current realizations of Neural Processing Units (NPU) that show a comparable or even higher power consumption profile than that of, e.g., physical layer System on Chips1 (SoC), new solutions beyond the conventional NPUs is needed. Otherwise, it would be difficult to justify for integration of transformer based models into commercial wireless systems.

Furthermore, due to thermal issues, integration of NPUs into wireless network strongly demands reduced power consumption. Unlike NPUs in data centers that enjoy relatively good cooling infrastructure, the cooling capacity within wireless network infrastructure is very limited. Considering notorious thermal issues of high-end NPUs [183], the thermal issues can be a show stopper for high performance LLM hardware [184].

To enable efficient computing for LLM like models, i.e., models based on transformers, the first approach is to simply optimize the LLMs and hence reduce computational load on the processing part [185]. However, the performance may not still be satisfactory for wireless applications. Hence, in following, hardware and hardware/software innovations that are tailored for LLMs are investigated.

Hardware/Software innovations for LLM computing

The attention mechanism which involves large vector-matrix multiplications (GEMV) is the most crucial and time consuming kernel in computations of LLMs [186], and hence most of accelerators focus primarily on efficient acceleration of GEMVs. The other limiting factor is large model sizes and heavy memory access, which brings up the notorious memory bottle-neck problem [184]. Various, techniques discussed here either target

the former or the later or the both through exploiting **redundancies** [187], **approximate computing** [188] and **sparsity** [189]. Later we delve into unorthodox computing solution that are specific to neural acceleration, i.e., **processing in memory** and **near-threshold computing** [190]. Notice, most of these solutions are orthogonal to each and their joint force may be applied to achieve maximum possible efficiency.

Due to large size of LLMs, data access to last level memory may become a bottleneck. DRAM space is less scare resource compared to DRAM bandwidth. General data compression methods in memory controller or neural networks specific model compression are helpful in improving performance, in particular that, neural networks weights and processing data tend to be quite compressible [191].

Exploiting redundancies observed in LLMs provides for optimization in design of LLM specific acceleration as it is done by [187]. Similarly, ELSA [192], omits non-essential redundant computations of relations in self-attention mechanism [192]. On the other hand, similiar to other neural network models, LLMs too are resilient against small errors in computations. Hence, approximate computing [184] is another approach the can be taken advantage of. In particular, notice that baseband computing in physical layer is rather forgivable with errors in data and computations. Exploiting this fact and considering the neural networks already have fault tolerance property, approximate computing technique can contribute. This has been the focus of [188], where a progressive approximate computing approach is proposed. However, as approximate computing introduces errors in the result, a conclusive evaluation of cost and benefits is difficult to draw [188]. Sparsity within LLMs also can be exploited for performance enhancements as well. In [189], by observing that the attention mechanism in language models inherently contains a large number of redundant connections that can be sparsified into a simpler model. In [189] a sparse attention model and associated hardware accelerator for exploiting the sparsity was proposed and investigated.

Alternative computing paradigms for LLM computations

Beyond exploring hardware-software co-design within the traditional computing framework, it is important to consider new and unconventional computing approaches. These approaches could hold the key to unlocking breakthrough solutions in energy efficiency. Here, we discuss two approaches that we believe have significant potential.

Processing In Memory: The re-emerging field of Processing In Memory (PIM), i.e., Processing Near Memory (PNM) and Processing Using Memory (PUM) have shown very promising results for large data manipulation. The former performs computing by devising computing elements close to the memory element and the later by using memory elements without adding extra computing circuity [193]. PNM and PUM remove the memory bottleneck of Von Neumann architecture and simultaneously act as a highly parallel Single Instruction Multiple Data (SIMD) or Multiple Instruction Multiple Data (MIMD) accelerators [194]; killing two birds with one stone. While there are already commercial PNM based platforms, e.g., UPMEM [195], PUM based systems are mostly under developments since those often rely on analog computations within the memory elements and suffer from reliability issues.

A commercial PNM based systems from SAMSUNG [186], demonstrated a performance on par with a high end Graphics Processing Unit (GPU), i.e., NVIDIA A100 and close to 2x improvement in energy efficiency when integrated into an AMD GPU [196]. Using PUM-based systems [197] and [198] demonstrated improve 10x to 100x improvement in vector-matrix operations. Due to ease of manufacturing and reliability of PNM systems, one can expect those to be adopted earlier by industry, hence, focusing on realizing of LLM on PNM seems more likely to end up in products.

PIM systems are generally limited to performing low resolution bit-serial operations [198]. Developing models that can be trained and inferred by using only few arithmetic bits is attractive as it makes a perfect match to memory-based computing. Such a promising alternative is introduced as end-to-end 1-bit LLM training and inference framework [199]. Considering possibility of performing bit-serial operations within memory-based computing systems [198], we believe integration of [199] framework can be considered a promising solution to investigate.

Near Threshold Computing: Due to quadratic relationship of dynamic power as the major component of power consumption in digital circuits with voltage, reducing operating voltage is very effective approach for increasing energy efficiency. Operating at voltages near the threshold voltage of transistors, promises up to 10x higher energy efficiency and is known as Near Threshold Computing (NTC). The achievable energy efficiency and trade off in performance is depicted in Fig. 13. As mentioned, although NTC needs to trade off the performance, however, that can be compensated by more parallelism.

---

1. A high-end wireless chip may consume only tens of Watts, while a highly optimized NPU that can handle real-time requests for processing an LLM consumes 200 W to 300 W [183].

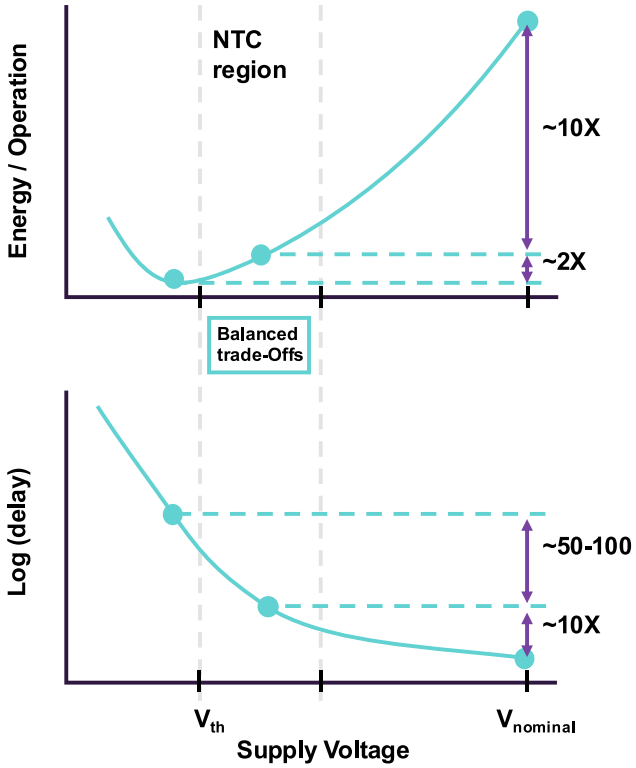
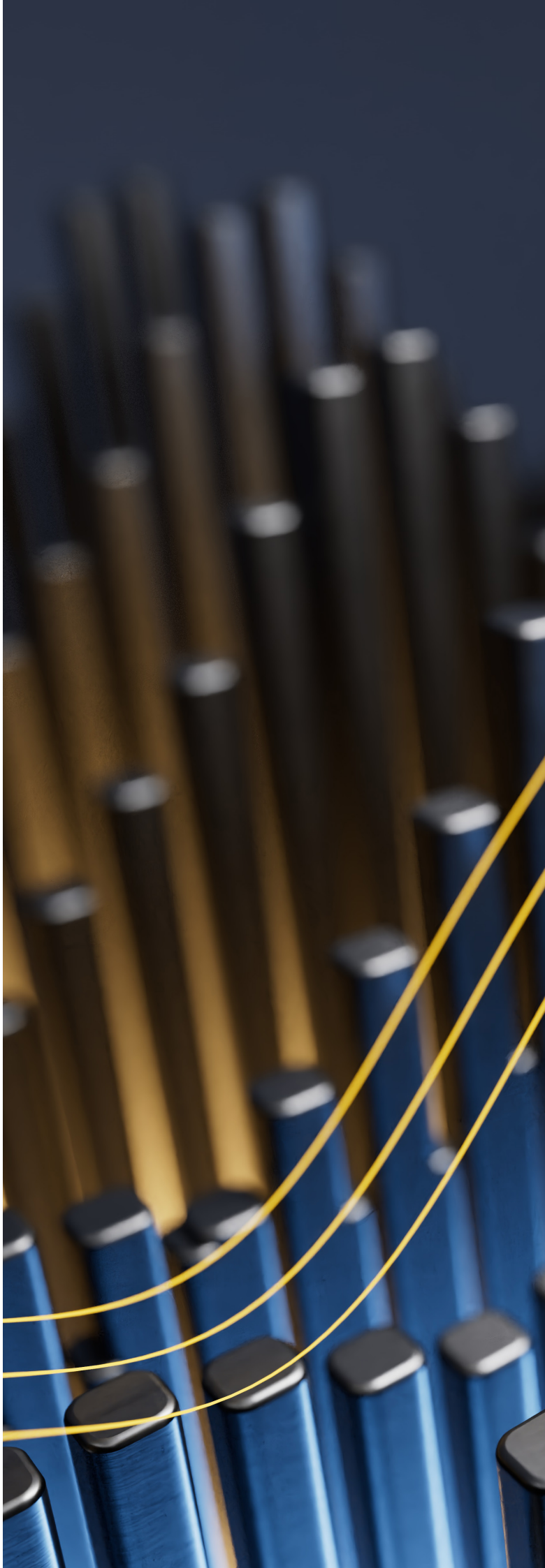
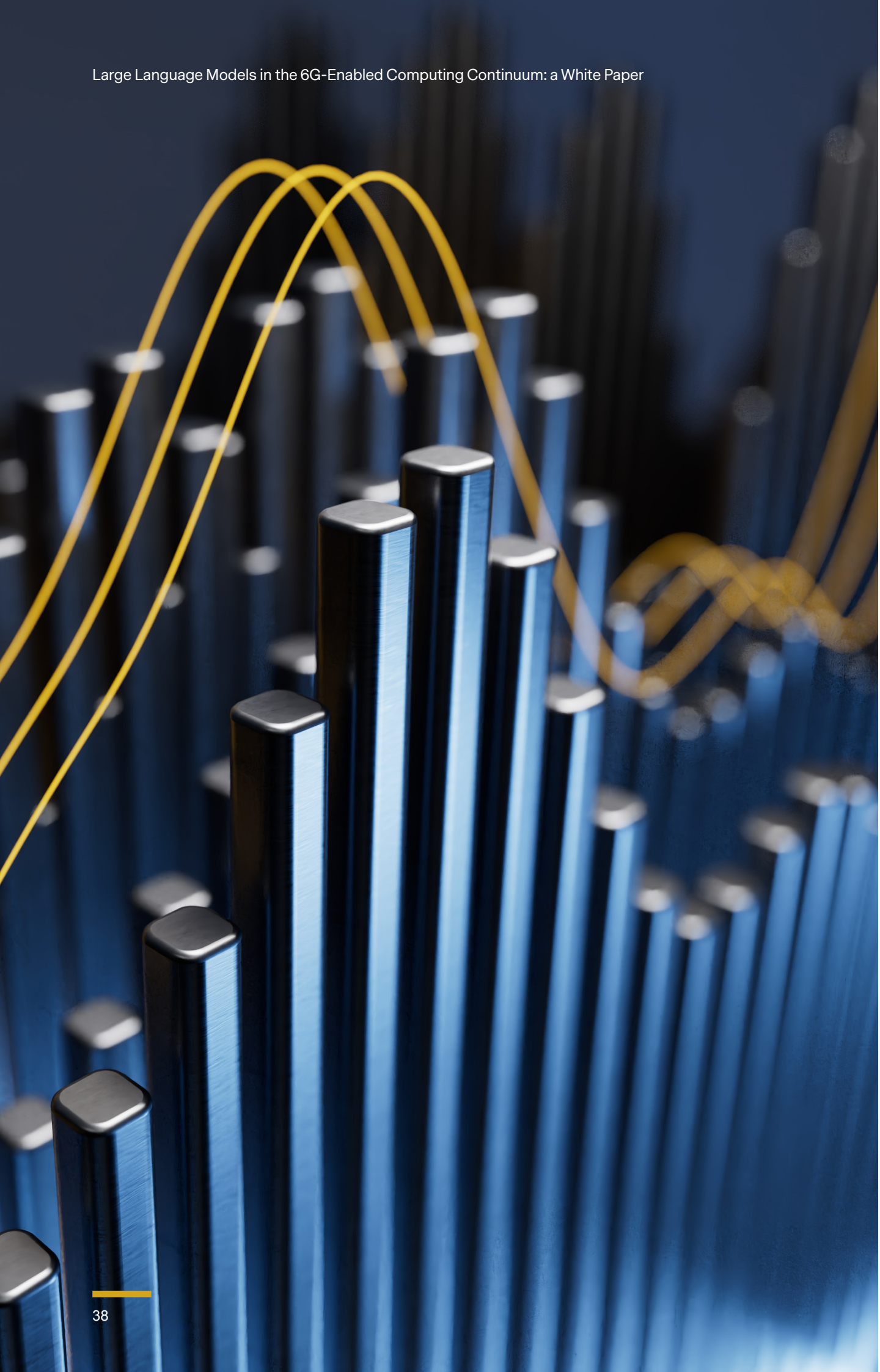


Fig. 13: Impact of voltage down scaling on energy efficiency and logic delay (adapted from [200] © IEEE).

Despite the substantial benefits of NTC, unfortunately, due to heightened sensitivity of the circuit at reduced voltages to Process, Voltage and Temperature (PVT) variations, it is extremely difficult to design near threshold voltage operating processors. Previously, some efforts using Timing Error Detection (TED) circuit were carried out [200], however, inserting TED circuitry into the processor incurs large design cost, overheads and complexity.

Fortunately, there is a unique opportunity for realization of NTC based neural accelerators by using specific mathematical operations that enables reliable computing even at reduced voltage regimes. This eliminates the need for complex TED based systems for enabling NTC. By integration of algorithmic error detection approaches such as Algorithm Based Fault Tolerance (ABFT) into matrix and convolution operations of neural models, reduced voltage operation has been made possible [201], [202], [203]. On the other hand successful integration of ABFT into various Convolutional Neural Networks (CNN) models were already demonstrated with very low computational and memory overhead, e.g., less than 4% [204]. We believe, due to heavy utilization of matrix arithmetic in LLMs, the same results are achievable with integration of ABFT into those models as it was shown [203]. This is an interesting research direction to investigate.





State of the Art & Applications

5

In the pre-print that forms the foundation of this white paper [205], we discussed how LLMs were anticipated to significantly advance 6G networks across various use cases and architectural aspects. However, since the publication of that paper, the state of the art has quickly become outdated due to a remarkable boom in research activities in this domain. The rapid evolution of the field is evidenced by the massive increase in publications, with a wide array of studies emerging that address both specific challenges and broader visions for integrating LLMs within the mobile network context. Fig. 14 tracks the number of publications per year, shows a significant increase from 2023 to 2024. It is important to note that the count for 2024 includes publications only up until the end of July, representing just over half the year. Despite this, the number of publications has nearly doubled compared to 2023, underscoring the growing interest and importance in exploring how LLMs can be integrated into telecommunications. This sharp rise suggests that by the end of 2024, the increase in publications could be even more substantial. This increase can be attributed to both the maturation of LLM technology and the increasing recognition of its potential impact on next-generation network architectures.

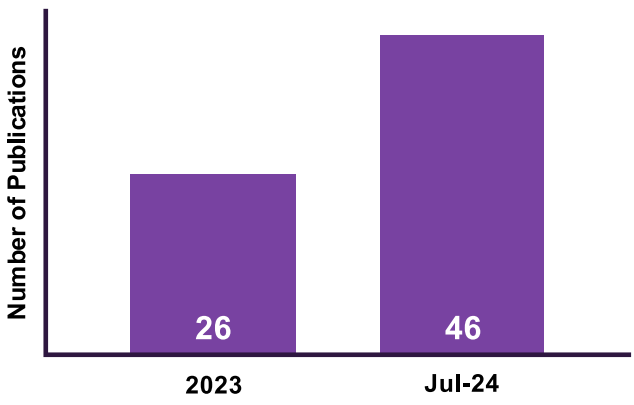


Fig. 14: Number of publications per year on Generative AI and LLMs in telecommunications (2023-2024). The 2024 data includes publications up until the end of July.

The current landscape of research is so dynamic that it has become increasingly difficult to keep track of the high volume of publications release weekly. These studies span multiple categories, reflecting the broad scope and depth of interest in this area. Given this surge, we rely on comprehensive sources, such as the *research library* of the *Large Generative AI Models in Telecom Emerging Technology Initiative (GenAINet ETI)* [206], which provides the most up-to-date collection of relevant publications per different categories.

Fig. 15 shows the categorization of publications, further highlighting the breadth of research activities in the GenAI and telecommunications. Categories such as *Reviews, Surveys, and Tutorials* and *Large Generative AI and Semantics/Effective Communications* have seen substantial contributions, reflecting the need for both foundational understanding and communication-focused applications of LLMs. Similarly, categories like *Edge Intelligence via Large Generative AI Models* and *Security, Privacy, and Resilience Aspects* show strong research interest, pointing to the practical and security challenges that come with deploying LLMs in telecom environments.

However, despite the extensive research activity, there are some areas where contributions appear to be lacking. For example, categories such as *Datasets, Demos, and Prototypes* have a relatively lower number of publications. This is noteworthy because datasets are critically important in advancing research and development in this field. High-quality datasets and robust prototypes are essential for benchmarking and validating new models and solutions, yet their scarcity suggests a potential gap that the research community needs to address.

Another notable absence is in the area of standardization. While research and development are crucial for innovation, ensuring that these advancements are aligned with industry standards is equally important. Entities such as the IETF, ETSI, and ITU-T play a pivotal role in establishing the protocols and frameworks that guide the integration of new technologies into existing systems. Without the support of these standardization bodies, there is a risk that innovations in LLMs and 6G networks may not be fully interoperable or aligned with global practices, potentially slowing down adoption and deployment. This highlights the need for more focused efforts in standardization to ensure that research outcomes are cutting-edge as well as also practically applicable and universally accepted.

This categorization of research also emphasizes the diversity of LLM applications within 6G networks. The spread across multiple categories indicates that researchers are not just focusing on a single aspect of integration but are exploring a wide array of possibilities, from network management and orchestration to green wireless and beyond. It is also clear that there are active research efforts in both directions: **6G for GenAI/LLMs** as well as **GenAI/LLMs for 6G**. However, the low representation in some critical categories suggests that while there is broad interest, some essential areas may still be underdeveloped, which could hinder long-term progress without focused attention.

In the context of GenAI/LLMs for 6G, recent research is exploring diverse applications that leverage AI to enhance network performance, user interaction, and system optimization. Recent literature reveals a convergence of challenges, solutions, and shared objectives across multiple domains, highlighting both the potential and the current limitations of AI-driven 6G networks.

Several studies have emphasized the transformative role of generative AI in enhancing 6G network management and optimization. For instance, Tao *et al*. [207] and Bariah *et al*.

[208] investigate the application of digital twins for network emulation and management. These studies illustrate how digital twins, supported by generative AI, can bridge the gap between data-driven and model-driven approaches, facilitating more accurate and adaptive network control. By enabling real-time synchronization and resource optimization, digital twins can address the complex orchestration needs of 6G, aligning with the broader trend of leveraging AI to manage increasingly dynamic network environments.

Wireless perception and AI-generated content form another key area of exploration, particularly in enhancing user interaction within 6G networks. Wang *et al*. [209] combine wireless perception (WP) with AI-generated content (AIGC) to guide digital content production. Their proposed WP-AIGC framework demonstrates how AI can be used to interpret and adapt to user behavior in real-time, significantly improving the quality and responsiveness of digital experiences. This approach enhances user engagement, exemplifying the bidirectional relationship between AI and 6G technologies. GenAI benefits from 6G's capabilities but also actively contributes to optimizing 6G network performance.

Immersive communication represents another frontier where LLMs and generative AI are pushing boundaries. Sehad *et al*. [210] explore the role of AI in enabling the Internet of Senses (IoS), creating highly immersive multi-sensory environments by reducing bandwidth requirements and synchronizing diverse media types. Similarly, Zhang *et al*. [211] introduce interactive generative AI agents for satellite networks, employing a mixture of experts approach to optimize transmission strategies, making them more responsive and engaging for users.

Healthcare applications, particularly those involving Human Digital Twins (HDTs), are also prominently featured in recent research. Chen *et al*. [212] and Chen *et al*. [213] provide comprehensive studies on the use of LLMs and generative AI to create HDTs that enhance personalized healthcare delivery. These models facilitate continuous health monitoring [214], diagnosis, and treatment personalization, leveraging the low-latency, high-reliability communication capabilities of 6G networks.

Generative AI's role in multimedia networks and tactical applications has also been extensively studied. Xu *et al*. [215] focus on integrating generative AI into mobile tactical multimedia networks, proposing novel content distribution and generation strategies that adapt to dynamic environments. Their work demonstrates the importance of AI-driven optimization for enhancing multimedia services in 6G, particularly in resource-constrained or tactical settings.

Edge intelligence and efficient model provisioning are critical to managing the computational demands of LLMs in 6G networks. Xu *et al*. [216] address this through a cached model-as-a-resource framework that optimizes resource management in space-air-ground integrated networks. By caching models closer to the data source, this approach reduces latency and enhances performance, crucial for real-time applications in 6G.

Finally, Du *et al*. [217] investigate the use of LLMs in FPGA-based hardware development for wireless communication. Their study highlights the potential of LLMs to generate complex HDL code for advanced signal processing algorithms, using in-context learning and

Chain-of-Thought prompting to overcome scheduling and multi-step reasoning challenges. This work illustrates the broader applicability of LLMs in hardware design, contributing to the rapid prototyping of 6G network components.

Across these diverse applications, several common challenges emerge. Key among them is the complexity of integrating AI models into existing and evolving network architectures. Whether in hardware design, content generation, or healthcare, researchers consistently highlight the need for advanced AI techniques to manage data, optimize resources, and maintain low latency. Additionally, the studies collectively identify the importance of efficient data and resource management, particularly in scenarios requiring real-time processing and interaction.

The proposed solutions often call for leveraging AI-driven optimization techniques, such as model provisioning, digital twins, and advanced prompting strategies, to enhance system performance. The underdeveloped areas in LLM and AI integration within 6G networks include security and privacy concerns, energy efficiency, and the lack of standardized frameworks for seamless AI deployment across diverse 6G applications [218]. The importance of retrieval-augmented generation (RAG) and domain-specific data, as highlighted in recent research, underscores the need for secure data handling and customization techniques to enhance the accuracy and trustworthiness of AI models in telecom and beyond. These gaps indicate that while research is progressing in areas like immersive communication, digital twins, and healthcare, critical aspects such as sustainable AI models, efficient resource management, and interoperability standards remain underexplored. Addressing these challenges is essential to ensure the long-term scalability and success of AI-driven 6G technologies, especially as GenAI continues to drive innovation across the telecom sector. However, and despite the identified challenges, the results from these studies consistently assume that integrated LLMs in 6G networks will provide significant productivity gains, enhanced user experiences, and improved system efficiency.

Expanding the Horizon: LLMs in the 6G Cloud-to-Edge Continuum

The transition toward 6G architectures is driven by evolving needs and technological advancements, particularly in the context of Large Language Models (LLMs). This shift is crucial for addressing the limitations of current infrastructure and for harnessing the potential of next-generation applications. Previous generations such as 5G brought enhanced connectivity and performance but were not specifically built to meet the needs of Artificial Intelligence (AI) [71]. Several challenges are considered when integrating AI and LLMs in a network architecture. Firstly, real-time interactions with LLMs such as autonomous systems and augmented reality applications, require ultra-low latency. Previous network generations can introduce latency that degrades user experience and system performance. 6G networks promise to significantly reduce latency to the sub-millisecond level, enabling near-instantaneous communication. Secondly, AI models process and generate vast amounts of data. To operate efficiently, they require high-speed data transmission that current net-

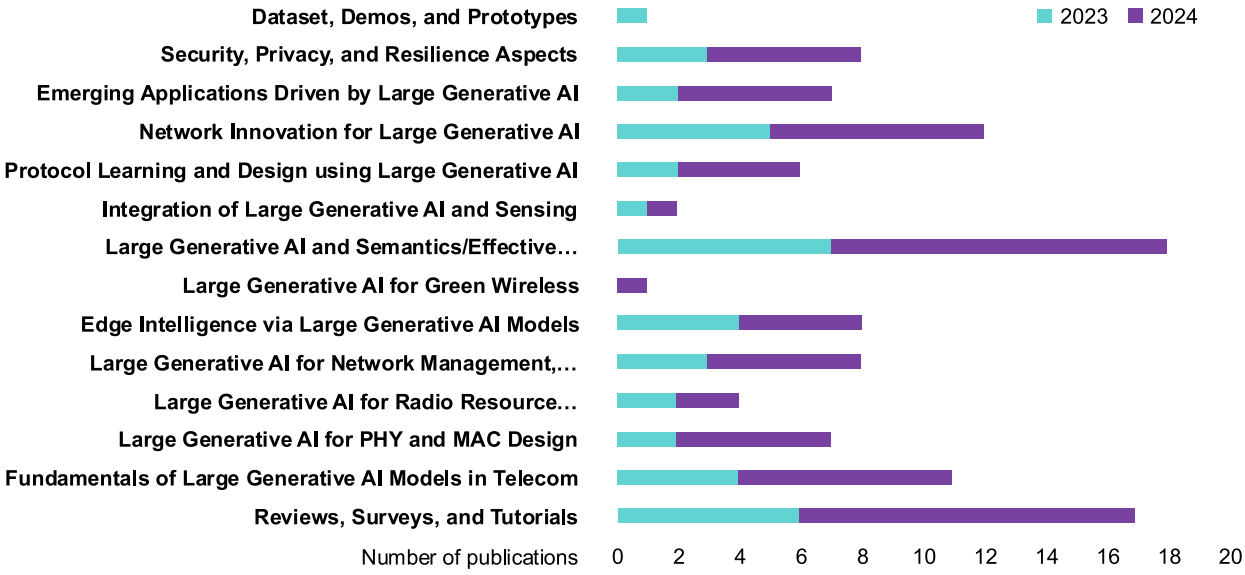


Fig. 15: Categorization of Publications on GenAI and LLMs in Telecommunications. This figure shows the distribution of research across various categories, highlighting key areas of focus and identifying potential gaps, such as in datasets and standardization.

works may struggle to provide, especially in densely populated or remote areas. With data rates expected to reach terabits per second [219] [220] [221], 6G will support the massive throughput needed for transferring large models and datasets, enabling more complex and data-intensive AI applications to operate smoothly and efficiently [222]. Thirdly, there's a growing need for decentralized processing to support applications requiring local context or immediate processing. Due to heavy data stream produced and consumed by AI algorithms, this need can hardly be met using 5G networks. Hopefully, 6G is designed to seamlessly integrate with edge computing, distributing computational resources closer to the data source. This setup minimizes latency and reduces bandwidth needs, enabling more responsive and context-aware LLM applications. Fourthly, AI applications often require constant internet access to function optimally. However, current networks can have coverage gaps or performance issues in certain areas, limiting the accessibility and reliability of these services. Promising near-universal coverage, including remote and underserved areas, 6G aims to provide consistent, high-quality connectivity. This ensures that LLM-powered applications and services are accessible everywhere. Finally, the complexity and scale of AI is rapidly increasing. Efficiently training and deploying models require substantial computational resources and advanced network capabilities. With its high-speed, low-latency communication and integration with edge computing, 6G can significantly enhance the capabilities of AI and Machine Learning (ML) operations. This supports the deployment of more sophisticated and larger-scale LLMs, driving innovation and enabling new applications.

The shift to 6G architecture addresses critical needs for supporting advanced AI and LLMs applications, offering significant improvements in latency, data throughput, edge computing integration, connectivity, and AI/ML scalability. While the integration of AI models with a 6G architecture offers numerous advantages, several technical obstacles need to be addressed to ensure seamless and efficient operation. Overcoming these obstacles is crucial for achieving efficient and effective distributed LLMs within the 6G ecosystem. The following section focuses on the various difficulties of using LLMs in the context of 6G.

Challenges and approaches to LLM integration in 6G networks

The fusion of LLMs and 6G architectures presents opportunities for enhanced NLP-based applications, dynamic information processing, and interactive user experiences. However, this integration is not without its challenges. This section enumerates the challenges encountered in harmonizing LLMs with the currenly devised 6G architectures and explores recent innovative solutions from the literature.

LLM Training Parallelism: performing efficient distribution and training of LLM across multiple computing nodes poses a significant challenge in maintaining model coherence. Additionally, ensuring synchronization and consistency across the distributed components is crucial yet complex. However, these challenges can be addressed through the implementation of parallelism techniques. Deep learning methods such as *data parallelism* [223], which involves keeping the entire network but training it on a subset of the training data, or *model parallelism* [224], where a sub-network with disjoint subsets of parameters is trained on each device, are some solutions used. The former requires managing bottlenecks, while the latter requires analyzing communication costs. Model parallelism techniques can be divided into two methods [225]: *intra-operator*, that focuses on parallelizing computations within individual operations, and *inter-operator* parallelism, that focuses on parallelizing computations across different layers. Efficient distribution and training of deep learning models across multiple computing nodes are crucial. 6G communication will enable faster and more reliable communication between computing nodes, and significantly accelerate the training process

Dynamic workload and resource management: dynamic resource allocation algorithms are required in 6G environments to better allocate resources (CPU, memory, accelerators) to LLMs based on real-time workload demands. Therefore, it is necessary to develop powerful workload prediction models to predict LLM resource requirements in order to allocate specific data on specific resources. Containerization and orchestration approaches could potentially solve this problem. In addition, it is also crucial to analyze the network architecture. In 6G communications, the network is expected to be more complex and heterogeneous. The challenge is to explore techniques for effective resource management in 6G networks specifically for the LLM. It can be solved through network softwarization [226] through the interaction of network functions virtualization (NFV) and software-defined networking (SDN). SDN and NFV manage LLM traffic by optimizing network configuration and implementing flexible and scalable network slicing.

Communication Overhead: taking into account the distributed location of resources is an advantage for optimization, but the increase in the number of data flows and the frequency of such communication may cause communication overhead, thus affecting the resource allocation efficiency of LLM. As a result, the delay caused is longer than expected. To solve this problem, [24] proposed a split inference method that shifts the computational load from edge servers to intermediate fog servers. These offloading methods, depending on the network architecture we consider [227], [228],

can be used to reduce latency, but also reduce energy impact. Communication overhead is particularly important in federated learning environments due to the potential number of communication rounds and the complexity of the aggregation stage. [229] proposed FedCPF, a solution in the context of vehicular network applications to optimize communication overhead in 6G communication networks by reducing communication rounds and improving aggregation strategies. Additionally, reducing the amount of data exchanged between nodes and using efficient communication protocols helps minimizing communication overhead.

Edge Computing Integration and Fine-Tuning: scaling LLM training across distributed edge environments presents complex challenges, particularly in preserving training efficiency and convergence. LLMs require significant resources for training and inference. While cloud environments can handle training, some companies may be reluctant to share data with LLM providers for fine-tuning. Effective solutions rely on efficient network management for coordinating edge devices and enabling distributed training and inference. This involves leveraging distributed computing and storage resources, with network virtualization guided by a central controller to orchestrate devices and coordinate fine-tuning and inference at the edge [24]. Robust fine-tuning strategies are essential for adapting distributed LLMs [230]. However, edge devices often lack the resources for extensive fine-tuning or inference. Promising techniques like *quantized learning*, *split edge learning*, and *parameter-efficient fine-tuning* can alleviate these challenges. Quantized learning, which approximates neural networks using low-bitwidth integers, reduces communication, training, and memory requirements, making it ideal for distributed computing [231]. Split edge learning reduces a device's training burden by splitting the model into two parts, with the cloud handling most of the training and the edge device managing the initial layers. This method, primarily designed for data privacy [232], also minimizes data transmission across networks. Lastly, parameter-efficient fine-tuning, such as LoRA [233], fine-tunes a small number of extra parameters instead of all model parameters, significantly reducing the number of trainable parameters, as demonstrated by reducing GPT-3's trainable parameters by 10,000 times [234]. These techniques enhance edge computing's ability to handle LLM tasks locally, reducing latency and minimizing data transmission to centralized servers.

Overfitting and Generalization: the distributed architecture of the cloud-to-edge continuum may lead to shortcomings such as overfitting. This hinders the model's ability to generalize across different datasets, because, for example, the model sticks too well to the training dataset, and it is difficult to generalize to a new dataset. Several solutions exist to overcome the overfitting, such that *network reduction*, increasing the training dataset, early stopping during the training phase, pruning and regularization techniques [235].

Data Distribution and Imbalance: ensuring a balanced distribution of training data across distributed nodes to prevent biases and maintain model accuracy is crucial. Non-iid data could lead LLMs to hallucinations [236] with inconsistent output content with real-world facts or user inputs. Data distribution also have a great impact on performances of federated learning. Non-IID data generally lead to bias in the global model, slower convergence of clients and communication overhead [237]. In Federated Learning (FL), one way to mitigate the data distribution challenge is to use *Clustered Federated Learning* (CFL) [238]. CFL enables clients to be grouped according to the data distribution and to generate personalized models at cluster level. These models are therefore better adapted and more efficient than a global model which is trained on non-IID data. A second approach is data augmentation which consists in generating new training samples without collecting new data [239]. *Data augmentation* is specially relevant when needing to increase the diversity of a dataset. It can therefore reduce the class imbalance. Finally, another approach is *ensemble models*. This approach combines several models trained on different data distribution to obtain a new one with a better generalization [240]. However, this challenge is a very dynamic area of ongoing research and techniques and improvements continue to emerge.

Privacy-Preserving Techniques: distributing LLMs on the cloud-to-edge continuum introduces new security challenges such as protecting data during communication, ensuring secure access to models, and preventing unauthorized access. Implementing privacy-preserving techniques such as *federated learning*, *secure multi-party computation*, and *differential privacy*, to protect sensitive data and ensure secure distributed processing is paramount. *Federated learning* [241] enables training models across decentralized edge devices or servers holding local data samples, without exchanging them. This approach is particularly useful in scenarios where centralized data collection is impractical or raises privacy concerns. But it does not suffice for preserving data privacy as FL present several vulnerabilities [242]. Maintaining data privacy therefore takes other approaches such as *secure multi-party computation* and *differential privacy* to ensure confidentiality. *Secure multi-party computation* (SMPC) is a cryptographic technique that enables multiple parties to jointly compute a function over their inputs while keeping those inputs private. Recently, approaches that adapt SMPC to machine learning for data privacy have been released [243]. *Differential privacy* is a privacy framework which consist in maintaining data privacy while sharing information about a group of individuals. One approach to

obtain privacy is to add artificial noise to model parameters [244]. The combination of these privacy-preserving techniques allows for collaborative learning and inference on the cloud-to-edge continuum while prioritizing individual privacy concerns.

Interpretability and Explainability: distributed LLMs may produce complex and difficult-to-interpret outputs, raising concerns about model transparency and interpretability. Incorporating interpretability techniques, developing explainability frameworks, and ensuring transparency in model decision-making processes is crucial for distributed LLMs. LLM explainability can be done by two means: the local and the global explainability [245]. The first concerns the explanation of the predictions generated by the LLM. The role of explanation is to clarify the process by which the model generated a specific classification. In contrast to local explanations, global explanations provide a deeper understanding of the internal mechanisms of language models. The goal of global explanations is to comprehend the encoded information and elucidate the knowledge and linguistic properties acquired by the individual components, including neurons, hidden layers, and larger modules.

As 6G networks have the potential to drive unprecedented levels of connectivity and computational capabilities, the native integration of LLMs into a distributed 6G architecture provides for unique opportunities in faster, privacy preserving and distributed training and fine-tuning of advanced models with applications in benefiting from communications and artificial intelligence.

Distributed applications across the Cloud-to-Edge continuum

We anticipate significant impacts of integrating LLMs into a 6G network spanning from cloud to edge computing. Several use cases that cannot be adequately addressed by the current 5G network [246] [226] [247].

For example, **Industry automation** involves using control systems and information technologies to manage processes and machinery, replacing human intervention, while smart environments encompass spaces like homes, cities, or industrial settings enhanced with embedded sensors, actuators, and interconnected devices that analyze data in real-time and respond intelligently to improve efficiency, convenience, safety, and sustainability. These environments leverage IoT, AI, data analytics, and communication networks to autonomously monitor, manage, and control various aspects. Integrating LLMs and 6G technology significantly enhances these settings by enabling communication across the cloud-to-edge continuum, where 6G's speed allows instant responses and LLMs provide contextual understanding for intuitive interactions. A major challenge in smart environments and industrial automation is achieving device interoperability [248]; LLMs help address this by generating natural language that facilitates device coordination. Efficient IoT device management is crucial for optimizing automated processes [249], and recent research highlights the use of LLMs in creating reasoning agents with advanced cognitive abilities that can manage IoT devices by breaking down complex tasks [250]. The introduction of 6G technology fosters high-speed, real-time data exchange, enabling synchronized operations and reducing downtime. In smart grids, LLMs can leverage their predictive capabilities to optimize grid operations by analyzing real-time data and suggesting adjustments to power distribution. The minimal latency of 6G (1ms) is essential for maintaining phase coherence between electricity suppliers [251].

Augmented reality (**AR**) overlays digital information, such as images, videos, or 3D models, onto the real-world environment, while **Virtual Reality** (**VR**) creates an immersive digital environment that simulates physical presence in a virtual world. The ultra-low latency and high data rates of 6G will significantly enhance AR and VR experiences by making them more interactive, realistic, and responsive, bridging the physical and virtual worlds. The high data rate of 6G connections [252] combined with LLMs allows for learning traffic and channel activities to monitor the connection between avatars in VR and physical humans in AR. Immersion remains a central challenge for the metaverse, and LLMs can play a crucial role in addressing this by providing dynamic and responsive dialogue, enhancing user engagement. In virtual worlds, Non-Playable Characters (NPCs) will benefit from LLMs, enabling them to communicate fluently in multiple languages and allowing natural interactions with users [253]. Furthermore, LLMs offer powerful language translation capabilities, although current models still face challenges with cultural-specific translations [254].

The tactile internet envisions an advanced form of the internet that delivers real-time haptic feedback, enabling highly responsive and interactive communication between humans and machines, as well as between machines themselves. This concept relies on ultra-low latency and high reliability, core features of 6G technology, to support applications like remote robotic control and haptic feedback systems. In these scenarios, 6G-enabled IoT device management ensures instantaneous responses and optimal performance. Haptic feedback involves two main components: *kinesthetic feedback*, which provides data on velocity, force, and position; and *tactile feedback*, which includes information about texture, surface characteristics, and friction [255] [256]. Haptic interfaces can also incorporate visual feedback and remote response capabilities [257], making the user experience consistent regardless of local or remote operation.

Implementing 6G for the tactile internet necessitates new network structures, such as decoupling control from the data plane and using Software-Defined Networking (SDN) for centralized control. This approach, along with Network Function Virtualization (NFV), enhances network scalability and reliability [258].

Holographic communication enables participants to interact using holograms—three-dimensional images formed by the interference of light beams. Although still in its early stages, this technology faces significant challenges, such as limited bandwidth. However, 6G technology's ultra-high speeds, minimal latency, and extensive device connectivity could dramatically enhance its potential. Holographic communication may require data transmission rates of up to 2Tbps per second, which current network infrastructures cannot support [259]. Therefore, data compression is crucial for enabling this form of communication. LLMs could serve as powerful tools for encoding and decoding various data types, such as spatial coordinates, color information, and depth maps from holographic images [260]. By compressing data into a compact form using language generation capabilities and decompressing it back, LLMs can handle complex data tasks. However, research on LLMs as encoders for non-textual data is limited, making it challenging to evaluate their efficiency compared to existing compression techniques [261].

EHealth (**electronic health**) leverages digital technologies to enhance healthcare efficiency, accessibility, and quality, particularly through cloud-to-edge architectures that provide real-time insights and monitoring [262]. Effective communication among diverse devices is critical, with 6G's ultra-high bandwidth reducing delays and enabling seamless, real-time interactions between patients and providers. Key advancements like sub-millisecond jitter precision [252] [246] and low-latency network slicing enhance remote teleoperation, allowing precise medical procedures to be conducted from afar [263]. The role of LLMs in analyzing unstructured complex medical data and providing insights could facilitate seamless communication between healthcare devices [212] by interpreting and standardizing data across diverse protocols [213], enhancing interoperability [264].

The issues and and applications discussed above all bring us to the central question of the usefulness of the solutions due to energy consumption. As outlined in Section , the integration of LLM networks into 6G systems also raises significant concerns regarding their energy usage and environmental impact [265]. Zhang *et al.* [266] highlights the need to optimise carbon emissions over the entire life-cycle of any AI-based network implementation. To reduce the energy consumption of LLMs, careful planning is essential. For example, energy production and task allocation emerge as crucial factors in reducing energy consumption.

Enhancing 6G Immersive Video Streaming Services with LLMs

Immersive communication is a key usage scenario highlighted in the ITU's framework for IMT-20302, focusing on the transmission of immersive videos like VR, AR, 360°, and point cloud videos, which present unique challenges compared to traditional media. These videos include multiview and free-viewpoint video [267], which allow users to switch perspectives but place heavy demands on networks [268], [269], [270]. 360° videos [271], stitched from multiple camera angles, require significant bandwidth, especially at high resolutions, and demand low latency to prevent user discomfort [272]. Volumetric videos, including point cloud [259] and light field methods, offer six degrees of freedom, creating fully immersive environments but require extensive bandwidth, particularly with high-resolution and multisensor setups [273].

The future of immersive streaming is closely linked with the capabilities of LLMs like GPT, which can enhance service quality in areas such as bandwidth prediction [274], FoV (field of view) forecasting [275], [276], [277], and encoding [278], [279]. LLMs can predict future bandwidth needs more accurately, improving network efficiency. They can also enhance FoV prediction, crucial for VR and volumetric videos, and refine video encoding to optimize transmission.

As 6G technology evolves, LLMs will play a crucial role in enhancing multimedia services, offering personalized user experiences, and improving infrastructure [280]. Their linguistic abilities enable better content caching, service discovery, and even content creation through natural language interaction. However, incorporating LLMs into 6G presents challenges, such as ensuring data privacy and security, managing real-time processing demands, achieving energy efficiency, and maintaining interoperability. Ethical concerns, such as AI biases and cultural impact, must also be addressed. Despite these challenges, LLMs offer promising benefits, including transforming user interfaces, improving accessibility, managing content more effectively, and revolutionizing education through enhanced interactivity and personalization.

2. https://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2030/Pages/default.aspx

6 Security and Resilience

The swift adoption of Generative AI (GenAI) and Large Language Models (LLMs) across diverse sectors, including education and healthcare, signifies a monumental leap forward in technological innovation. These advancements offer unprecedented opportunities for enhancing learning experiences, streamlining information processing, and facilitating more effective healthcare solutions. Yet, alongside these benefits, the rapid proliferation of GenAI and LLMs has unveiled a critical, yet often overlooked, dimension: the emergence of security vulnerabilities. As the ecosystem encompassing both offline and online models continues to grow, incorporating a myriad of tools, browser extensions, and third-party applications, the potential for security risks escalates correspondingly. This expansion not only broadens the attack surface but also introduces complex challenges in safeguarding these technologies against exploitation. In the era of 6G and beyond, where connectivity and computational power are greatly enhanced, the avenues for adversaries to infiltrate and manipulate LLMs for nefarious purposes have multiplied. This evolving landscape necessitates a concerted effort to address these security concerns, ensuring the safe and ethical utilization of GenAI and LLMs.

As we navigate these advancements, it becomes imperative to develop robust security measures and protocols that can shield these technologies from potential threats, safeguarding the integrity of the innovations that stand to revolutionize sectors as vital also as the use to develop 6G. Due to the white paper's limited capacity for content, the extended version for this section can be found from [281].

In this section, our examination zeroes in on the security dimensions of LLMs, through the lens of potential adversaries. We delve into their aims and tactics, aiming to provide a thorough analysis of recognized security vulnerabilities associated with LLMs. Our exploration will unfold a detailed threat taxonomy that classifies various adversarial behaviors, offering insights into the array of security challenges. Furthermore, our investigation extends to the strategic incorporation of LLMs into cybersecurity measures undertaken by defensive teams, commonly referred to as blue teams. This integration is pivotal for bolstering defense mechanisms against sophisticated cyber threats. Building upon this foundation, we introduce and consider the concept of LLMSecOps inspired from Security Operations (SecOps) within practical scenarios, with a particular emphasis on its relevance and application in the burgeoning 6G landscape. An integral part of our discourse also revolves around the innovative convergence of LLMs with blockchain technology. We posit that this fusion holds the promise of pioneering next-generation, autonomously operating security solutions. Our objective is to craft a comprehensive cybersecurity strategy that spans the entire computing spectrum. By doing so, we aim to significantly reinforce the digital security infrastructure, ensuring a robust defense against emerging and evolving cyber threats.

Our comprehensive analysis, drawing from academic research, proof-of-concept studies, and renowned cybersecurity resources like the Open Web Application Security Project (OWASP), aims to equip LLM stakeholders with a detailed, actionable road map. This guide focuses on enhancing defense strategies informed by threats to LLM applications. Furthermore, the development of a threat taxonomy, specifically for GenAI and LLMs, will significantly enhance the robustness of novel frameworks like synergizing LLMs or autonomous LLM agent swarms. By categorizing potential adversarial behaviors, this taxonomy empowers the framework to proactively address security vulnerabilities, thereby strengthening the security and resilience of the 6G network ecosystem.

Attacks - Red teaming

This section looks at current vulnerabilities in the field and develops a comprehensive taxonomy, differentiating between the types, objectives, and strategies. This taxonomy will be instrumental in informing and guiding the application of LLMs in the 6G computing continuum.

Recently, OWASP has convened a multidisciplinary team of experts from a broad spectrum of disciplines, such as security, Artificial Intelligence (AI), software development, and industry leadership [282]. This coalition's objective is to systematically identify and underscore the critical security and safety challenges that both developers and security professionals need to be aware of when they are integrating LLMs into application development. Below is the preliminary compilation of critical vulnerability categories related to AI applications developed using LLMs: 1) Prompt Injection, 2) Insecure Output Handling, 3) Threats of Training Data Poisoning, 4) Model Denial of Service Attacks, 5) Supply Chain Concerns, 6) Disclosure of Sensitive Information, 7) Design Flaws in Insecure Plugins, 8) Excessive Agency in Models, 9) Overreliance on AI Models, 10) Model Theft.

Furthermore, numerous review studies have also aimed to explore the limitations, challenges, potential risks, and opportunities presented by GenAI in the realms of cybersecurity and privacy [283], [284]. According to Yao et al. [285], these vulnerabilities can broadly be categorized into two main types: AI-inherent vulnerabilities and non-AI inherent vulnerabilities, detailed in Table 1.

Defense - Blue teaming

In the last few years, the exploration of LLMs for cybersecurity operations has significantly advanced. Yao et al.[285] conducted a detailed examination and analysis of 279 research papers from 2021 to 2023, investigating the relationship between LLMs and security and privacy concerns before pointing out strategies for LLM training safety. Moreover, to enhance our understanding of the versatility of LLMs across different cybersecurity operations, this section provides a review of existing research results applicable to cyber defense (i.e., blue) teams.

Strategies for LLM training safety

Due to the border security aspects of non-AI inherent vulnerabilities, [285]'s discussion focuses on strategies to improve LLM training safety. Below is the set of strategies to mitigate LLM vulnerabilities regarding AI inherent.

LLM training: The development of LLMs encompasses intricate decisions regarding model architectures, the selection and preparation of training data, and the adoption of specific optimization techniques. Each of these components plays a crucial role in ensuring the security and privacy of LLMs throughout their lifecycle.

- **Model architectures and privacy preservation:** The management of storage and organization within LLM architectures is paramount for maintaining data privacy. Recent studies have shown the effectiveness of incorporating differential privacy techniques during the training phase to safeguard user data [315]. Additionally, enhancing LLMs' resilience against adversarial attacks has been a focus of ongoing research [316].
- **Knowledge integration for enhanced trust:** Incorporating external knowledge sources, such as knowledge graphs [317], into LLMs can significantly improve their trustworthiness and cognitive robustness [318]. These enhancements not only contribute to the models' understanding of complex concepts but also bolster their defenses against misleading information.
- **Corpora cleaning for bias reduction:** The quality of the training corpora is fundamental to preventing bias and ensuring high-quality data input. Rigorous corpora cleaning processes are essential for eliminating biases and enhancing the contextuality and accuracy of the training data [319], [320].
- **Optimization techniques for secure learning:** Optimization strategies influence how LLMs interpret and learn from data, directly impacting their security. Adversarial training methods have been developed to train LLMs to withstand malicious inputs [321], [322]. Furthermore, aligning LLMs' objectives with safety principles through human feedback has emerged as a promising approach to mitigate unintended harmful behaviors [323], [324].

LLM inference: Deploying LLMs within systems that interact with users in real time necessitates a comprehensive security strategy. This strategy should encompass three critical phases: prompt pre-processing, abnormal detection, and response post-processing. By meticulously implementing safeguards at each phase, we can significantly enhance the security of LLM interactions, potentially unlocking new possibilities for LLM connectivity and distributed applications.

- **Prompt pre-processing:** The initial phase focuses on mitigating risks from potentially malicious user inputs, commonly associated with jailbreak attacks. Strategies include: (1) Instruction Manipulation Prevention: Implementing checks to identify and neutralize attempts to alter instructions in a way that could compromise the system [325]; (2) Defensive Demonstrations: Utilizing examples of secure and compliant interactions to guide the LLM away from fulfilling harmful requests [326]; (3) Purification of Inputs: Applying techniques to cleanse input data of any elements that could lead to undesirable outputs [327].
- **Malicious detection:** This phase involves a thorough analysis of the LLM's outputs based on the given inputs to identify any prompt injection threats or backdoored instructions. Recent advancements include:

(1) Backdoored Instruction Detection: Techniques proposed by Sun et al. [328] focus on identifying hidden malicious commands embedded within seemingly benign inputs; (2) Abnormal and Poisoned Instruction Detection: [329] introduce methods to detect and mitigate the impact of abnormal or poisoned instructions, ensuring the integrity of the LLM's outputs.

- **Response post-processing:** Before presenting the generated responses to users, a final verification step is crucial, particularly, studies [330], [331], pointing assessing harmfulness and confidence, suggest mechanisms for evaluating the potential harm and reliability of LLM responses, ensuring that outputs are both safe and contextually appropriate.

Taxonomy and LLMSecOps applications

Specifically, Sultana et al. [284] systematically assessed the role of LLMs in cyber operations, exploring their utility in key cyber defense tasks. By analyzing literature on network defense, including Cyber Threat Intelligence (CTI) analysis, log management, anomaly detection, and incident response, they developed a four-tier taxonomy. This taxonomy organizes the extensive capabilities of LLMs into distinct cyber operations categories, closely aligned with the widely adopted version 1.1 of the National Institute of Standards and Technology (NIST) Cybersecurity Framework [332].

- **Identify** focuses on LLMs for identifying and classifying threats from open-source CTI, enhancing early threat intelligence efforts.

Vulnerability Type	Description
AI-Related Vulnerabilities	
Adversarial Attacks	Manipulate input data to affect performance, includes data poisoning [286], [287], [288] and backdoor attacks [289], [290], [291].
Inference Attacks	Deduce sensitive information about the model and its data, includes attribute inference attacks [292], [293], [294] and membership inferences [295], [296], [297], [298].
Extraction Attacks	Obtain specific resources or information, includes model stealing [299], [300], gradient leakage [301], and training data extraction [302], and [303] points out the possibility to replicate the model without accessing the original model data.
Bias and Unfair Exploitation	Arises from biases in training data or model fine-tuning [304], leading to misinformation [305], [306] and reinforcing stereotypes.
Instruction Tuning Attack	Manipulate models to execute unintended actions or bypass limitations, includes DoS [307], prompt injection [308], [309], [310], and jailbreaking [311], [312].
Zero-day Attacks	Backdoor vulnerabilities that act as "sleeper agents," activated by specific triggers.
Non-AI-Related Vulnerabilities	
Remote Code Execution	Execute arbitrary code remotely by exploiting vulnerabilities, leading to data theft, system control, and disruption [313].
Side Channel	Gather information through observable patterns or phenomena, not directly exploiting LLM vulnerabilities.
Insecure Plugin	Target vulnerabilities in plugins due to poor security practices or design flaws [314], leading to data theft or unauthorized actions [282].

TABLE 1: Summary of AI and Non-AI-Related Vulnerabilities

Project	Technology	Goal	Notable Outcomes
PentestGPT [343]	GPT-3.5, GPT-4	Enhance penetration testing	Improved task completion by 228.6% (GPT-3.5) and 58.6% (GPT-4)
PAC-GPT [344]	GPT-3	Generate synthetic network traffic	High accuracy in generating realistic network activities
TSTEM [345]	BERT, Longformer, BERT-NER	Real-time CTI collection	Over 98% accuracy in IOC identification and extraction
GPT-2C [346]	GPT-2	Improve IDS via log analysis	89% inference accuracy on honeypot logs
LogBERT [347]	BERT	Anomaly detection in system logs	Outperforms in anomaly detection across datasets
LogPPT [348]	RoBERTa	Log parsing with few-shot learning	Superior parsing accuracy and efficiency
LogBot [349]	GPT-3	Enhance chatbot performance in cybersecurity	Over 99% accuracy in anomaly detection
Cyber Sentinel [350]	GPT-4	Cybersecurity dialogue system	Integrates LLMs for proactive / reactive security measures
HuntGPT [351]	GPT-3.5 Turbo, Random Forest	Intrusion detection dashboard	Enhances decision-making with XAI and conversational AI

TABLE 2: Summary of Cybersecurity Innovations Using LLMs

- **Protect** uses LLMs for vulnerability scans, security assessments, and automating defense strategies, bolstering network security.
- **Detect** applies LLMs to detect vulnerabilities, extract malware, and classify attacks, highlighting their role in early threat identification.
- **Respond** leverages LLMs in incident response and recovery, aiding in analysis and strategic planning postincident.

Furthermore, the section presents an in-depth analysis of diverse systems and tools designed to leverage LLMs with the aim of enhancing cybersecurity operations, across the Identify, Protect, Detect, and Respond phases of the NIST framework. This exploration is focused on understanding how these technological advancements are integrated and their significant contribution to the cybersecurity domain. We specifically focus on demonstrating the capabilities of LLMs within the SecOps framework, emphasizing their role in the advancing 6G edge-cloud continuum. The section highlights the creative applications of LLMs in strengthening security infrastructures and enhancing response strategies.

LLMSecOps in the 6G era

The integration of AI into communication networks, notably within the 6G era, signifies a transformative shift towards "AI Interconnect" and "AI-native Telecom" paradigms or autonomous LLM agent swarms. This evolution, as highlighted in recent studies [1], [205], introduces a dual spectrum of vulnerabilities: those inherent to AI and those not specific to AI technologies [285]. In addressing these challenges, the insights provided by OWASP [282] and the governance surrounding LLMs emerge as pivotal considerations for advancing secure communication networks. The incorporation of LLMs marks a significant advancement in the telecommunication sector's vision. Nevertheless, ensuring the security and trustworthiness of LLM usage necessitates enhanced verification measures.

As research on the 6G computing continuum is still in its early phases, there is an expectation that the resulting reference architecture will integrate Key Enabling Technologies (KETs) like Network Function Virtualization (NFV), edge intelligence, Software-Defined Networking (SDN) across 5G Core, Cloud, and Edge networks. Furthermore, the system's architecture will be aligned with international standards, ensuring not only its full autonomy but also seamless interoperability with existing legacy systems [342]. In this scenario, LLMs will offer support for intelligent decision-making, evolving SecOps into fully autonomous cognitive LLMSecOps services seamlessly integrated within network functions.

Intent-based networking.
Intent-based networking (IBN) [343] is an emerging paradigm that aims to automate network configurations using AI. The basic idea of IBN is to enable network administrators to manage complex networks with business intents. Business intents can be high-level business objectives of organizations. For example, an organization can prioritize one type of traffic, such as video traffic, over other types of traffic, such as text. IBN requires developments in AI, such as LLMs, to effectively convert such intents into configurations. 6G will utilize LLMs to enable run-time network configurations through high-level intents to simplify human-network interactions and smoothen the deployment of new services. However, IBNs can have several security challenges [344], that can be described along with the process or workflow of IBNs. The Internet Engineering Task Force (IETF) describes intents to be abstract, declarative, and vendor agnostic set of rules that can be deployed through several steps of a properly defined process [345]. A pictorial presentation of the process and flow of IBN is presented in Fig. 16.

The first component or process in IBNs is intent profiling where a network administrator expresses intent to an IBN system. This process must be user-friendly and the system must facilitate the user for meaningful intents. The second step is intent translation or compilation, where intents are transformed into low-level network configurations. LLMs can play a major role in these steps, to effectively transform service requests into network configurations. A major security risk here is that if the LLMs are compromised, the whole network can be compromised, for instance through malignant configurations. Sensitive traffic in this case can be diverted to malicious nodes for compromising privacy and security. However, if the next step, intent resolution is properly carried out, miss-configurations can be recognized. Since miss-configurations are responsible for a majority of network security challenges arising from human-network interactions [346], LLMs-based configurations can minimize such vulnerabilities, given correct intent profiling carried out through LLMs. The next steps in IBN, intent activation which provides the necessary services intended by intents, and intent as-

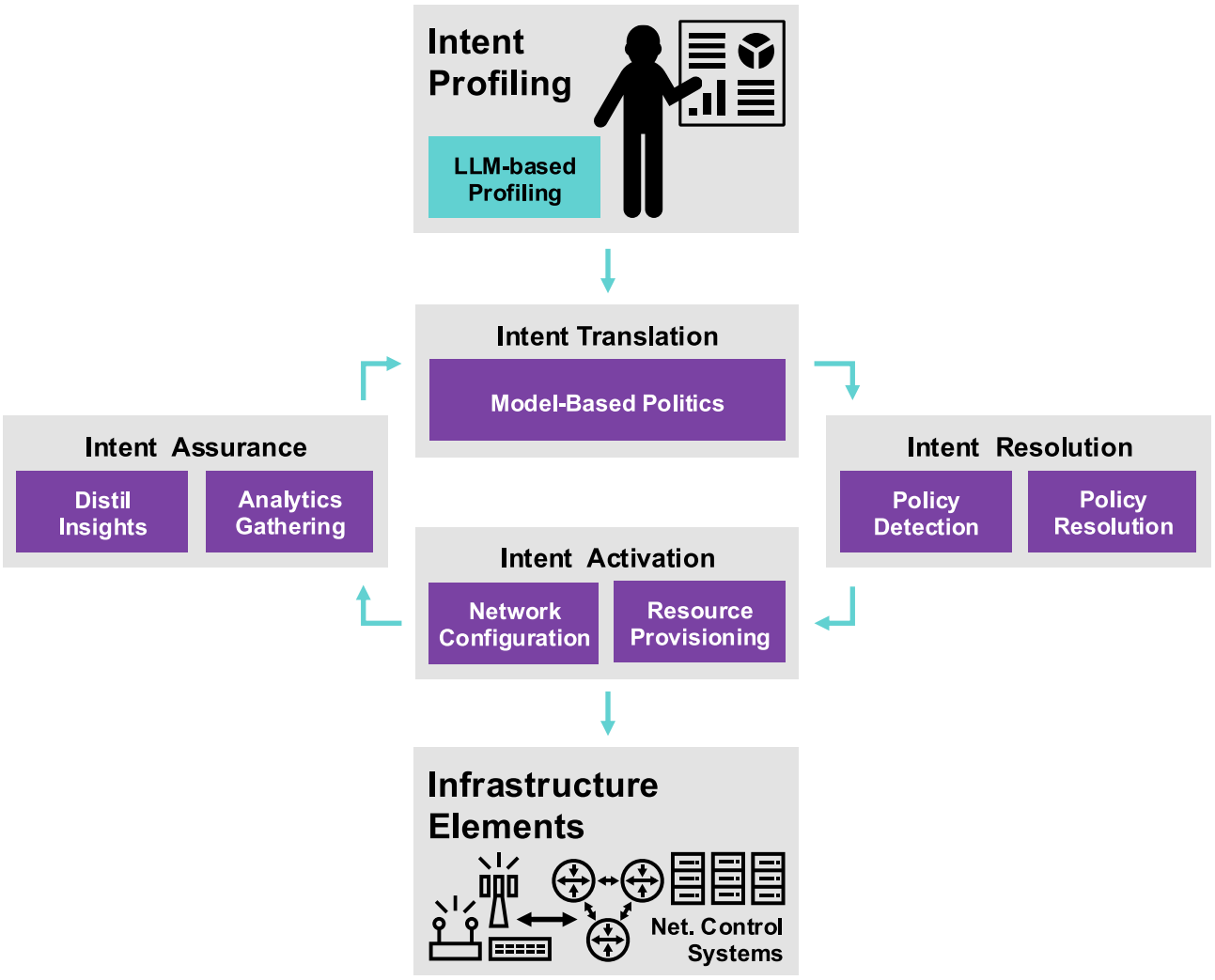


Fig. 16: Process and components of IBN.

surance which monitors the intents throughout its life cycle to ensure that the network behaves as intended, depend on the initial steps. Therefore, the security of LLMs-based intent profiling and translation is extremely important for IBN, and thus 6G.

Network Data Analytics Function.

The 3GPP Network Data Analytics Function (NWDAF) [347], envisioned as a key component of the 6G computing continuum's service base [205], is the analytics hub that establishes a foundation for AI/ML-driven data analytics operations and services [78]. This hub aims to ensure their complete integration and interoperability across the network. NWDAF is designed to aggregate and analyze data related to network efficiency, User Equipment (UE) behavior, service usage patterns, network Operations, Administration, and Maintenance (OAM) spanning the computing continuum, including WiFi, WAN, 5G Core, Cloud, and Edge networks [348]. Specifically, this function is designed to perform analytical inferences, provide services for training machine learning models, and make data accessible via the Nnwdaf AnalyticsInfo service. This enables native downstream LLMSecOps services to acquire targeted analytics from NWDAF [78].

Zero-Touch Network 5G/6G Security.

The Zero-touch Network and Service Management (ZSM) is focused on revolutionizing network management towards a fully automated, flexible, and efficient approach. This extends across all mobile network generations, emphasizing the necessity for networks to autonomously self-configure, self-monitor, self-heal, and self-optimize without human intervention [349]. This automation is crucial as networks evolve, especially with the complexity introduced by 5G and future 6G technologies, supporting a vast number of connected devices and services with diverse network requirements [350], [351]. ZSM contains advanced technologies like AI and ML for intelligent decision-making, as well as SDN and NFV for simplifying network management, indicating a shift towards networks that can independently adapt to and defend against cyber threats, ensuring robust and resilient future telecommunications infrastructure.

The integration of Zero-touch Network Management (ZTM) into the ZSM framework is essential to improve network management's security and efficiency in 5G and 6G networks. The ZSM initiative aims for full end-to-end network automation, and ZTM plays a key role in achieving this goal [352]. The framework can enable end-to-end network slicing and AI-based security mechanisms, ensuring a secure infrastructure that caters to diverse service requirements. Furthermore, the adoption of AI and ML optimizes network operations and introduces new security paradigms, addressing challenges posed by SDN, NFV, MEC, and network slicing. These advancements are crucial for the creation of secure, autonomous systems capable of proactively identifying and mitigating threats, thus safeguarding the integrity and resilience of future mobile networks [353].

The anticipated complexity of future 6G networks will escalate to levels where conventional analytical and numerical simulation methods will become impractical [353]. This necessitates the ZSM framework reference architecture to embed data protection mechanisms for use, transit, and storage. This ensures an elevated security standard across management functions, services, and infrastructure resources while safeguarding data security and integrity [354]. Therefore, integrating Zero-Touch with LLMs can enhance the security of 6G networks. This integration could help create a future where networks can autonomously defend against cyber threats through intelligent, adaptive, and automated mechanisms [355]. These systems would continuously learn from new data, adapt to emerging threats, and implement security measures without manual configuration or intervention [356]. As a result, a robust and resilient 6G infrastructure can be ensured.

Autonomous LLM Agent Swarms

This section examines the current use of LLMs in decentralized defense applications and uses these technologies as a basis to propose a forward-looking perspective on the

future of distributed LLM or LLM multi-agent systems. This analysis is crucial to supporting the AI Interconnect's commitment to integrating security and trust principles from the outset, thereby creating a more robust and trustworthy framework for future technological deployments.

The transition to distributed LLMs.

The prevailing model for LLMs is predominantly centralized, managed by singular organizations. This centralization introduces critical challenges, including discrepancies in model design and the utilization of potentially biased or sub-optimal training data. To address these issues, a decentralized architecture for LLMs is proposed, leveraging a network of LLMs to validate responses and offer a diversity of perspectives. This decentralization not only mitigates the concerns associated with centralized models, such as privacy risks, usage restrictions, and dataset biases, but also fosters greater openness and inclusivity in contributions. Gao et al. [357] advocate for a peer-to-peer decentralized network of LLMs, positing that such a structure could enhance robustness and trustworthiness, driving the model towards greater impartiality and openness. Their proposal underscores the potential of decentralized LLMs to overcome the inherent limitations of traditional, centralized models. Similarly, Tang et al. [358] contribute to this discourse by highlighting the advantages of distributing certain computational tasks to the client side.

This approach not only serves to preserve privacy but also optimizes the processes of pre-training, inference, and fine-tuning of LLMs, offering a novel perspective on the deployment and utilization of LLMs in a manner that prioritizes data security and user privacy.

Security and trust in distributed LLMs

Blockchain technology emerges as a pivotal infrastructure in fostering collaborative AI model development and interconnecting LLMs to establish a decentralized AI marketplace, especially the formation of blockchain-based LLM multi-agent systems [359]. It primarily offers a mechanism to cultivate trust through the integrity and availability of secure, decentralized knowledge bases during LLM interactions. Moreover, blockchain's contribution to sustainability through incentive mechanisms and reputation systems sets a precedent for further integration and development.

Gong [360] introduces an innovative concept of a decentralized LLM framework built upon blockchain technology, aiming to imbue LLMs with dynamic capabilities. This model suggests that blockchain not only transitions LLMs from centralized to decentralized architectures but also facilitates their real-time, continuous training. Furthermore, Gong highlights how blockchain can underpin decentralized datasets and economic incentives, thereby fostering an open, collaborative environment for LLM model contributors and validators. This openness and trust paradigm shift, enabled by blockchain, is pivotal for the evolution of dynamic LLMs.

Another noteworthy contribution by [361] underscores blockchain's potential in affirming data rights and reshaping profit distribution mechanisms, further emphasizing the technology's transformative impact on LLM ecosystems. To address security concerns within distributed LLM systems, [362] has introduced an innovative approach that integrates Trusted Execution Environments (TEEs) to create a secure, distributed LLM framework based on model slicing. Their methodology not only emphasizes the importance of maintaining communication performance and model accuracy but also introduces a novel way of securing the most vulnerable segments of the training model. By deploying the sensitive parts of the model within TEEs at either the sending or receiving ends of the model exchange process, and safeguarding this exchange through robust encryption and decryption mechanisms, Huang et al. offer a promising solution to the dual challenges of ensuring security and preserving the integrity of distributed LLM systems.

Autonomous defense framework.

Upon establishing secure and trustworthy LLM and their integration into autonomous swarms of LLM agents, the concept of an autonomous defense framework leveraging LLMs for enhanced cybersecurity emerges as a compelling area of interest. In light of this perspective, we have delineated a comprehensive overview of an autonomous defense framework, employing these swarms of LLM agents, organized into a four-tier taxonomy designed for cyber operations. While this initial framework sketch does not explicitly detail the incorporation of technologies such as blockchain or TEEs, their critical roles in fortifying the framework's trustworthiness and security are implicitly recognized. The integration of such technologies is envisaged to significantly contribute to the robustness and efficacy of the autonomous defense framework, ensuring a secure and resilient cyber operational environment.

The sequence diagram depicted in Figure 17 represents an illustration of an autonomous defense framework utilizing LLM agent swarms. This framework is designed to identify and respond to adversarial actions instigated by LLM, such as PentestGPT [333] or PAC-GPT [334], within cyber environments. Its primary focus is the protection of 6G edge-cloud infrastructure through the deployment of explainable and actionable AI, along with conversational agents designed for human interaction (e.g, HuntGPT [341] and Cyber Sentinel [340]). The processes are segmented into four main components reflecting the cyber defense lifecycle: Attack Surface, Initial Detection and Analysis, Response Generation and Execution, and Decision Support and Communication.

Open research issues

Drawing from our comprehensive summarization and analysis, this section concludes with a curated collection of research questions aimed at exploring the secure and safe utilization of LLMs within the 6G edge-cloud continuum. These pivotal research inquiries are categorized into three primary areas: (RQ 1.) Ensuring safety during the training of LLMs; (RQ 2.) Optimizing the integration of LLMs within SecOps to evolve into an effective LLMSecOps framework; and (RQ 3.) Investigating the trustworthiness and security mechanisms in autonomous LLM agent swarms, alongside their potential applications. This structured inquiry seeks to address the critical aspects of LLM deployment, emphasizing safety, efficiency, and security in a rapidly advancing technological landscape.

7 Conclusion

LLMs are rapidly evolving. They are expected to become more autonomous and capable, finding wider adoption and enabling new functionalities across industries. This will redefine human-AI collaboration in ways we are only just beginning to understand and shape that could span from revolutionizing manufacturing in Industry 5.0 to enhancing personalized healthcare, streamlining logistics, or transforming education.

Furthermore, the rapid acceleration in the development of LLMs intersects directly with the emerging world of 6G networks. As we have extensively explored in this paper, the integration of these emerging AI technologies into the 6G architectural framework offers both untapped potential and inherent challenges.

In particular, the co-evolution of both LLMs and 6G appears fascinating. LLMs' ability to reason (or replicate reasoning in decision making), plan (or support planning), and grasp complex situations (in a variety of operational conditions) will likely exceed our current expectations by the time 6G networks are fully deployed. At the same time, 6G technology itself is under development, so the specific needs and functionalities of the network might change as new applications and use cases emerge, and with a number of functionalities that will actually be enabled by AI to create a true "Intelligent Internet of Intelligent Things". Therefore this co-evolution presents a unique situation and there are challenges to be considered. For instance, it is necessary to make sure LLM/GPT capabilities keep pace with the evolving needs of 6G. Early integration strategies might require adjustments as both technologies mature. Additionally, unforeseen use cases for LLMs within the 6G network might emerge, requiring the network architecture to adapt and accommodate these new possibilities.

The combined advancements of LLMs and 6G could unlock entirely new applications and functionalities that we can't even imagine today. This co-evolution has the potential to revolutionize areas like intelligent network management, self-optimizing infrastructure, and hyper-personalized user experiences. Furthermore, as LLMs become more autonomous and collaborative, the way humans interact with AI within 6G networks will transform. We might see a shift towards shared decision-making, where humans leverage AI for complex analysis and rapid decision-making while providing strategic guidance and overall objectives.

However, through our discussions, it has become evident that the integration of such technologies does not involve merely technicalities but encompasses vast landscapes of requirements and design considerations. While 6G remains predominantly in the research domain, the pace at which these models are emerging may very well outpace the development of their corresponding standards and regulations. The current absence of universally accepted standards and a consistent regulatory framework adds layers to this complexity. The situation highlights the need for clearer guidelines as we try to integrate different approaches.

Recognizing these multifaceted challenges, this paper has aimed to provide a fresh conceptual overview, shedding light on the novel intersections of such AI models with the 6G ecosystem. Furthermore, our work ventured into the practical applications and intricacies of LLMs and GPTs, offering insights and potential pathways for their seamless integration into future mobile networks.

In light of our exploration, we believe that the integration of GPTs and LLMs with 6G networks has immense promise, and a collective push towards establishing shared practices and frameworks is crucial. Such an initiative can help us navigate the potential of GenAI in 6G responsibly and with an open acknowledgment of the challenges ahead.

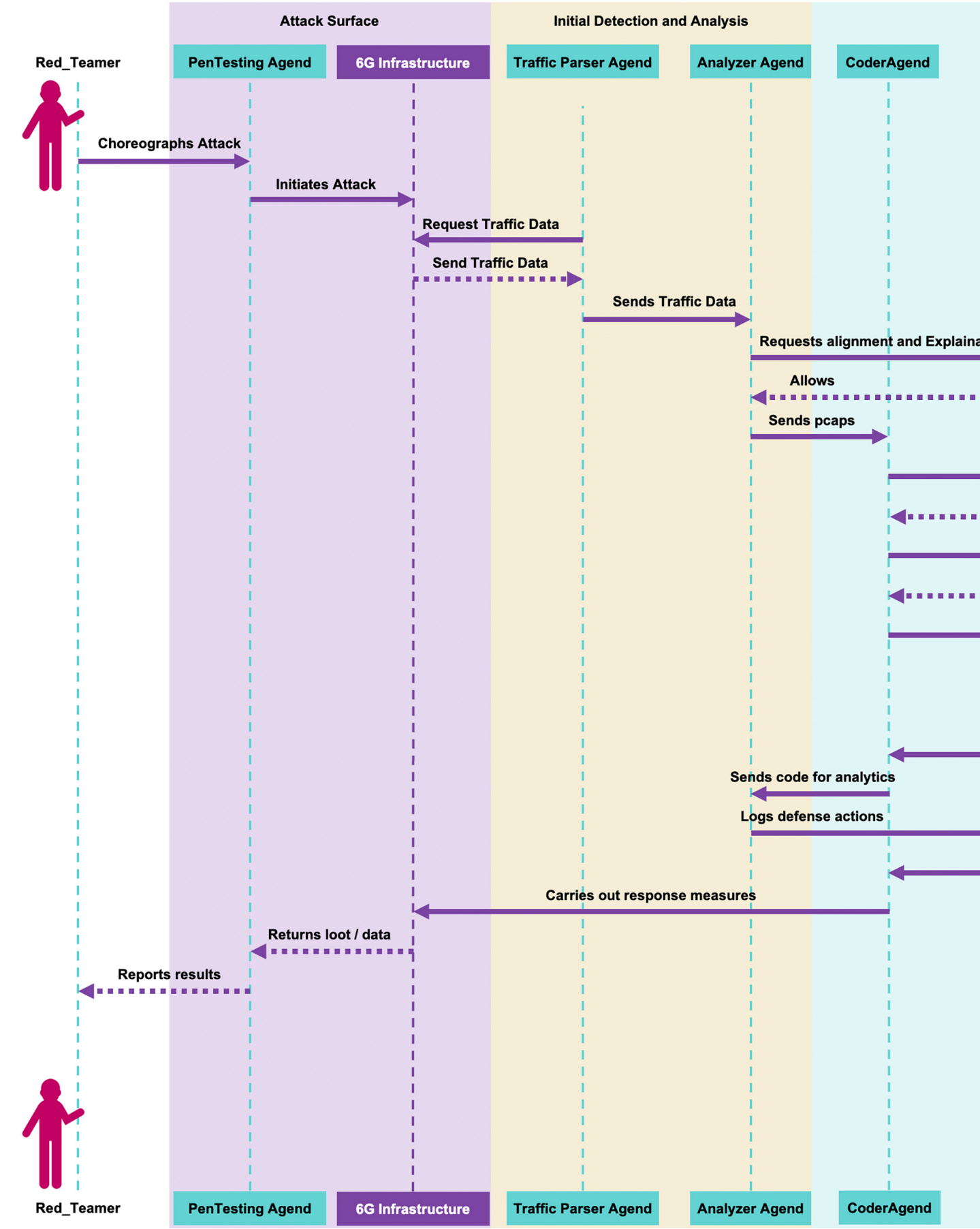
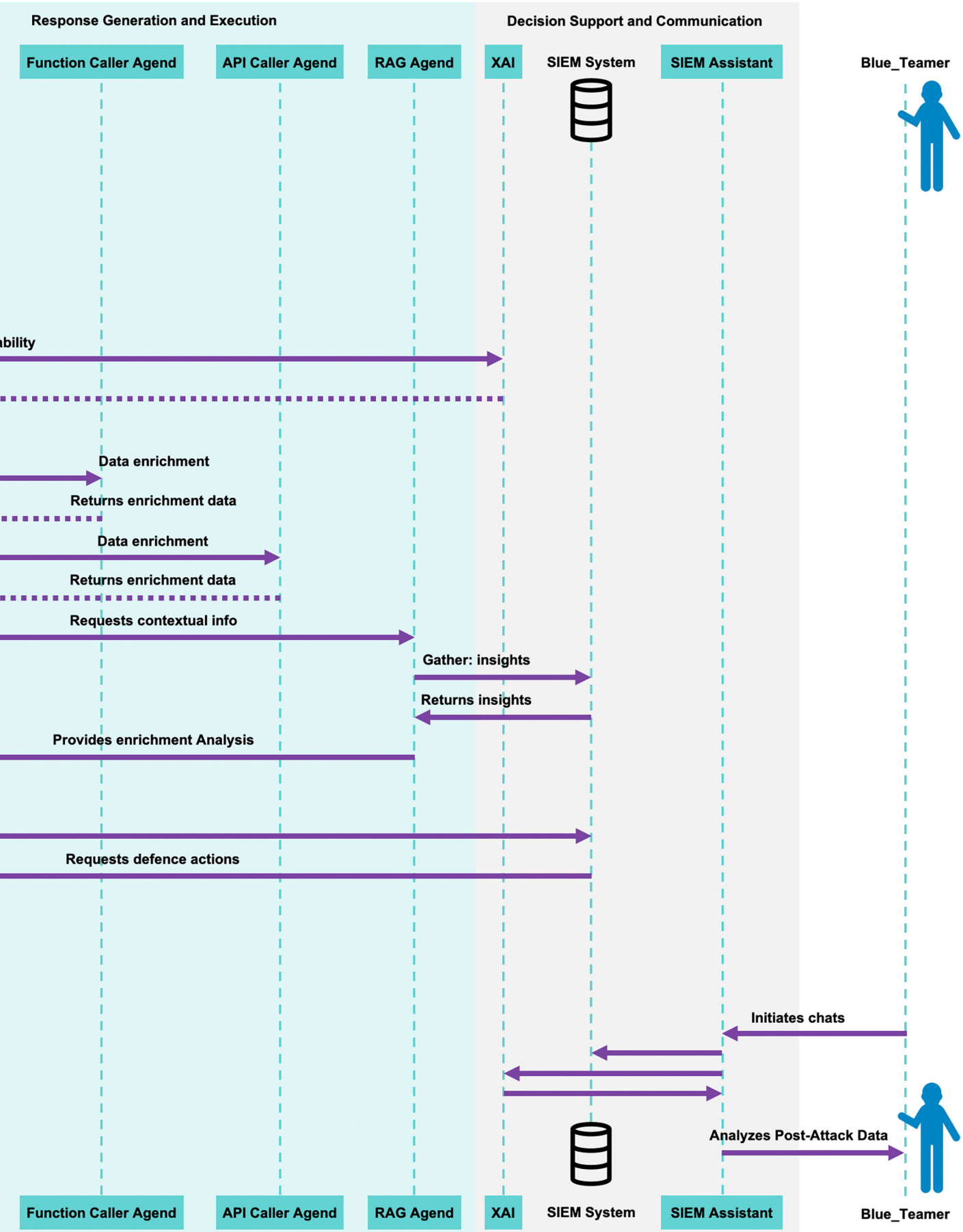


Fig. 17: Autonomous defense with LLM agent swarms.


Response Generation and Execution
Decision Support and Communication
Function Caller Agend
API Caller Agend
RAG Agend
XAI
SIEM System
SIEM Assistant
Blue_Teamer
ability
Data enrichment
Returns enrichment data
Data enrichment
Returns enrichment data
Requests contextual info
Gather: insights
Returns insights
Provides enrichment Analysis
Requests defence actions
Initiates chats
Analyzes Post-Attack Data
Function Caller Agend
API Caller Agend
RAG Agend
XAI
SIEM System
SIEM Assistant
Blue_Teamer

References

[1] R. Britto, T. Murphy, M. Iovene, L. Jonsson, M. Erol-Kantarci, and B. Kovács, "Telecom ai native systems in the age of generative ai–an engineering perspective," arXiv preprint arXiv:2310.11770, 2023.

[2] Nokia White Paper, "Toward a 6g ai-native air interface," accessed October 2023. [Online]. Available: ***https://onestore.nokia.com/asset/210299***

[3] 6G Flagship, "6g research visions white paper series," accessed October 2023. [Online]. Available: ***https://www.6gflagship.com/white-papers/***

[4] M. Latva-aho, K. Leppänen et al., "Key drivers and research challenges for 6g ubiquitous wireless intelligence," 6G Flagship – White Paper Series, 2019.

[5] M. A. Uusitalo, P. Rugeland, M. R. Boldi, E. C. Strinati, P. Demestichas, M. Ericson, G. P. Fettweis, M. C. Filippou, A. Gati, M.-H. Hamon et al., "6g vision, value, use cases and technologies from european 6g flagship project hexa-x," IEEE Access, vol. 9, pp. 160 004–160 020, 2021.

[6] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill et al., "On the opportunities and risks of foundation models," arXiv preprint arXiv:2108.07258, 2021.

[7] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong et al., "A survey of large language models," arXiv preprint arXiv:2303.18223, 2023.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., "Language models are few-shot learners," Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020.

[9] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A prompt pattern catalog to enhance prompt engineering with chatgpt," arXiv preprint arXiv:2302.11382, 2023.

[10] International Telecommunication Union (ITU), "Imt towards 2030 and beyond," ITU-R Study Group 5 (SG 5), Tech. Rep., September 2023. [Online]. Available: ***https://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2030/Pages/default.aspx***

[11] Ericsson White Paper, "Defining ai native: A key enabler for advanced intelligent telecom networks," accessed October 2023. [Online]. Available: ***https://www.ericsson.com/49341a/assets/local/reports-papers/white-papers/ai-native.pdf***

[12] L. Lovén, T. Leppänen, E. Peltonen, J. Partala, E. Harjula, P. Porambage, M. Ylianttila, and J. Riekki, "Edgeai: A vision for distributed, edge-native artificial intelligence in future 6g networks," 6G Wireless Summit, March 24-26, 2019 Levi, Finland, 2019.

[13] D. Xu, T. Li, Y. Li, X. Su, S. Tarkoma, T. Jiang, J. Crowcroft, and P. Hui, "Edge intelligence: Empowering intelligence to the edge of network," Proceedings of the IEEE, vol. 109, no. 11, pp. 1778– 1837, 2021.

[14] E. Peltonen, M. Bennis, M. Capobianco, M. Debbah, A. Ding, F. Gil-Castiñeira, M. Jurmu, T. Karvonen, M. Kelanti, A. Kliks et al., "6g white paper on edge intelligence," arXiv preprint arXiv:2004.14850, 2020.

[15] E. Peltonen, I. Ahmad, A. Aral, M. Capobianco, A. Y. Ding, F. Gil-Castineira, E. Gilman, E. Harjula, M. Jurmu, T. Karvonen et al., "The many faces of edge intelligence," IEEE Access, vol. 10, pp. 104 769–104 782, 2022.

[16] T. Meuser, L. Lovén, M. Bhuyan, S. G. Patil, S. Dustdar, A. Aral, S. Bayhan, C. Becker, E. De Lara, A. Y. Ding et al., "Revisiting edge ai: Opportunities and challenges," IEEE Internet Computing, vol. 28, no. 4, pp. 49–59, 2024.

[17] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, "6g wireless systems: Vision, requirements, challenges, insights, and opportunities," Proceedings of the IEEE, vol. 109, no. 7, pp. 1166–1199, 2021.

[18] O. Gheibi, D. Weyns, and F. Quin, "Applying machine learning in self-adaptive systems: A systematic literature review," ACM Transactions on Autonomous and Adaptive Systems (TAAS), vol. 15, no. 3, pp. 1–37, 2021.

[19] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," arXiv preprint arXiv:2210.03629, 2022.

[20] D. Townend, R. Husbands, S. D. Walker, and A. Sutton, "Challenges and opportunities in wireless fronthaul," IEEE Access, 2023.

[21] E. ISG-mWT, "5g wireless backhaul/x-haul," vol, vol. 1, pp. 1–24, 2018.

[22] A. Karapantelakis, P. Alizadeh, A. Alabassi, K. Dey, and A. Nikou, "Generative ai in mobile networks: a survey," Annals of Telecommunications, pp. 1–19, 2023.

[23] L. Bariah, Q. Zhao, H. Zou, Y. Tian, F. Bader, and M. Debbah, "Large language models for telecom: The next big thing?" arXiv preprint arXiv:2306.10249, 2023.

[24] Z. Lin, G. Qu, Q. Chen, X. Chen, Z. Chen, and K. Huang, "Pushing large language models to the 6g edge: Vision, challenges, and opportunities," arXiv preprint arXiv:2309.16739, 2023.

[25] S. Yrjola, P. Ahokangas, M. Matinmikko-Blue, R. Jurva, V. Kant, P. Karppinen, M. Kinnula, H. Koumaras, M. Rantakokko, V. Ziegler et al., "White paper on business of 6g," arXiv preprint arXiv:2005.06400, 2020.

[26] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," arXiv preprint arXiv:2306.13549, 2023.

[27] F. Jiang, Y. Peng, L. Dong, K. Wang, K. Yang, C. Pan, and X. You, "Large ai model empowered multimodal semantic communications," arXiv preprint arXiv:2309.01249, 2023.

[28] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "Next-gpt: Any-to-any multimodal llm," arXiv preprint arXiv:2309.05519, 2023.

[29] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," Proceedings of the IEEE, vol. 103, no. 9, pp. 1449–1477, 2015.

[30] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6g: Ai empowered wireless networks," IEEE communications magazine, vol. 57, no. 8, pp. 84–90, 2019.

[31] A. Tamkin, M. Brundage, J. Clark, and D. Ganguli, "Understanding the capabilities, limitations, and societal impact of large language models," arXiv preprint arXiv:2102.02503, 2021.

[32] "OpenAI," accessed October 2023. [Online]. Available: ***https://openai.com/***

[33] "LangChain," accessed October 2023. [Online]. Available: ***https://www.langchain.com/***

[34] "O-RAN ALLIANCE e.V," accessed October 2023. [Online]. Available: ***https://www.o-ran.org/***

[35] A. Y. Ding, E. Peltonen, T. Meuser, A. Aral, C. Becker, S. Dustdar, T. Hiessl, D. Kranzmüller, M. Liyanage, S. Maghsudi et al., "Roadmap for edge ai: A dagstuhl perspective," pp. 28–33, 2022.

[36] European Commission, "Ethics guidelines for trustworthy AI - Shaping Europe's digital future," accessed October 2023. [Online]. Available: ***https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai***

[37] ——, "Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts," Eur Comm, vol. 106, pp. 1–108, 2021.

[38] Y. E. Wang, G.-Y. Wei, and D. Brooks, "Benchmarking tpu, gpu, and cpu platforms for deep learning," arXiv preprint arXiv:1907.10701, 2019.

[39] J. Lin, W.-M. Chen, H. Cai, C. Gan, and S. Han, "Mcunetv2: Memory-efficient patch-based inference for tiny deep learning," arXiv preprint arXiv:2110.15352, 2021.

[40] J. Lin, L. Zhu, W.-M. Chen, W.-C. Wang, C. Gan, and S. Han, "On-device training under 256kb memory," Advances in Neural Information Processing Systems, vol. 35, pp. 22 941–22 954, 2022.

[41] M. Matinmikko-Blue, S. Aalto, M. I. Asghar, H. Berndt, Y. Chen, S. Dixit, R. Jurva, P. Karppinen, M. Kekkonen, M. Kinnula et al., "White paper on 6g drivers and the un sdgs," arXiv preprint arXiv:2004.14695, 2020.

[42] United Nations. Department of Economic and Social Affairs, The Sustainable Development Goals: Report 2022. UN, 2022.

[43] M. Katz, T. Paso, K. Mikhaylov, L. Pessoa, H. Fontes, L. Hakola, J. Leppäniemi, E. Carlos, G. Dolmans, J. Rufo et al., "Towards truly sustainable iot systems: the superiot project," Journal of Physics: Photonics, vol. 6, no. 1, p. 011001, 2024.

[44] V. Ziegler, H. Viswanathan, H. Flinck, M. Hoffmann, V. Räisänen, and K. Hätönen, "6g architecture to connect the worlds," IEEE Access, vol. 8, pp. 173 508–173 520, 2020.

[45] M. A. Uusitalo, M. Ericson, B. Richerzhagen, E. U. Soykan, P. Rugeland, G. Fettweis, D. Sabella, G. Wikström, M. Boldi, M. H. Hamon et al., "Hexa-x the european 6g flagship project," in 2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit). IEEE, 2021, pp. 580–585.

[46] C.-J. Wu, R. Raghavendra, U. Gupta, B. Acun, N. Ardalani, K. Maeng, G. Chang, F. Aga, J. Huang, C. Bai et al., "Sustainable ai: Environmental implications, challenges and opportunities," Proceedings of Machine Learning and Systems, vol. 4, pp. 795–813, 2022.

[47] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale et al., "Llama 2: Open foundation and fine-tuned chat models," arXiv preprint arXiv:2307.09288, 2023.

[48] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann et al., "Palm: Scaling language modeling with pathways," arXiv preprint arXiv:2204.02311, 2022.

[49] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "Carbon emissions and large neural network training," arXiv preprint arXiv:2104.10350, 2021.

[50] D. Patterson, J. Gonzalez, U. Hölzle, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. R. So, M. Texier, and J. Dean, "The carbon footprint of machine learning training will plateau, then shrink," Computer, vol. 55, no. 7, pp. 18–28, 2022.

[51] J. Dodge, T. Prewitt, R. Tachet des Combes, E. Odmark, R. Schwartz, E. Strubell, A. S. Luccioni, N. A. Smith, N. DeCario, and W. Buchanan, "Measuring the carbon intensity of ai in cloud instances," in Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 1877–1894.

[52] T. Sipola, J. Alatalo, T. Kokkonen, and M. Rantonen, "Artificial intelligence in the iot era: A review of edge ai hardware and software," in 2022 31st Conference of Open Innovations Association (FRUCT). IEEE, 2022, pp. 320–331.

[53] S. A. Khowaja, P. Khuwaja, and K. Dev, "Chatgpt needs spade (sustainability, privacy, digital divide, and ethics) evaluation: A review," arXiv preprint arXiv:2305.03123, 2023.

[54] K. K. Ramachandran, K. K. K, A. Semwal, S. P. Singh, A. A. Al-Hilali, and M. B. Alazzam, "Ai-powered decision making in management: A review and future directions," in 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2023, pp. 82–86.

[55] S. Stoykova and N. Shakev, "Artificial intelligence for management information systems: Opportunities, challenges, and future directions," Algorithms, vol. 16, no. 8, 2023. [Online]. Available: ***https://www.mdpi.com/1999-4893/16/8/357***

[56] Y. Bao, W. Gong, and K. Yang, "A literature review of human–ai synergy in decision making: From the perspective of affordance actualization theory," Systems, vol. 11, no. 9, 2023. [Online]. Available: ***https://www.mdpi.com/2079-8954/11/9/442***

[57] N. Borazjanizadeh and S. T. Piantadosi, "Reliable reasoning beyond natural language," arXiv preprint arXiv:2407.11373, 2024.

[58] H. Taherdoost, "Deep learning and neural networks: Decision-making implications," Symmetry, vol. 15, no. 9, 2023. [Online]. Available: ***https://www.mdpi.com/2073-8994/15/9/1723***

[59] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley, "A survey of multi-objective sequential decision-making," J. Artif. Int. Res., vol. 48, no. 1, p. 67–113, oct 2013.

[60] J. C. C. Tesolin, A. M. Demori, D. F. C. Moura, and M. C. Cavalcanti, "Enhancing heterogeneous mobile network management based on a well-founded reference ontology," Future Generation Computer Systems, vol. 149, pp. 577–593, 2023. [Online]. Available: ***https://www.sciencedirect.com/science/ article/pii/S0167739X23003084***

[61] B. Liu, J. Luo, and X. Su, "The framework of 6g self-evolving networks and the decision-making scheme for massive iot," Applied Sciences, vol. 11, no. 19, 2021. [Online]. Available: ***https://www.mdpi.com/2076-3417/11/19/9353***

[62] Y. Xiao, G. Shi, Y. Li, W. Saad, and H. Poor, "Toward self-learning edge intelligence in 6g," IEEE Communications Magazine, vol. 58, no. 12, pp. 34–40, Dec. 2020, publisher Copyright: © 1979-2012 IEEE.

[63] A. Islam and K. Chang, "Real-time ai-based informational decision-making support system utilizing dynamic text sources," Applied Sciences, vol. 11, no. 13, 2021. [Online]. Available: ***https://www.mdpi.com/2076-3417/11/13/6237***

[64] L. Mumford, The City in History: Its Origins, Its Transformations, and Its Prospects, ser. A Harbinger Book. Harcourt, Brace & World, 1961.

[65] W. Saad, M. Bennis, and M. Chen, "A vision of 6g wireless systems: Applications, trends, technologies, and open research problems," IEEE network, vol. 34, no. 3, pp. 134–142, 2019.

[66] S. Russell, Human compatible: AI and the problem of control. Penguin Uk, 2019.

[67] L. Floridi and J. Cowls, "A unified framework of five principles for ai in society," Machine learning and the city: Applications in architecture and urban design, pp. 535–545, 2022.

[68] Presidency of Council of the European Union, "Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts," 2024. [Online]. Available: ***https://data.consilium.europa. eu/doc/document/ST-5662-2024-INIT/en/pdf***

[69] European Data Protection Board, "Guidelines 1/2020 on processing personal data in the context of connected vehicles and mobility related applications," 2020.

[70] Ada Lovelace Institute, "The data will see you now: Datafication and the boundaries of health," 2020, accessed August 2024. [Online]. Available: ***https://www.adalovelaceinstitute.org/wp-content/uploads/2020/11/The-data-will-see-you-now-Ada-Lovelace-Institute-Oct-2020.pdf***

[71] R. Li, Z. Zhao, X. Zhou, G. Ding, Y. Chen, Z. Wang, and H. Zhang, "Intelligent 5g: When cellular networks meet artificial intelligence," IEEE Wireless communications, vol. 24, no. 5, pp. 175–183, 2017.

[72] 3GPP, "System architecture for the 5g system (5gs)," 2023, accessed October 2023. [Online]. Available: ***https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3370***

[73] ——, "5g system; network exposure function southbound services; stage 3," 2023, accessed October 2023. [Online]. Available: ***https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3681***

[74] L. Morand, "Openapis for the service-based architecture," 3GPP, accessed October 2023. [Online]. Available: ***https://www.3gpp. org/technologies/openapis-for-the-service-based-architecture***

[75] E. Zeydan, J. Mangues-Bafalluy, J. Baranda, M. Requena, and Y. Turk, "Service based virtual ran architecture for next generation cellular systems," IEEE Access, vol. 10, pp. 9455–9470, 2022.

[76] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi-access edge computing: A survey of the emerging 5g network edge cloud architecture and orchestration," IEEE Communications Surveys & Tutorials, vol. 19, no. 3, pp. 1657– 1681, 2017.

[77] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions," IEEE Communications Surveys & Tutorials, vol. 20, no. 3, pp. 2429–2453, 2018.

[78] M. A. Garcia-Martin, M. Gramaglia, and P. Serrano, "Network automation and data analytics in 3gpp 5g systems," IEEE Network, 2023.

[79] A. Ghosh, A. Maeder, M. Baker, and D. Chandramouli, "5g evolution: A view on 5g cellular technology beyond 3gpp release 15," IEEE access, vol. 7, pp. 127 639–127 651, 2019.

[80] M. Yao, M. Sohul, V. Marojevic, and J. H. Reed, "Artificial intelligence defined 5g radio access networks," IEEE Communications Magazine, vol. 57, no. 3, pp. 14–20, 2019.

[81] Y. L. Lee, D. Qin, L.-C. Wang, and G. H. Sim, "6g massive radio access networks: Key applications, requirements and challenges," IEEE Open Journal of Vehicular Technology, vol. 2, pp. 54–66, 2020.

[82] B. Brik, H. Chergui, L. Zanzi, F. Devoti, A. Ksentini, M. S. Siddiqui, X. Costa-Pérez, and C. Verikoukis, "A survey on explainable ai for 6g o-ran: Architecture, use cases, challenges and research directions," arXiv preprint arXiv:2307.00319, 2023.

[83] A. Mudvari, N. Makris, and L. Tassiulas, "Ml-driven scaling of 5g cloud-native rans," in 2021 IEEE Global Communications Conference (GLOBECOM). IEEE, 2021, pp. 1–6.

[84] S. K. Singh, R. Singh, and B. Kumbhani, "The evolution of radio access network towards open-ran: Challenges and opportunities," in 2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW). IEEE, 2020, pp. 1–6.

[85] O-RAN Software Community, "D release (jul 2021)," 2023, accessed October 2023. [Online]. Available: ***https://wiki.o-ran-sc.org/pages/viewpage.action?pageId=20878658***

[86] M. Polese, L. Bonati, S. D'oro, S. Basagni, and T. Melodia, "Understanding o-ran: Architecture, interfaces, algorithms, security, and research challenges," IEEE Communications Surveys & Tutorials, 2023.

[87] O-RAN Working Group 2, "O-ran ai/ml workflow description and requirements 1.03," O-RAN, Tech. Rep. O-RAN.WG2.AIML-v01.03 Technical Specification, August 2021.

[88] L. Lovén, H. F. Shahid, L. Nguyen, E. Harjula, O. Silvén, S. Pirttikangas, and M. B. Lo´pez, "Semantic slicing across the distributed intelligent 6g wireless networks," in 2023 20th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON). IEEE, 2023, pp. 79–84.

[89] The O-RAN Software Community (SC) Documentation, "Ric message router documentation," accessed October 2023. [Online]. Available: ***https://docs.o-ran-sc.org/projects/o-ran-sc-ric-plt-lib-rmr***

[90] S. Tarkoma, Publish/Subscribe Systems: Design and Principles. United States: Wiley, Aug. 2012.

[91] J. Mcnamara, D. Camps-Mur, M. Goodarzi, H. Frank, L. Chinchilla-Romero, F. Cañellas, A. Fernandéz-Fernandéz, and S. Yan, "Nlp powered intent based network management for private 5g networks," IEEE Access, 2023.

[92] A. Leivadeas and M. Falkner, "A survey on intent based networking," IEEE Communications Surveys & Tutorials, 2022.

[93] S. Guo, Y. Wang, S. Li, and N. Saeed, "Semantic importance-aware communications using pre-trained language models," IEEE Communications Letters, 2023.

[94] J. Park, S.-W. Ko, J. Choi, S.-L. Kim, and M. Bennis, "Towards semantic communication protocols for 6g: From protocol learning to language-oriented approaches," arXiv preprint arXiv:2310.09506, 2023.

[95] H. Nam, J. Park, J. Choi, M. Bennis, and S.-L. Kim, "Language-oriented communication with semantic coding and knowledge distillation for text-to-image generation," arXiv preprint arXiv:2309.11127, 2023.

[96] R. Calo, "Artificial intelligence policy: A primer and roadmap," UCDL Rev., vol. 51, p. 399, 2017.

[97] T. Zhang and S. Mao, "An introduction to the federated learning standard," GetMobile: Mobile Computing and Communications, vol. 25, pp. 18–22, 1 2022. [Online]. Available: ***https://dl.acm.org/doi/10.1145/3511285.3511291***

[98] D. Kreuzberger, N. Kuhl, and S. Hirschl, "Machine learning operations (mlops): Overview, definition, and architecture," IEEE Access, vol. 11, pp. 31 866–31 879, 2023.

[99] Ambrosys, "Machine learning terminology in the light of physics," 2023. [Online]. Available: ***https://medium.com/@ambrosys/machine-learning-terminology-in-the-light-of-physics-854967d653bb***

[100] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 4 2018. [Online]. Available: ***https://ojs.aaai.org/index.php/AAAI/article/view/11694***

[101] A. Holzinger, E. Weippl, A. M. Tjoa, and P. Kieseberg, "Digital transformation for sustainable development goals (sdgs) - a security, safety and privacy perspective on ai," vol. 12844 LNCS. Springer Science and Business Media Deutschland GmbH, 2021, pp. 1–20.

[102] T. Mohammed, C. Joe-Wong, R. Babbar, and M. D. Francesco, "Distributed inference acceleration with adaptive dnn partitioning and offloading," in IEEE INFOCOM 2020 - IEEE Conference on Computer Communications, 2020, pp. 854–863.

[103] A. Parthasarathy and B. Krishnamachari, "Defer: Distributed edge inference for deep neural networks," 2022 14th International Conference on COMmunication Systems & NETworkS (COMSNETS), pp. 749–753, 2022. [Online]. Available: ***https://api.semanticscholar.org/CorpusID:245940198***

[104] N. Li, A. Iosifidis, and Q. Zhang, "Collaborative edge computing for distributed cnn inference acceleration using receptive field-based segmentation," Computer Networks, vol. 214, p. 109150, 2022. [Online]. Available: ***https://www.sciencedirect.com/science/article/pii/S1389128622002638***

[105] ——, "Distributed deep learning inference acceleration using seamless collaboration in edge computing," in ICC 2022 - IEEE International Conference on Communications, 2022, pp. 3667–3672.

[106] C. Hu, W. Bao, D. Wang, and F. Liu, "Dynamic adaptive dnn surgery for inference acceleration on the edge," in IEEE INFO-COM 2019 - IEEE Conference on Computer Communications, 2019, pp. 1423–1431.

[107] A. Bakhtiarnia, N. Milošević, Q. Zhang, D. Bajović, and A. Iosifidis, "Dynamic split computing for efficient deep edge intelligence," in ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.

[108] B. Zhang, T. Xiang, H. Zhang, T. Li, S. Zhu, and J. Gu, "Dynamic dnn decomposition for lossless synergistic inference," in 2021 IEEE 41st International Conference on Distributed Computing Systems Workshops (ICDCSW). IEEE, 2021, pp. 13–20.

[109] A. Bakhtiarnia, Q. Zhang, and A. Iosifidis, "Single-layer vision transformers for more accurate early exits with less overhead," Neural Networks, vol. 153, pp. 461–473, 2022. [Online]. Available: ***https://www.sciencedirect.com/science/article/pii/S0893608022002532***

[110] A. Borzunov, M. Ryabinin, A. Chumachenko, D. Baranchuk, T. Dettmers, Y. Belkada, P. Samygin, and C. A. Raffel, "Distributed inference and fine-tuning of large language models over the internet," Advances in Neural Information Processing Systems, vol. 36, 2024.

[111] R. Y. Aminabadi, S. Rajbhandari, A. A. Awan, C. Li, D. Li, E. Zheng, O. Ruwase, S. Smith, M. Zhang, J. Rasley et al., "Deepspeed-inference: enabling efficient inference of transformer models at unprecedented scale," in SC22: International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, 2022, pp. 1–15.

[112] G.-I. Yu, J. S. Jeong, G.-W. Kim, S. Kim, and B.-G. Chun, "Orca: A distributed serving system for Transformer-Based generative models," in 16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22). Carlsbad, CA: USENIX Association, Jul. 2022, pp. 521–538. [Online]. Available: ***https://www.usenix.org/conference/osdi22/presentation/yu***

[113] B. Wu, Y. Zhong, Z. Zhang, G. Huang, X. Liu, and X. Jin, "Fast distributed inference serving for large language models," arXiv preprint arXiv:2305.05920, 2023.

[114] G. Al-Atat, A. Fresa, A. P. Behera, V. N. Moothedath, J. Gross, and J. P. Champati, "The case for hierarchical deep learning inference at the network edge," in Proceedings of the 1st International Workshop on Networked AI Systems, 2023, pp. 1–6.

[115] J. C. V. Moothedath and J. Gross, "Getting the best out of both worlds: Algorithms for hierarchical inference at the edge," IEEE Transactions on Transactions on Machine Learning in Communications and Networking, 2024.

[116] B. Yang, L. He, N. Ling, Z. Yan, G. Xing, X. Shuai, X. Ren, and X. Jiang, "Edgefm: Leveraging foundation model for open-set learning on the edge," arXiv preprint arXiv:2311.10986, 2023.

[117] D. Xu, W. Yin, X. Jin, Y. Zhang, S. Wei, M. Xu, and X. Liu, "Llm-cad: Fast and scalable on-device large language model inference," arXiv preprint arXiv:2309.04255, 2023.

[118] Z. Sun, A. T. Suresh, J. H. Ro, A. Beirami, H. Jain, and F. Yu, "Spectr: Fast speculative decoding via optimal transport," Advances in Neural Information Processing Systems, vol. 36, 2024.

[119] Y. Wang, K. Chen, H. Tan, and K. Guo, "Tabi: An efficient multilevel inference system for large language models," in Proceedings of the Eighteenth European Conference on Computer Systems, 2023, pp. 233–248.

[120] L. Chen, M. Zaharia, and J. Zou, "Frugalgpt: How to use large language models while reducing cost and improving performance," arXiv preprint arXiv:2305.05176, 2023.

[121] Z. Yao, R. Yazdani Aminabadi, M. Zhang, X. Wu, C. Li, and Y. He, "Zeroquant: Efficient and affordable post-training quantization for large-scale transformers," Advances in Neural Information Processing Systems, vol. 35, pp. 27 168–27 183, 2022.

[122] E. Frantar and D. Alistarh, "Sparsegpt: Massive language models can be accurately pruned in one-shot," in International Conference on Machine Learning. PMLR, 2023, pp. 10 323–10 337.

[123] H. Jiang, Q. Wu, C.-Y. Lin, Y. Yang, and L. Qiu, "LLMLingua: Compressing prompts for accelerated inference of large language models," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 13 358–13 376. [Online]. Available: ***https://aclanthology.org/2023.emnlp-main.825***

[124] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1–38, 2023.

[125] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, "A survey of the state of explainable ai for natural language processing," arXiv preprint arXiv:2010.00711, 2020.

[126] P. Hager, F. Jungmann, K. Bhagat, I. Hubrecht, M. Knauer, J. Vielhauer, R. Holland, R. Braren, M. Makowski, G. Kaisis et al., "Evaluating and mitigating limitations of large language models in clinical decision making," medRxiv, pp. 2024–01, 2024.

[127] Y. Cao, Y. Kang, and L. Sun, "Instruction mining: High-quality instruction data selection for large language models," arXiv preprint arXiv:2307.06290, 2023.

[128] A. Razdaibiedina, Y. Mao, R. Hou, M. Khabsa, M. Lewis, and A. Almahairi, "Progressive prompts: Continual learning for language models," arXiv preprint arXiv:2301.12314, 2023.

[129] S. Ott, K. Hebenstreit, V. Liévin, C. E. Hother, M. Moradi, M. Mayrhauser, R. Praas, O. Winther, and M. Samwald, "Thoughtsource: A central hub for large language model reasoning data," Scientific Data, vol. 10, no. 1, p. 528, 08 2023. [Online]. Available: ***https://doi.org/10.1038/s41597-023-02433-3***

[130] Y. Wang and Y. Zhao, "Tram: Benchmarking temporal reasoning for large language models," 2023.

[131] A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, and Y. Fang, "Evaluating very long-term conversational memory of llm agents," arXiv preprint arXiv:2402.17753, 2024.

[132] B. Jin, G. Liu, C. Han, M. Jiang, H. Ji, and J. Han, "Large Language Models on Graphs: A Comprehensive Survey," arXiv preprint arXiv:2312.02783, 2023.

[133] S. Pan, Y. Zheng, and Y. Liu, "Integrating graphs with large language models: Methods and prospects," IEEE Intelligent Systems, 2024.

[134] S. Xiong, A. Payani, R. Kompella, and F. Fekri, "Large language models can learn temporal reasoning," arXiv preprint arXiv:2401.06853, 2024.

[135] M. Galkin, X. Yuan, H. Mostafa, J. Tang, and Z. Zhu, "Towards foundation models for knowledge graph reasoning," in International Conference on Learning Representations, 2023.

[136] H. Xie, D. Zheng, J. Ma, H. Zhang, V. N. Ioannidis, X. Song, Q. Ping, S. Wang, C. Yang, Y. Xu, B. Zeng, and T. Chilimbi, "Graph-aware language model pre-training on a large graph corpus can help multiple graph applications," in ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023.

[137] Y. Tian, H. Song, Z. Wang, H. Wang, Z. Hu, F. Wang, N. V. Chawla, and P. Xu, "Graph neural prompting with large language models," in AAAI Conference on Artificial Intelligence, 2024.

[138] J. Baek, A. F. Aji, and A. Saffari, "Knowledge-augmented language model prompting for zero-shot knowledge graph question answering," arXiv preprint arXiv:2306.04136, 2023.

[139] R. Ye, C. Zhang, R. Wang, S. Xu, and Y. Zhang, "Language is all a graph needs," in Conference of the European Chapter of the Association for Computational Linguistics, 2024.

[140] M. J. Buehler, "Generative retrieval-augmented ontologic graph and multiagent strategies for interpretive large language model-based materials design," ACS Engineering Au, 2023.

[141] J. Jiang, K. Zhou, Z. Dong, K. Ye, W. X. Zhao, and J.-R. Wen, "Structgpt: A general framework for large language model to reason over structured data," arXiv preprint arXiv:2305.09645, 2023.

[142] M. Kang, S. Lee, J. Baek, K. Kawaguchi, and S. J. Hwang, "Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks," Advances in Neural Information Processing Systems, vol. 36, 2024.

[143] T. Baldazzi, L. Bellomarini, S. Ceri, A. Colombo, A. Gentili, and E. Sallinger, "Fine-tuning large enterprise language models via ontological reasoning," arXiv preprint arXiv:2306.10723, 2023.

[144] J. Sun, C. Xu, L. Tang, S. Wang, C. Lin, Y. Gong, H.-Y. Shum, and J. Guo, "Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph," arXiv preprint arXiv:2307.07697, 2023.

[145] Y. Wen, Z. Wang, and J. Sun, "Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models," arXiv preprint arXiv:2308.09729, 2023.

[146] Y. Tian, H. Song, Z. Wang, H. Wang, Z. Hu, F. Wang, N. V. Chawla, and P. Xu, "Graph neural prompting with large language models," 2023.

[147] L. Luo, Y.-F. Li, G. Haffari, and S. Pan, "Reasoning on graphs: Faithful and interpretable large language model reasoning," arXiv preprint arXiv:2310.01061, 2023.

[148] X. Guan, Y. Liu, H. Lin, Y. Lu, B. He, X. Han, and L. Sun, "Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting," arXiv preprint arXiv:2311.13314, 2023.

[149] W. Zhao, Y. Liu, T. Niu, Y. Wan, P. S. Yu, S. Joty, Y. Zhou, and S. Yavuz, "Divknowqa: Assessing the reasoning ability of llms via open-domain question answering over knowledge base and text," arXiv preprint arXiv:2310.20170, 2023.

[150] C. Wang, Y. Xu, Z. Peng, C. Zhang, B. Chen, X. Wang, L. Feng, and B. An, "keqing: knowledge-based question answering is a nature chain-of-thought mentor of llm," arXiv preprint arXiv:2401.00426, 2023.

[151] M. Mountantonakis and Y. Tzitzikas, "Using multiple rdf knowledge graphs for enriching chatgpt responses," arXiv preprint arXiv:2304.05774, 2023.

[152] Y. Gao, R. Li, J. Caskey, D. Dligach, T. Miller, M. M. Churpek, and M. Afshar, "Leveraging a medical knowledge graph into large language models for diagnosis prediction," arXiv preprint arXiv:2308.14321, 2023.

[153] L. Luo, T.-T. Vu, D. Phung, and G. Haffari, "Systematic assessment of factual knowledge in large language models," arXiv preprint arXiv:2310.11638, 2023.

[154] W. Wang, J. Shi, Z. Tu, Y. Yuan, J.-t. Huang, W. Jiao, and M. R. Lyu, "The earth is flat? unveiling factual errors in large language models," arXiv preprint arXiv:2401.00761, 2024.

[155] V. Pallagani, K. Roy, B. Muppasani, F. Fabiano, A. Loreggia, K. Murugesan, B. Srivastava, F. Rossi, L. Horesh, and A. Sheth, "On the prospects of incorporating large language models (llms) in automated planning and scheduling (aps)," arXiv preprint arXiv:2401.02500, 2024.

[156] Z. Liu, Y. Zhang, P. Li, Y. Liu, and D. Yang, "Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization," arXiv preprint arXiv:2310.02170, 2023.

[157] B. Ni and M. J. Buehler, "Mechagents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge," Extreme Mechanics Letters, p. 102131, 2024.

[158] S. Hong, X. Zheng, J. Chen, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, and C. Wu, "Metagpt: Meta programming for multi-agent collaborative framework," 2023.

[159] G. Chen, S. Dong, Y. Shu, G. Zhang, J. Sesay, B. F. Karlsson, J. Fu, and Y. Shi, "Autoagents: A framework for automatic agent generation," arXiv preprint arXiv:2309.17288, 2023.

[160] J. C.-Y. Chen, S. Saha, E. Stengel-Eskin, and M. Bansal, "Magdi: Structured distillation of multi-agent interaction graphs improves reasoning in smaller language models," arXiv preprint arXiv:2402.01620, 2024.

[161] Z. Tang, R. Wang, W. Chen, K. Wang, Y. Liu, T. Chen, and L. Lin, "Towards causalgpt: A multi-agent approach for faithful knowledge reasoning via promoting causal consistency in llms," arXiv preprint arXiv:2308.11914, 2023.

[162] S. Agashe, Y. Fan, and X. E. Wang, "Evaluating multi-agent coordination abilities in large language models," arXiv preprint arXiv:2310.03903, 2023.

[163] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, "Large language model based multi-agents: A survey of progress and challenges," arXiv preprint arXiv:2402.01680, 2024.

[164] A. Saleh, R. Morabito, S. Dustdar, S. Tarkoma, S. Pirttikangas, and L. Lovén, "Towards message brokers for generative ai: Survey, challenges, and opportunities," 2024.

[165] Y. Cheng, C. Zhang, Z. Zhang, X. Meng, S. Hong, W. Li, Z. Wang, Z. Wang, F. Yin, J. Zhao et al., "Exploring large language model based intelligent agents: Definitions, methods, and prospects," arXiv preprint arXiv:2401.03428, 2024.

[166] Y. Shen, Y. Shi, J. Zhang, and K. B. Letaief, "Graph neural networks for scalable radio resource management: Architecture design and theoretical analysis," IEEE Journal on Selected Areas in Communications, 2021.

[167] T. Chen, X. Zhang, M. You, G. Zheng, and S. Lambotharan, "A GNN-based supervised learning framework for resource allocation in wireless IoT networks," IEEE Internet of Things Journal, 2022.

[168] Z. He, L. Wang, H. Ye, G. Y. Li, and B.-H. F. Juang, "Resource allocation based on graph neural networks in vehicular communications," in IEEE Global Communications Conference, 2020.

[169] S. K. Mani, Y. Zhou, K. Hsieh, S. Segarra, R. Chandra, and S. Kandula, "Enhancing network management using code generated by large language models," arXiv preprint arXiv:2308.06261, 2023.

[170] Z. Wang, Y. Zhou, Y. Zou, Q. An, Y. Shi, and M. Bennis, "A graph neural network learning approach to optimize ris-assisted federated learning," IEEE Transactions on Wireless Communications, 2023.

[171] B. Fatemi, J. Halcrow, and B. Perozzi, "Talk like a graph: Encoding graphs for large language models," in International Conference on Learning Representations, 2024.

[172] Z. Zhang, X. Wang, Z. Zhang, H. Li, Y. Qin, S. Wu, and W. Zhu, "LLM4DyG: Can Large Language Models Solve Problems on Dynamic Graphs?" arXiv preprint arXiv:2310.17110, 2023.

[173] Y. Wang, R. Jiao, C. Lang, S. S. Zhan, C. Huang, Z. Wang, Z. Yang, and Q. Zhu, "Empowering Autonomous Driving with Large Language Models: A Safety Perspective," arXiv preprint arXiv:2312.00812, 2023.

[174] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan, "3D-LLM: Injecting the 3D World into Large Language Models," Conference on Neural Information Processing Systems, 2023.

[175] L. Wei, Z. He, H. Zhao, and Q. Yao, "Unleashing the power of graph learning through LLM-based autonomous agents," arXiv preprint arXiv:2309.04565, 2023.

[176] L. Wen, D. Fu, X. Li, X. Cai, M. Tao, P. Cai, M. Dou, B. Shi, L. He, and Y. Qiao, "Dilu: A knowledge-driven approach to autonomous driving with large language models," in International Conference on Learning Representations, 2024.

[177] C. Sun, J. Han, W. Deng, X. Wang, Z. Qin, and S. Gould, "3D-GPT: Procedural 3d modeling with large language models," arXiv preprint arXiv:2310.12945, 2023.

[178] K. Tonchev, A. Ivanov, N. Neshov, A. Manolova, and V. Poulkov, "Learning graph convolutional neural networks to predict radio environment maps," in International Symposium on Wireless Personal Multimedia Communications, 2022.

[179] K. Tan, D. Bremner, J. Le Kernec, Y. Sambo, L. Zhang, and M. A. Imran, "Graph neural network-based cell switching for energy optimization in ultra-dense heterogeneous networks," Scientific Reports, 2022.

[180] M. Nerini and B. Clerckx, "Overhead-free blockage detection and precoding through physics-based graph neural networks: LIDAR data meets ray tracing," IEEE Wireless Communications Letters, 2023.

[181] X. Zhou, M. Liu, B. L. Zagar, E. Yurtsever, and A. C. Knoll, "Vision language models in autonomous driving and intelligent transportation systems," arXiv preprint arXiv:2310.14414, 2023.

[182] S. Wasserkrug, L. Boussioux, D. d. Hertog, F. Mirzazadeh, I. Birbil, J. Kurtz, and D. Maragno, "From large language models and optimization to decision optimization copilot: A research manifesto," arXiv preprint arXiv:2402.16269, 2024.

[183] H. Amrouch, G. Zervakis, S. Salamin, H. Kattan, I. Anagnostopoulos, and J. Henkel, "Npu thermal management," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 39, no. 11, pp. 3842–3855, 2020.

[184] C. Kachris, "A survey on hardware accelerators for large language models," arXiv preprint arXiv:2401.09890, 2024.

[185] X. Yang, B. Yan, H. Li, and Y. Chen, "Retransformer: Reram-based processing-in-memory architecture for transformer acceleration," in Proceedings of the 39th International Conference on Computer-Aided Design, 2020, pp. 1–9.

[186] S.-H. Kang, S. Lee, and K. Sohn, "The era of generative artificial intelligence: In-memory computing perspective," in 2023 International Electron Devices Meeting (IEDM). IEEE, 2023, pp. 1–4.

[187] N. Yang, T. Ge, L. Wang, B. Jiao, D. Jiang, L. Yang, R. Majumder, and F. Wei, "Inference with reference: Lossless acceleration of large language models," arXiv preprint arXiv:2304.04487, 2023.

[188] H. Wang, Z. Zhang, and S. Han, "Spatten: Efficient sparse attention architecture with cascade token and head pruning," in 2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA). IEEE, 2021, pp. 97–110.

[189] L. Lu, Y. Jin, H. Bi, Z. Luo, P. Li, T. Wang, and Y. Liang, "Sanger: A co-design framework for enabling sparse attention using reconfigurable architecture," in MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture, 2021, pp. 977–991.

[190] M. Isaev, N. McDonald, and R. Vuduc, "Scaling infrastructure to support multi-trillion parameter llm training," in Architecture and System Support for Transformer Models (ASSYST@ ISCA 2023), 2023.

[191] V. Young, S. Kariyappa, and M. K. Qureshi, "Enabling transparent memory-compression for commodity memory systems," in 2019 IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE, 2019, pp. 570–581.

[192] T. J. Ham, Y. Lee, S. H. Seo, S. Kim, H. Choi, S. J. Jung, and J. W. Lee, "Elsa: Hardware-software co-design for efficient, lightweight self-attention mechanism in neural networks," in 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA). IEEE, 2021, pp. 692–705.

[193] R. Mao, B. Wen, A. Kazemi, Y. Zhao, A. F. Laguna, R. Lin, N. Wong, M. Niemier, X. S. Hu, X. Sheng et al., "Experimentally validated memristive memory augmented neural network with efficient hashing and similarity search," Nature communications, vol. 13, no. 1, p. 6284, 2022.

[194] G. F. Oliveira, A. Olgun, A. G. Yağlıkçı, F. Bostancı, J. Gómez-Luna, S. Ghose, and O. Mutlu, "Mimdram: An end-to-end processing-using-dram system for high-throughput, energy-efficient and programmer-transparent multiple-instruction multiple-data processing," arXiv preprint arXiv:2402.19080, 2024.

[195] F. Devaux, "The true processing in memory accelerator," in 2019 IEEE Hot Chips 31 Symposium (HCS). IEEE Computer Society, 2019, pp. 1–24.

[196] L. Su and S. Naffziger, "1.1 innovation for the next decade of compute efficiency," in 2023 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2023, pp. 8–12.

[197] Q. Zeng, J. Liu, M. Jiang, J. Lan, Y. Gong, Z. Wang, Y. Li, C. Li, J. Ignowski, and K. Huang, "Realizing in-memory baseband processing for ultra-fast and energy-efficient 6g," IEEE Internet of Things Journal, 2023.

[198] N. Hajinazar, G. F. Oliveira, S. Gregorio, J. D. Ferreira, N. M. Ghiasi, M. Patel, M. Alser, S. Ghose, J. Gómez-Luna, and O. Mutlu, "Simdram: A framework for bit-serial simd processing using dram," in Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, 2021, pp. 329–345.

[199] S. Ma, H. Wang, L. Ma, L. Wang, W. Wang, S. Huang, L. Dong, R. Wang, J. Xue, and F. Wei, "The era of 1-bit llms: All large language models are in 1.58 bits," arXiv preprint arXiv:2402.17764, 2024.

[200] R. G. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge, "Near-threshold computing: Reclaiming moore's law through energy efficient integrated circuits," Proceedings of the IEEE, vol. 98, no. 2, pp. 253–266, 2010.

[201] M. Safarpour, R. Inanlou, and O. Silvén, "Algorithm level error detection in low voltage systolic array," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 69, no. 2, pp. 569–573, 2021.

[202] M. Safarpour, T. Z. Deng, J. Massingham, L. Xun, M. Sabokrou, and O. Silvén, "Low-voltage energy efficient neural inference by leveraging fault detection techniques," in 2021 IEEE Nordic Circuits and Systems Conference (NorCAS). IEEE, 2021, pp. 1–5.

[203] M. Safarpour, L. Xun, G. V. Merrett, and O. Silvén, "A high-level approach for energy efficiency improvement of fpgas by voltage trimming," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 41, no. 10, pp. 3548–3552, 2021.

[204] K. Zhao, S. Di, S. Li, X. Liang, Y. Zhai, J. Chen, K. Ouyang, F. Cappello, and Z. Chen, "Ft-cnn: Algorithm-based fault tolerance for convolutional neural networks," IEEE Transactions on Parallel and Distributed Systems, vol. 32, no. 7, pp. 1677–1689, 2020.

[205] S. Tarkoma, R. Morabito, and J. Sauvola, "Ai-native interconnect framework for integration of large language model technologies in 6g systems," arXiv preprint arXiv:2311.05842, 2023.

[206] IEEE Communications Society, "Large generative ai models in telecom emerging technologies initiative," 2024, accessed August 2024. [Online]. Available: ***https://genainet.committees.comsoc.org/home-2/***

[207] Z. Tao, W. Xu, Y. Huang, X. Wang, and X. You, "Wireless network digital twin for 6g: Generative ai as a key enabler," IEEE Wireless Communications, vol. 31, no. 4, pp. 24–31, August 2024.

[208] L. Bariah and M. Debbah, "The interplay of ai and digital twin: Bridging the gap between data-driven and model-driven approaches," IEEE Wireless Communications, 2024.

[209] J. Wang, H. Du, D. Niyato, Z. Xiong, J. Kang, S. Mao, and X. S. Shen, "Guiding ai-generated digital content with wireless perception," IEEE Wireless Communications, 2024.

[210] N. Sehad, L. Bariah, W. Hamidouche, H. Hellaoui, R. Jäntti, and M. Debbah, "Generative ai for immersive communication: The next frontier in internet-of-senses through 6g," arXiv preprint arXiv:2404.01713, 2024.

[211] R. Zhang, H. Du, Y. Liu, D. Niyato, J. Kang, Z. Xiong, A. Jamalipour, and D. I. Kim, "Interactive generative ai agents for satellite networks through a mixture of experts transmission," arXiv preprint arXiv:2404.09134, 2024.

[212] J. Chen, C. Yi, H. Du, D. Niyato, J. Kang, J. Cai, and X. Shen, "A revolution of personalized healthcare: Enabling human digital twin with mobile aigc," IEEE Network, 2024.

[213] J. Chen, Y. Shi, C. Yi, H. Du, J. Kang, and D. Niyato, "Generative ai-driven human digital twin in iot-healthcare: A comprehensive survey," arXiv preprint arXiv:2401.13699, 2024.

[214] L. Nguyen, P. Susarla, A. Mukherjee, M. Lage, C. Alvarez Casado, X. Wu, O. Silvén, D. B. Jayagopi, and M. Bordallo Lo´pez, "Non-contact multimodal indoor human monitoring systems: A survey," Information Fusion, vol. 110, p. 102457, 2024.

[215] M. Xu, D. Niyato, J. Kang, Z. Xiong, S. Guo, Y. Fang, and D. I. Kim, "Generative ai-enabled mobile tactical multimedia networks: Distribution, generation, and perception," arXiv preprint arXiv:2401.06386, 2024.

[216] M. Xu, D. Niyato, H. Zhang, J. Kang, Z. Xiong, S. Mao, and Z. Han, "Cached model-as-a-resource: Provisioning large language model agents for edge intelligence in space-air-ground integrated networks," arXiv preprint arXiv:2403.05826, 2024.

[217] Y. Du, S. C. Liew, K. Chen, and Y. Shao, "The power of large language models for wireless communication system development: A case study on fpga platforms," arXiv preprint arXiv:2307.07319, 2023.

[218] X. Lin, L. Kundu, C. Dick, M. A. C. Galdon, J. Vamaraju, S. Dutta, and V. Raman, "A primer on generative ai for telecom: From theory to practice," arXiv preprint arXiv:2408.09031, 2024.

[219] S. Nayak and R. Patgiri, "6g communication: Envisioning the key issues and challenges," EAI Endorsed Transactions on Internet of Things, vol. 6, no. 24, p. 166959, Feb. 2021.

[220] W. Saad, M. Bennis, and M. Chen, "A vision of 6g wireless systems: Applications, trends, technologies, and open research problems," IEEE Network, vol. 34, pp. 134–142, 2019.

[221] P. Yang, Y. Xiao, M. Xiao, and S. Li, "6g wireless communications: Vision and potential techniques," IEEE Netw., vol. 33, no. 4, pp. 70–75, 2019.

[222] L. Chang, Z. Zhang, P. Li, S. Xi, W. Guo, Y. Shen, Z. Xiong, J. Kang, D. Niyato, X. Qiao, and Y. Wu, "6g-enabled edge AI for metaverse: Challenges, methods, and future research directions," CoRR, vol. abs/2204.06192, 2022.

[223] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States, P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012, pp. 1106–1114.

[224] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. Ranzato, A. W. Senior, P. A. Tucker, K. Yang, and A. Y. Ng, "Large scale distributed deep networks," in Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States, P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012, pp. 1232–1240.

[225] F. Brakel, U. Odyurt, and A.-L. Varbanescu, "Model parallelism on distributed infrastructure: A literature review from theory to llm case-studies," 2024.

[226] M. W. Akhtar, S. A. Hassan, R. Ghaffar, H. Jung, S. Garg, and M. S. Hossain, "The shift to 6g communications: Vision and requirements," Human-centric Computing and Information Sciences, vol. 10, pp. 1–27, 2020.

[227] B. Mao, F. Tang, Y. Kawamoto, and N. Kato, "Ai models for green communications towards 6g," IEEE Communications Surveys & Tutorials, vol. 24, no. 1, pp. 210–247, 2021.

[228] Y. Song, Z. Mi, H. Xie, and H. Chen, "Powerinfer: Fast large language model serving with a consumer-grade gpu," arXiv preprint arXiv:2312.12456, 2023.

[229] S. Liu, J. Yu, X. Deng, and S. Wan, "Fedcpf: An efficient-communication federated learning approach for vehicular edge computing in 6g communication networks," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 2, pp. 1616–1629, 2021.

[230] B. Zhang, Z. Liu, C. Cherry, and O. Firat, "When scaling meets llm finetuning: The effect of data, model and finetuning method," 2024.

[231] J. Yang, X. Shen, J. Xing, X. Tian, H. Li, B. Deng, J. Huang, and X.-s. Hua, "Quantization networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 7308–7316.

[232] C. Thapa, P. C. M. Arachchige, S. Camtepe, and L. Sun, "Splitfed: When federated learning meets split learning," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 8, 2022, pp. 8485–8493.

[233] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," arXiv preprint arXiv:2106.09685, 2021.

[234] Z. Hu, L. Wang, Y. Lan, W. Xu, E.-P. Lim, L. Bing, X. Xu, S. Poria, and R. K.-W. Lee, "Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models," in The 2023 Conference on Empirical Methods in Natural Language Processing, 2023.

[235] A. Celik and A. M. Eltawil, "At the dawn of generative ai era: A tutorial-cum-survey on new frontiers in 6g wireless intelligence," IEEE Open Journal of the Communications Society, p. 1–1, 2024. [Online]. Available: ***http://dx.doi.org/10.1109/OJCOMS.2024.3362271***

[236] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," ArXiv, 2023.

[237] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," arXiv preprint arXiv:1907.02189, 2019.

[238] M. Morafah, S. Vahidian, W. Wang, and B. Lin, "Flis: Clustered federated learning via inference similarity for non-iid data distribution," IEEE Open Journal of the Computer Society, vol. 4, pp. 109–120, 2023.

[239] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A survey of data augmentation approaches for nlp," arXiv preprint arXiv:2105.03075, 2021.

[240] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," Engineering Applications of Artificial Intelligence, vol. 115, p. 105151, 2022.

[241] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," Knowledge-Based Systems, vol. 216, p. 106775, 2021.

[242] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," Future Generation Computer Systems, vol. 115, pp. 619–640, 2021.

[243] B. Knott, S. Venkataraman, A. Hannun, S. Sengupta, M. Ibrahim, and L. van der Maaten, "Crypten: Secure multi-party computation meets machine learning," Advances in Neural Information Processing Systems, vol. 34, pp. 4961–4973, 2021.

[244] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," IEEE transactions on information forensics and security, vol. 15, pp. 3454–3469, 2020.

[245] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du, "Explainability for large language models: A survey," ACM Trans. Intell. Syst. Technol., vol. 15, no. 2, feb 2024.

[246] A. Shahraki, M. Abbasi, M. J. Piran, M. Chen, and S. Cui, "A comprehensive survey on 6g networks: Applications, core services, enabling technologies, and future challenges," CoRR, vol. abs/2101.12475, 2021.

[247] L. Chang, Z. Zhang, P. Li, S. Xi, W. Guo, Y. Shen, Z. Xiong, J. Kang, D. Niyato, X. Qiao et al., "6g-enabled edge ai for metaverse: Challenges, methods, and future research directions," Journal of Communications and Information Networks, vol. 7, no. 2, pp. 107– 121, 2022.

[248] G. Dileep, "A survey on smart grid technologies and applications," Renewable energy, vol. 146, pp. 2589–2625, 2020.

[249] A. Deshpande, P. Pitale, and S. Sanap, "Industrial automation using internet of things (iot)," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 5, no. 2, pp. 266–269, 2016.

[250] K. Wang, Y. Lu, M. Santacroce, Y. Gong, C. Zhang, and Y. Shen, "Adapting llm agents through communication," arXiv preprint arXiv:2310.01444, 2023.

[251] G. P. Fettweis, "The tactile internet: Applications and challenges," IEEE vehicular technology magazine, vol. 9, no. 1, pp. 64–70, 2014.

[252] Z. Hou, C. She, Y. Li, D. Niyato, M. Dohler, and B. Vucetic, "Intelligent communications for tactile internet in 6g: Requirements, technologies, and challenges," IEEE Communications Magazine, vol. 59, no. 12, pp. 82–88, 2021.

[253] H. Duan, J. Li, S. Fan, Z. Lin, X. Wu, and W. Cai, "Metaverse for social good: A university campus prototype," in Proceedings of the 29th ACM international conference on multimedia, 2021, pp. 153–161.

[254] B. Yao, M. Jiang, D. Yang, and J. Hu, "Empowering llm-based machine translation with cultural awareness," arXiv preprint arXiv:2305.14328, 2023.

[255] A. Aijaz, "Toward human-in-the-loop mobile networks: A radio resource allocation perspective on haptic communications," IEEE Transactions on Wireless Communications, vol. 17, no. 7, pp. 4493– 4508, 2018.

[256] D. Van Den Berg, R. Glans, D. De Koning, F. A. Kuipers, J. Lugtenburg, K. Polachan, P. T. Venkata, C. Singh, B. Turkovic, and B. Van Wijk, "Challenges in haptic communications over the tactile internet," IEEE Access, vol. 5, pp. 23 502–23 518, 2017.

[257] Y. Lu and X. Zheng, "6g: A survey on technologies, scenarios, challenges, and the related issues," Journal of Industrial Information Integration, vol. 19, p. 100158, 2020.

[258] K. Antonakoglou, X. Xu, E. Steinbach, T. Mahmoodi, and M. Dohler, "Toward haptic communications over the 5g tactile internet," IEEE Communications Surveys & Tutorials, vol. 20, no. 4, pp. 3034–3059, 2018.

[259] A. Clemm, M. T. Vega, H. K. Ravuri, T. Wauters, and F. De Turck, "Toward truly immersive holographic-type communication: Challenges and solutions," IEEE Communications Magazine, vol. 58, no. 1, pp. 93–99, 2020.

[260] B. Jin, G. Liu, C. Han, M. Jiang, H. Ji, and J. Han, "Large language models on graphs: A comprehensive survey," arXiv preprint arXiv:2312.02783, 2023.

[261] U. Jayasankar, V. Thirumal, and D. Ponnurangam, "A survey on data compression techniques: From the perspective of data quality, coding schemes, data type and applications," Journal of King Saud University-Computer and Information Sciences, vol. 33, no. 2, pp. 119–140, 2021.

[262] B. Farahani, M. Barzegari, F. S. Aliee, and K. A. Shaik, "Towards collaborative intelligent iot ehealth: From device to fog, and cloud," Microprocessors and Microsystems, vol. 72, p. 102938, 2020.

[263] W. Tian, M. Fan, C. Zeng, Y. Liu, D. He, and Q. Zhang, "Telerobotic spinal surgery based on 5g network: the first 12 cases," Neurospine, vol. 17, no. 1, p. 114, 2020.

[264] Y. Li, H. Wang, H. Yerebakan, Y. Shinagawa, and Y. Luo, "Enhancing health data interoperability with large language models: A fhir study," arXiv preprint arXiv:2310.12989, 2023.

[265] C. de Alwis, P. Kumar, Q. Pham, K. Dev, A. Kalla, M. Liyanage, and W. Hwang, "Towards 6g: Key technological directions," ICT Express, vol. 9, no. 4, pp. 525–533, 2023.

[266] P. Zhang, Y. Xiao, Y. Li, X. Ge, G. Shi, and Y. Yang, "Towards net-zero carbon emissions in network AI for 6g and beyond," CoRR, vol. abs/2401.01007, 2024.

[267] G. Tech, Y. Chen, K. Müller, J.-R. Ohm, A. Vetro, and Y.-K. Wang, "Overview of the multiview and 3d extensions of high efficiency video coding," IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, no. 1, pp. 35–49, 2015.

[268] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan, "High-quality streamable free-viewpoint video," ACM Transactions on Graphics (ToG), vol. 34, no. 4, pp. 1–13, 2015.

[269] Z. Liu, G. Cheung, J. Chakareski, and Y. Ji, "Multiple description coding and recovery of free viewpoint video for wireless multipath streaming," IEEE Journal of Selected Topics in Signal Processing, vol. 9, no. 1, pp. 151–164, 2014.

[270] C. Fehn, "Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3d-tv," in Stereoscopic displays and virtual reality systems XI, vol. 5291. SPIE, 2004, pp. 93–104.

[271] A. Yaqoob, T. Bi, and G.-M. Muntean, "A survey on adaptive 360 video streaming: Solutions, challenges and opportunities," IEEE Communications Surveys & Tutorials, vol. 22, no. 4, pp. 2801–2838, 2020.

[272] Huawei-iLab, "Cloud vr network solution white paper," 2018. [Online]. Available: ***http://www.huawei.com***

[273] J. Karafin and B. Bevensee, "25-2: On the support of light field and holographic video display technology," in SID Symposium Digest of Technical Papers, vol. 49, no. 1. Wiley Online Library, 2018, pp. 318–321.

[274] T. Azmin, M. Ahmadinejad, and N. Shahriar, "Bandwidth prediction in 5g mobile networks using informer," in 2022 13th International Conference on Network of the Future (NoF). IEEE, 2022, pp. 1–9.
[
275] J. Li, L. Han, C. Zhang, Q. Li, and Z. Liu, "Spherical convolution empowered viewport prediction in 360 video multicast with limited fov feedback," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 19, no. 1, pp. 1–23, 2023.

[276] F.-Y. Chao, C. Ozcinar, and A. Smolic, "Transformer-based long-term viewport prediction in 360° video: Scanpath is all you need." in MMSP, 2021, pp. 1–6.

[277] Y. Cheng and F. Lu, "Gaze estimation using transformer," in 2022 26th International Conference on Pattern Recognition (ICPR). IEEE, 2022, pp. 3341–3347.

[278] J. Xiang, K. Tian, and J. Zhang, "Mimt: Masked image modeling transformer for video compression," in The Eleventh International Conference on Learning Representations, 2022.

[279] F. Mentzer, G. Toderici, D. Minnen, S.-J. Hwang, S. Caelles, M. Lucic, and E. Agustsson, "Vct: A video compression transformer," arXiv preprint arXiv:2206.07307, 2022.

[280] P. Zhou, L. Wang, Z. Liu, Y. Hao, P. Hui, S. Tarkoma, and J. Kangasharju, "A survey on generative ai and llm for video generation, understanding, and streaming."

[281] T. Nguyen, H. Nguyen, A. Ijaz, S. Sheikhi, A. V. Vasilakos, and P. Kostakos, "Large language models in 6g security: challenges and opportunities," 2024.

[282] (2023) OWASP Top 10 for Large Language Model Applications. [Online]. Available: https://owasp.org/ www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v1 1.pdf

[283] M. Gupta, C. Akiri, K. Aryal, E. Parker, and L. Praharaj, "From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy," IEEE Access, 2023.

[284] M. Sultana, A. Taylor, L. Li, and S. Majumdar, "Towards evaluation and understanding of large language models for cyber operation automation," in 2023 IEEE Conference on Communications and Network Security (CNS). IEEE, 2023, pp. 1–6.

[285] Y. Yao, J. Duan, K. Xu, Y. Cai, E. Sun, and Y. Zhang, "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly," 2023.

[286] Y. Wan, S. Zhang, H. Zhang, Y. Sui, G. Xu, D. Yao, H. Jin, and L. Sun, "You see what i want you to see: poisoning vulnerabilities in neural code search," in Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2022, pp. 1233–1245.

[287] R. Schuster, C. Song, E. Tromer, and V. Shmatikov, "You autocomplete me: Poisoning vulnerabilities in neural code completion," in 30th USENIX Security Symposium (USENIX Security 21), 2021, pp. 1559–1575.

[288] S. Shan, W. Ding, J. Passananti, H. Zheng, and B. Y. Zhao, "Prompt-specific poisoning attacks on text-to-image generative models," arXiv preprint arXiv:2310.13828, 2023.

[289] J. Li, Y. Yang, Z. Wu, V. Vydiswaran, and C. Xiao, "Chatgpt as an attack tool: Stealthy textual backdoor attack via blackbox generative model trigger," arXiv preprint arXiv:2304.14475, 2023.

[290] H. Yao, J. Lou, and Z. Qin, "Poisonprompt: Backdoor attack on prompt-based large language models," arXiv preprint arXiv:2310.12439, 2023.

[291] H. Yang, K. Xiang, M. Ge, H. Li, R. Lu, and S. Yu, "A comprehensive overview of backdoor attacks in large language models within communication networks," 2023.

[292] N. Kandpal, K. Pillutla, A. Oprea, P. Kairouz, C. A. Choquette-Choo, and Z. Xu, "User inference attacks on large language models," arXiv preprint arXiv:2310.09266, 2023.

[293] C. Song and A. Raghunathan, "Information leakage in embedding models," in Proceedings of the 2020 ACM SIGSAC conference on computer and communications security, 2020, pp. 377–390.

[294] R. Staab, M. Vero, M. Balunović, and M. Vechev, "Beyond memorization: Violating privacy via inference with large language models," arXiv preprint arXiv:2310.07298, 2023.

[295] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in 2017 IEEE symposium on security and privacy (SP). IEEE, 2017, pp. 3–18.

[296] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in 2022 IEEE Symposium on Security and Privacy (SP). IEEE, 2022, pp. 1897–1914.

[297] C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot, "Label-only membership inference attacks," in International conference on machine learning. PMLR, 2021, pp. 1964–1974.

[298] J. Duan, F. Kong, S. Wang, X. Shi, and K. Xu, "Are diffusion models vulnerable to membership inference attacks?" 2023.

[299] M. Juuti, S. Szyller, S. Marchal, and N. Asokan, "Prada: protecting against dnn model stealing attacks," in 2019 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2019, pp. 512–527.

[300] S. Kariyappa, A. Prakash, and M. K. Qureshi, "Maze: Data-free model stealing attack using zeroth-order gradient estimation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13 814–13 823.

[301] M. Balunović, D. Dimitrov, N. Jovanović, and M. Vechev, "Lamp: Extracting text from gradients with language model priors," Advances in Neural Information Processing Systems, vol. 35, pp. 7641–7654, 2022.

[302] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson et al., "Extracting training data from large language models," in 30th USENIX Security Symposium (USENIX Security 21), 2021, pp. 2633–2650.

[303] J.-B. Truong, P. Maini, R. J. Walls, and N. Papernot, "Data-free model extraction," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021, pp. 4771–4780.

[304] Z. Talat, A. Névéol, S. Biderman, M. Clinciu, M. Dey, S. Longpre, S. Luccioni, M. Masoud, M. Mitchell, D. Radev et al., "You reap what you sow: On the challenges of bias evaluation under multilingual settings," in Proceedings of BigScience Episode# 5– Workshop on Challenges & Perspectives in Creating Large Language Models, 2022, pp. 26–41.

[305] J. Su, T. Y. Zhuo, J. Mansurov, D. Wang, and P. Nakov, "Fake news detectors are biased against texts generated by large language models," arXiv preprint arXiv:2309.08674, 2023.

[306] A. Urman and M. Makhortykh, "The silence of the llms: Cross-lingual analysis of political bias and false information prevalence in chatgpt, google bard, and bing chat," 2023.

[307] A. Namer, P. Kulkarni, E. Jeansson, B. Maltzman, and H. Vagts, "Automatically detecting expensive prompts and configuring firewall rules to mitigate denial of service attacks on large language models," 2024.

[308] S. Abdelnabi, K. Greshake, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection," in Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, 2023, pp. 79–90.

[309] Y. Liu, G. Deng, Y. Li, K. Wang, T. Zhang, Y. Liu, H. Wang, Y. Zheng, and Y. Liu, "Prompt injection attack against llmintegrated applications," arXiv preprint arXiv:2306.05499, 2023.

[310] Y. Liu, Y. Jia, R. Geng, J. Jia, and N. Z. Gong, "Prompt injection attacks and defenses in llm-integrated applications," arXiv preprint arXiv:2310.12815, 2023.

[311] P. Taveekitworachai, F. Abdullah, M. C. Gursesli, M. F. Dewantoro, S. Chen, A. Lanata, A. Guazzini, and R. Thawonmas, "Breaking bad: Unraveling influences and risks of user inputs to chatgpt for game story generation," in International Conference on Interactive Digital Storytelling. Springer, 2023, pp. 285–296.

[312] G. Deng, Y. Liu, Y. Li, K. Wang, Y. Zhang, Z. Li, H. Wang, T. Zhang, and Y. Liu, "Jailbreaker: Automated jailbreak across multiple large language model chatbots," arXiv preprint arXiv:2307.08715, 2023.

[313] T. Liu, Z. Deng, G. Meng, Y. Li, and K. Chen, "Demystifying rce vulnerabilities in llm-integrated apps," arXiv preprint arXiv:2309.02926, 2023.

[314] I. Drori and D. Te'eni, "Human-in-the-loop ai reviewing: Feasibility, opportunities, and risks," Journal of the Association for Information Systems, vol. 25, no. 1, pp. 98–109, 2024.

[315] X. Li, F. Tramèr, P. Liang, and T. Hashimoto, "Large language models can be strong differentially private learners," 2022.

[316] K. Zhu, J. Wang, J. Zhou, Z. Wang, H. Chen, Y. Wang, L. Yang, W. Ye, Y. Zhang, N. Z. Gong, and X. Xie, "Promptbench: Towards evaluating the robustness of large language models on adversarial prompts," 2023.

[317] A. Zafar, V. B. Parthasarathy, C. L. Van, S. Shahid, A. I. khan, and A. Shahid, "Building trust in conversational ai: A comprehensive review and solution architecture for explainable, privacy-aware systems using llms and knowledge graph," 2023.

[318] O. J. Romero, J. Zimmerman, A. Steinfeld, and A. Tomasic, "Synergistic integration of large language models and cognitive architectures for robust ai: An exploratory analysis," in Proceedings of the AAAI Symposium Series, vol. 2, no. 1, 2023, pp. 396–405.

[319] P. Ganesh, H. Chang, M. Strobel, and R. Shokri, "On the impact of machine learning randomness on group fairness," ser. FAccT '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 1789–1800.

[320] N. Ousidhoum, X. Zhao, T. Fang, Y. Song, and D.-Y. Yeung, "Probing toxic content in large pre-trained language models," C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 4262–4274.

[321] M. Ivgi and J. Berant, "Achieving model robustness through discrete adversarial training," 2021.

[322] J. Y. Yoo and Y. Qi, "Towards improving adversarial training of nlp models," 2021.

[323] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in Advances in Neural Information Processing Systems, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 27 730–27 744.

[324] Z. Yuan, H. Yuan, C. Tan, W. Wang, S. Huang, and F. Huang, "Rrhf: Rank responses to align language models with human feedback without tears," 2023.

[325] J. Kirchenbauer, J. Geiping, Y. Wen, M. Shu, K. Saifullah, K. Kong, K. Fernando, A. Saha, M. Goldblum, and T. Goldstein, "On the reliability of watermarks for large language models," 2023.

[326] Z. Wei, Y. Wang, and Y. Wang, "Jailbreak and guard aligned language models with only few in-context demonstrations," 2023.

[327] L. Li, D. Song, and X. Qiu, "Text adversarial purification as defense against adversarial attacks," 2023.

[328] X. Sun, X. Li, Y. Meng, X. Ao, L. Lyu, J. Li, and T. Zhang, "Defending against backdoor attacks in natural language generation," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 4, pp. 5257–5265, Jun. 2023.

[329] Z. Xi, T. Du, C. Li, R. Pang, S. Ji, J. Chen, F. Ma, and T. Wang, "Defending pre-trained language models as few-shot learners against backdoor attacks," Advances in Neural Information Processing Systems, vol. 36, 2024.

[330] M. Xiong, Z. Hu, X. Lu, Y. Li, J. Fu, J. He, and B. Hooi, "Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms," 2023.

[331] M. Phute, A. Helbling, M. Hull, S. Peng, S. Szyller, C. Cornelius, and D. H. Chau, "Llm self defense: By self examination, llms know they are being tricked," 2023.

[332] National Institute of Standards and Technology, "Nist cybersecurity framework," ***https://www.nist.gov/cyberframework***, 2024, accessed: 2024-02-20.

[333] G. Deng, Y. Liu, V. Mayoral-Vilches, P. Liu, Y. Li, Y. Xu, T. Zhang, Y. Liu, M. Pinzger, and S. Rass, "Pentestgpt: An llm-empowered automatic penetration testing tool," arXiv preprint arXiv:2308.06782, 2023.

[334] D. K. Kholgh and P. Kostakos, "Pac-gpt: A novel approach to generating synthetic network traffic with gpt-3," IEEE Access, 2023.

[335] P. Balasubramanian, S. Nazari, D. K. Kholgh, A. Mahmoodi, J. Seby, and P. Kostakos, "Tstem: A cognitive platform for collecting cyber threat intelligence in the wild," arXiv preprint arXiv:2402.09973, 2024. [Online]. Available: ***https://arxiv.org/pdf/2402.09973.pdf***

[336] F. Setianto, E. Tsani, F. Sadiq, G. Domalis, D. Tsakalidis, and P. Kostakos, "Gpt-2c: A parser for honeypot logs using large pre-trained language models," in Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2021, pp. 649–653.

[337] H. Guo, S. Yuan, and X. Wu, "Logbert: Log anomaly detection via bert," in 2021 international joint conference on neural networks (IJCNN). IEEE, 2021, pp. 1–8.

[338] V.-H. Le and H. Zhang, "Log parsing with prompt-based few-shot learning," arXiv preprint arXiv:2302.07435, 2023.

[339] P. Balasubramanian, J. Seby, and P. Kostakos, "Transformer-based llms in cybersecurity: An in-depth study on log anomaly detection and conversational defense mechanisms," in 2023 IEEE International Conference on Big Data (BigData). IEEE, 2023, pp. 3590–3599.

[340] M. Kaheh, D. K. Kholgh, and P. Kostakos, "Cyber sentinel: Exploring conversational agents in streamlining security tasks with gpt-4," arXiv preprint arXiv:2309.16422, 2023.

[341] T. Ali and P. Kostakos, "Huntgpt: Integrating machine learning-based anomaly detection and explainable ai with large language models (llms)," arXiv preprint arXiv:2309.16021, 2023.

[342] T. Taleb, C. Benzäıd, M. Bordallo Lopez, K. Mikhaylov, S. Tarkoma, P. Kostakos, N. H. Mahmood, P. Pirinen, M. Matinmikko-Blue, M. Latva-Aho et al., "6g system architecture: A service of services vision," ITU journal on future and evolving technologies, vol. 3, no. 3, pp. 710–743, 2022.

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

[344] I. Ahmad, J. Malinen, F. Christou, P. Porambage, A. Kirstädter, and J. Suomalainen, "Security in intent-based networking: Challenges and solutions," Authorea Preprints, 2023.

[345] A. Clemm, L. Ciavaglia, L. Z. Granville, and J. Tantsura, "Intent-based networking-concepts and definitions," IRTF draft work-in-progress, 2020.

[346] I. Ahmad, S. Shahabuddin, T. Kumar, J. Okwuibe, A. Gurtov, and M. Ylianttila, "Security for 5g and beyond," IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3682–3722, 2019.

[347] European Telecommunications Standards Institute, "5G; 5G System; Network Data Analytics Services; Stage 3," European Telecommunications Standards Institute (ETSI), Technical Specification TS 129 520 V15.3.0, Apr. 2019, 3GPP TS 29.520 version 15.3.0 Release 15. [Online]. Available: ***https://www.etsi.org/deliver/etsits/129500129599/129520/15.03.0060/ts129520v150300p.pdf***.

[348] ——, "5G; Procedures for the 5G System (5GS)," European Telecommunications Standards Institute (ETSI), Technical Specification TS 123 502 V16.7.0, Jan. 2021, 3GPP TS 23.502 version 16.7.0 Release 16. [Online]. Available: ***https://www.etsi.org/deliver/etsits/123500123599/123502/16.07.0060/ts123502v160700p.pdf***.

[349] N. F. S. de Sousa, D. L. Perez, C. E. Rothenberg, and P. H. Gomes, "End-to-end service monitoring for zero-touch networks," Journal of ICT Standardization, pp. 91–112, 2021.

[350] C. De Alwis, A. Kalla, Q.-V. Pham, P. Kumar, K. Dev, W.- J. Hwang, and M. Liyanage, "Survey on 6g frontiers: Trends, applications, requirements, technologies and future research," IEEE Open Journal of the Communications Society, vol. 2, pp. 836– 886, 2021.

[351] S. Sheikhi and P. Kostakos, "Ddos attack detection using unsupervised federated learning for 5g networks and beyond," in 2023 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit). IEEE, 2023, pp. 442–447.

[352] J. Gallego-Madrid, R. Sanchez-Iborra, P. M. Ruiz, and A. F. Skarmeta, "Machine learning-based zero-touch network and service management: A survey," Digital Communications and Networks, vol. 8, no. 2, pp. 105–123, 2022.

[353] E. Coronado, R. Behravesh, T. Subramanya, A. Fernandéz-Fernandéz, S. Siddiqui, X. Costa-Pérez, and R. Riggio, "Zero touch management: A survey of network automation solutions for 5g and 6g networks," IEEE Communications Surveys & Tutorials, 2022.

[354] M. Liyanage, Q.-V. Pham, K. Dev, S. Bhattacharya, P. K. R. Maddikunta, T. R. Gadekallu, and G. Yenduri, "A survey on zero touch network and service management (zsm) for 5g and beyond networks," Journal of Network and Computer Applications, vol. 203, p. 103362, 2022.

[355] A. B. Z. Mahmoodi, S. Sheikhi, E. Peltonen, and P. Kostakos, "Autonomous federated learning for distributed intrusion detection systems in public networks," IEEE Access, vol. 11, pp. 121 325– 121 339, 2023.

[356] S. Sheikhi and P. Kostakos, “Cyber threat hunting using unsupervised federated learning and adversary emulation,” in 2023 IEEE International Conference on Cyber Security and Resilience (CSR). IEEE, 2023, pp. 315–320.

[357] Y. Gao, Z. Song, and J. Yin, “Gradientcoin: A peer-to-peer decentralized large language models,” 2023.

[358] Z. Tang, Y. Wang, X. He, L. Zhang, X. Pan, Q. Wang, R. Zeng, K. Zhao, S. Shi, B. He, and X. Chu, “Fusionai: Decentralized training and deploying llms with massive consumer-level gpus,” 2023.

[359] S. Han, Q. Zhang, Y. Yao, W. Jin, Z. Xu, and C. He, “Llm multi-agent systems: Challenges and open problems,” 2024.

[360] Y. Gong, “Dynamic large language models on blockchains,” 2023.

[361] H. Luo, J. Luo, and A. V. Vasilakos, “Bc4llm: Trusted artificial intelligence when blockchain meets large language models,” 2023.

[362] W. Huang, Y. Wang, A. Cheng, A. Zhou, C. Yu, and L. Wang, “A fast, performant, secure distributed training framework for large language model,” 2024.

Large Language Models in the 6G-Enabled Computing Continuum: a White Paper

List of authors

Markus Abel, Ambrosys GmbH, Germany · Ijaz Ahmad, VTT Technical Research Centre of Finland, Finland · Constantino Alvarez Casado, Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland · Rico Berner, Ambrosys GmbH, Germany · Mickaël Bettinelli, LISTIC, University Savoie Mont Blanc, France · Kaj Mikael Björk, Centre For Intelligent Computing (CIC), University of Turku, Finland · Michele Capobianco, Capobianco, Italy · James Gross, KTH Royal Institute of Technology, Sweden · Tri Hong Nguyen, Department of Computer Science, Aalto University, Finland · Pan Hui, Hong Kong University of Science and Technology (GuangZhou), China · Panos Kostakos, European Public Prosecutors Office (EPPO), Luxembourg · Abhishek Kumar, Center for Ubiquitous Computing (UBICOMP), University of Oulu, Finland · Mika-Petri Laakkonen, Oulu University of Applied Sciences, Finland · Xiaoli Liu, University of Helsinki, Finland · Zhi Liu, The University of Electro-Communications (UEC), Japan · Le Nguyen, Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland · Huong Nguyen, Center for Ubiquitous Computing (UBICOMP), University of Oulu, Finland · Basak Ozan Ozparlak, Ozyegin University, Türkiye · Ville Pietiläinen, University of Lapland, Finland · Susanna Pirttikangas, Center for Ubiquitous Computing (UBICOMP), University of Oulu, Finland · Stéphan Plassart, LISTIC, University Savoie Mont Blanc, France · Sampo Pyysalo, University of Turku, Finland · Soheyb Ribouh, LITIS, Université de Rouen Normandie, France · Jari Rinne, University of Lapland, Finland · Mehdi Safanpour, Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland · Alaa Saleh, Center for Ubiquitous Computing (UBICOMP), University of Oulu, Finland · Saeid Sheikhi, Center for Ubiquitous Computing (UBICOMP), University of Oulu, Finland · Olli Silvén, Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland · Harry Souris, Silo AI, Finland · Xiang Su, University of Helsinki, Finland · Roope Suomalainen, Silo AI, Finland · Athanasios V. Vasilakos, University of Agder, Norway · Aleksandr Zavodovski, Ericsson, Finland · Qi Zhang, Aarhus University, Denmark · Peng Yuan Zhou, Aarhus University, Denmark · Alireza Zourmand, Häme University of Applied Sciences (HAMK), Finland

Editors:

Lauri Lovén, Center for Ubiquitous Computing (UBICOMP), University of Oulu, Finland · Miguel Bordallo López, Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland · Roberto Morabito, EURECOM, France · Jaakko Sauvola, Empirical Software Engineering in Software, Systems, and Services Research Unit (M3S), University of Oulu, Finland · Sasu Tarkoma, University of Helsinki, Finland

Reviewers:

Marja Matinmikko-Blue, Infotech Oulu and Centre for Wireless Communications (CWC), University of Oulu, Finland · Marcos Katz, Centre for Wireless Communications (CWC), University of Oulu, Finland · Shahriar Shahabuddin, Oklahoma State University, US · Mehdi Bennis, Centre for Wireless Communications (CWC), University of Oulu, Finland

6G Flagship, University of Oulu, Finland · January 2025

6G Research Visions, No. 14, 2025

ISSN 2669-9621 (print) · ISSN 2669-963X (online)

ISBN 978-952-62-4375-7 (print) · ISBN 978-952-62-4376-4 (online)

6G FLAGSHIP
UNIVERSITY OF OULU

6gflagship.com