Advancing Generative AI Implementation in Telecommunications Networks

EXECUTIVE SUMMARY

Generative Artificial Intelligence (GenAI) is a rapidly evolving technology that will transform the telecommunications industry by enhancing the efficiency, personalization, and security of network operations across domains such as Radio Access Networks (RAN), core networks, devices, and network management. This white paper surveys key use cases of GenAI, including RAN optimization, digital twins, network slicing, and AI-enhanced troubleshooting. While GenAI shows great potential, the telecommunications sector lacks focused studies on its specific applications. This white paper addresses these gaps by analyzing real-world use cases and offering recommendations to facilitate the integration of GenAI technologies into telecom networks, aiming to improve operational efficiency and foster future innovations. It provides actionable recommendations that will drive the telecommunications industry forward while complementing the ongoing efforts by Standards Development Organizations (SDOs).

FOREWORD

As a leading technology and solutions development organization, the Alliance for Telecommunications Industry Solutions (ATIS) brings together the top global ICT companies to advance the industry's most pressing business priorities. ATIS' nearly 200 member companies are currently working to address the all-Internet Protocol (IP) transition, 5G, NF virtualization, big data analytics, cloud services, device solutions, emergency services, M2M, cyber security, network evolution, quality of service, billing support, operations, and much more. These priorities follow a fast-track development lifecycle — from design and innovation through standards, specifications, requirements, business use cases, software toolkits, open source solutions, and interoperability testing.

ATIS is accredited by the American National Standards Institute (ANSI). The organization is the North American Organizational Partner for the 3rd Generation Partnership Project (3GPP), a founding Partner of the oneM2M global initiative, a member of and major U.S. contributor to the International Telecommunication Union (ITU), as well as a member of the Inter-American Telecommunication Commission (CITEL). For more information, visit www.atis.org.

TABLE OF CONTENTS

Executive Summary i

Foreword ii

1 Introduction 1

2 GenAI in RANs 2

Radio Channel Modeling 2

Digital Twin RAN 2

Channel State Information Feedback for Massive MIMO 2

Beamforming 2

Spectrum Management and Sharing 2

Case Study: Toward a Wireless Physical-Layer Foundation Model 3

3 GenAI in Core Networks 6

Overview of Core Network Use Cases 6

Combatting Voice Call Fraud 6

4 GenAI in Device 7

On-Board Device AI — Imagining Next Mobile Experiences 7

Case Study: Implementing AI-Generated Content (AIGC) on Resource-Constrained Devices 8

5 GenAI in Network Management 10

Overview of Network Management Use Cases 10

Customer Support — Ticket-based Troubleshooting 10

Case Study: Leveraging AI for Enhanced Ticket-Based Troubleshooting 10

Case Study: Reinforcement Learning with LLM Interaction for GenAI in Network Management 13

Neuro-Symbolic AI for Enhanced Reasoning and Decision-Making 15

Case Study: Compliance with Upcoming Regulations 15

Case Study: Translating Policies between Different Constituencies 16

Network Operations Use Cases 17

6 GenAI Across Domains 19

Overview of Cross-Domain Use Cases 19

Network Slicing 19

Datasets for Developing Telecom-Specific Language Models 19

Case Study: Adapting Language Models for Telecom Applications 20

Retrieval Augmented Generation-Based AI Chatbot for Telecom 22

Cognitive Digital Twins 23

7 Gap Analysis 25

Technical and Infrastructure Challenges 25

Operational and Strategic Needs 26

Regulatory and Ethical Considerations 27

Toward Responsible AI 30

Transparency 31

8 Recommendations 32

9 Conclusion 33

TABLE OF CONTENTS

10 Appendix A: Background on Semantics and Cognition 34

11 Acronyms and Abbreviations 36

12 References 38

Acknowledgements 41

Copyright and Disclaimer 44

1. INTRODUCTION

GenAI is a category of AI systems designed to create various new content, such as text, images, sounds, animations, videos, three-dimensional (3D) models, and computer code. It can leverage a wide range of AI/Machine Learning (ML) techniques, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), diffusion models, reinforcement learning, and transformers, among others. A prominent example of GenAI applications is OpenAI's ChatGPT, which is a Large Language Model (LLM) based on the Generative Pre-trained Transformer (GPT).

GenAI development is moving quickly, transforming business operations, service offerings, and productivity across industries, including telecommunications. GenAI has the potential to reinvent the telecommunication industry, enabling customers to use telecommunication services more easily and efficiently. Concurrently, it enables service providers to offer more secure, reliable, and personalized services.

Although much of the focus of this white paper is on wireless networks, it is important to recognize that the applicability of GenAI extends across various types of telecom networks — wireline, fiber, and satellite — in addition to wireless. These networks can benefit from AI-enhanced network automation, real-time data analytics, traffic optimization, and dynamic network configurations. GenAI can also assist in predictive maintenance, customer service chatbots, and adaptive resource allocation, regardless of the specific network type.

Despite various telecom SDOs discussing AI/ML, such as the comprehensive work conducted by the 3rd Generation Partnership Project (3GPP) [1], focused studies on GenAI in telecommunications remain limited. To address this gap, ATIS launched a new AI Network Applications Working Group, which conducted a first-of-its-kind study on GenAI in telecommunications, surveying GenAI use cases across the network. This work aims to assess key use cases for network applications, address critical gaps, and provide recommendations to help advance AI implementation across networks.

The rest of this white paper is organized as follows:

- Sections 2, 3 4, and 5 survey GenAI use cases in telecommunications for RAN, core, device, and management domains, respectively.
- Section 6 examines GenAI use cases that span across domains.
- Section 7 provides a gap analysis on the application of GenAI in telecommunications.
- Section 8 highlights key recommendations for applying GenAI across networks, followed by concluding remarks in Section 9.

2. GenAI in RANs

GenAI holds promise in optimizing and enhancing RANs. This section surveys several selected use cases for GenAI in RAN. The discussion exemplifies the role of GenAI techniques in RAN but is not meant to be exhaustive.

Radio Channel Modeling

Radio channel modeling is a fundamental requirement in the design, evaluation, and optimization of wireless communication systems. Stochastic channel models have been widely used, especially for the link and system performance evaluation in standards bodies such as 3GPP. Deterministic radio channel modeling with ray tracing is another popular approach, gaining increasing popularity for emerging use cases such as digital twin, Integrated Sensing and Communication (ISAC), and Reconfigurable Intelligence Surface (RIS).

The objective of the radio channel modeling is to generate wireless channel data for use in system design, evaluation, and optimization. To this end, GenAI techniques can complement traditional channel modeling methods by generating synthetic data that mimics real-world radio channel conditions in a comprehensive and diverse set of scenarios. For example, in [2], a GAN consisting of a channel data generator and a channel data discriminator was trained on raw channel measurement data. After training, the resulting channel data generator was used to generate channel data for the target application scenario.

Digital Twin RAN

Digital twin RAN is anticipated to play a key role in future RAN optimization by enabling planning, simulating, and monitoring in a virtual twin RAN. A digital twin consists of a physical part, a digital part, and a connection between them. Building a digital twin RAN requires 3D models of objects and scenes in the physical environment.

Applying generative models of 3D content is considered one of the most promising ways to synthesize 3D shapes and scenes [3]. GenAI can analyze various data sources, including geographical information, building structures, and signal propagation models, to generate highly accurate 3D maps of the radio environment [4]. This assists in predicting signal coverage, path loss, and interference, aiding in RAN planning and optimization. In some cases, generating a radio coverage map for RAN planning and optimization may be sufficient. To this end, [5] proposed a conditional GAN to estimate fine-resolution radio maps from sparse radio strength measurements.

Channel State Information Feedback for Massive MIMO

Massive Multiple-input Multiple-Output (MIMO) is an intrinsic, standardized technology component in 5G. Although massive MIMO has been deployed successfully in a commercial RAN, there is much room for further improvement, including overhead reduction and accuracy improvement in Channel State Information (CSI) feedback.

Using AI/ML to enhance CSI feedback for massive MIMO in 5G is a key use case under investigation in 3GPP [6]. The basic idea of AI/ML-enabled CSI feedback is to perform non-linear compression of the CSI to a lower-dimensional latent representation by an encoder at the device and decompress the latent representation by a decoder at the base station. Various GenAI techniques have been studied to improve the CSI feedback performance for massive MIMO, such as deep convolutional GAN [7] and transformer-based architectures [8].

Beamforming

5G can operate at a wide range of frequencies, ranging from sub-6 GHz to millimeter wave frequencies. To support operation over such a wide range of carrier frequencies, 5G has been designed to utilize beam-based operation, where both base stations and devices can use transmit and receive beamforming for all channels and signals [9].

GenAI can be applied to beamforming in RANs by assisting in the optimization, adaptation, and improvement of beamforming techniques. It can be used to adjust antenna configurations to maximize signal strength and quality in changing radio environments. For example, [10] utilized a GAN to represent the radio channel as a low-dimensional manifold, where the optimum beamformers were searched, and [11] also proposed a transformer-based beamforming design where separate transformer encoders were utilized to implement different parts of hybrid beamforming optimization.

Spectrum Management and Sharing

The radio frequency spectrum is a finite and valuable resource. 5G RAN operates across a broad spectrum, including low-, mid-, and high-frequency bands, while co-existing with non-terrestrial networks, radar systems, and Wi-Fi in certain bands. GenAI can be applied to spectrum management and sharing so that the limited spectrum resource is used optimally, allowing for more users, better service quality, and increased data transfer capacity within the available frequency bands. For example, [12] developed a GAN for a spectrum sensing problem, where the objective was to detect the presence of an emitter from spectrum measurements, and [13] detected abnormalities inside the

radio spectrum by learning generative models, including a conditional GAN and a dynamic Bayesian network.

Overall, by utilizing GenAI techniques, RAN can become more adaptive, efficient, and capable of self-optimization and self-learning, ultimately leading to improved network performance, reliability, and user experience.

Case Study: Toward a Wireless Physical-Layer Foundation Model

Background and Motivation

Very recently, key AI domains have undergone a paradigm shift with the introduction of foundation models. A foundation model acts as a template for building AI systems in which a model trained on a large amount of unlabeled data can be adapted to many different downstream tasks, even with different types of data. For example, a foundation model for image generation can cope with a wide variety of styles and shapes. Upon showing just a few new samples of a new style, the foundation model is adapted toward a new downstream task (e.g., generation of images using the new style) without significant effort required.

LLMs have already been popularized in wireless networks. Examples of tasks include answering healthcare-related questions, classifying 3GPP working groups based on technical specifications, answering telecom-related questions, network configurations, etc. However, because these models lack an understanding of physical wireless signals, the development of a Wireless Physical Layer Foundation Model (WPFM) is crucial and provides a semantic understanding of the physical wireless environment. Unlike text-based configurations in the higher layers of wireless networks, the realization of a WPFM requires dealing with complex and dynamic data in the form of time series. The adoption of ML innovations can be significantly accelerated with a model that can understand and represent these time series.

Trained foundation models using architectures such as transformer networks will eliminate the need for exhaustive research, domain expertise, and expensive data collection because they enable improved representation learning across multiple downstream tasks. It will be possible to integrate WPFMs with LLMs, which focus on higher layers of the network stack. In these layers, pre-trained LLMs, such as ChatGPT, bidirectional encoder representations from transformers (BERT), and Llama (Large Language model Meta AI), have been very recently investigated and can already perform general tasks well.

Challenges

In building a wireless physical layer foundation model, the following challenges are identified:

Challenge 1: Novel methods are required for wireless pre-training tasks

Modern ML architectures are capable of handling complex pattern learning and scaling of many parameters. The transformer neural network, for example, excels in efficiently learning sequential data with parallel training. However, in wireless networks, the transformer architecture is missing its potential by being trained to perform individual tasks. Such approaches have a more simplified implementation compared to foundation transformer models, which instead require effective un-/self-supervised pre-training tasks before being adapted to multiple downstream tasks. However, due to the limited data availability in the past and challenges in supporting heterogeneous data types, there is a gap in the wireless domain for effectively creating such pre-training tasks.

Challenge 2: WPFMs need to support the embedding of heterogeneous wireless time series with different lengths, sampling rates, and data types

Creating a foundation model for wireless time series, such as In-phase and Quadrature (IQ) samples and Channel Impulse Responses (CIRs), diverges significantly from models in the Natural Language Processing (NLP) and vision domains due to the data's heterogeneous temporal nature. Developing a WPFM demands a deep understanding of time-sensitive heterogeneous data dynamics (i.e., supporting samples from use cases with different lengths), sampling rates, and data types (IQ, received signal strength indicator, CIR, metadata, etc.), setting it apart as a compelling and distinctive domain within the broader AI landscape.

Challenge 3: WPFMs need human-understandable interaction and prompt-based optimization

The semantic description of wireless networks and their environments has been a recent topic for standalone wireless ML models. Combining textual descriptions with the first two challenges in a foundation model allows AI to be used in many downstream applications and opens new possibilities. For example, by combining the embedded information from WPFMs with such semantic descriptions, LLMs may autonomously understand wireless data from the physical layer, enabling more sophisticated prompt-based applications (e.g., using chain-of-thought) and optimization. The state of the art lacks the ability for wireless optimization with LLM and WPFM combined and is limited to LLM-based configurations using standardization and specification documents.

A Wireless Physical-Layer Foundation Model Strategic Framework

To address these challenges, [25] presented a WPFM framework. This foundation model, illustrated in Figure 1, has a simple but transformative goal: It can understand many types of physical wireless signals (in time series form) and metadata, while enabling network configurations and optimization and allowing user interaction. Using this framework and large openly available datasets, WPFM can be enabled to adapt new downstream tasks.

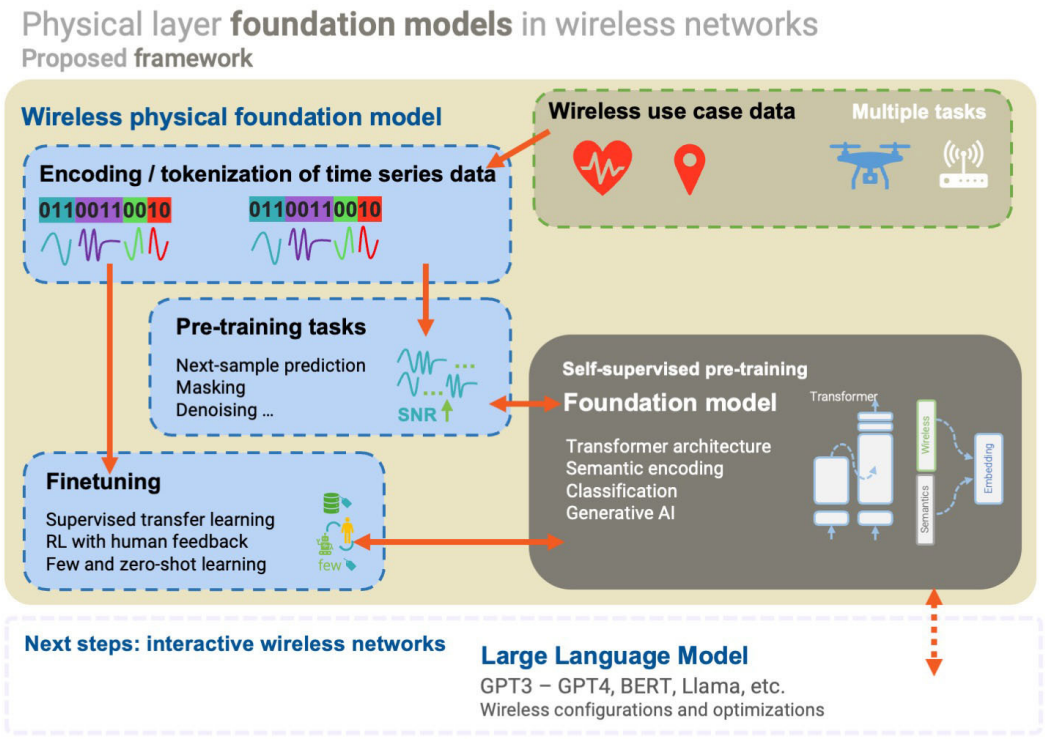


Figure 1: The Different Framework Strategies of a WPFM [25]

Embedding and tokenization of physical-layer time-series

One strategy is to use byte-pair encoding to merge frequently occurring signal or pulse IQ sample pairs until a fixed input size of the model is obtained. Another strategy is to represent known pulses and short-term patterns in time series as tokens. Alternatively, time-variable-based tokenization can be used to tokenize time series where the token uniformity inductive bias would act on the variable dimension.

In addition to encoding, it is essential to define the vocabulary size to support variable input lengths while avoiding unnecessary complexity. For instance, by applying tokenization techniques such as those utilized in models like BERT or GPT, a uniform sample distribution could be achieved, regardless of the sampling rate. This results in a similar embedding space even when comparing signals sampled at different rates or when subjected to input noise.

In wireless use cases, these techniques allow CIRs with larger and shorter lengths to be understood by the framework, as well as technology recognition using IQ streams with different sampling rates and sample lengths. The shape of pulses or sine waves contains information about the wireless technology and modulation used or represents reflections in CIR data. These waves or pulses can be represented as tokens.

Effective self-supervised pre-training tasks for a WPFM

Pre-training tasks are essential because they allow the model (e.g., a transformer) to learn general features and representations from a vast amount of diverse data before fine tuning on a specific task. Allowing self-supervised pre-training is especially interesting in the wireless domain, which suffers from a low volume of labeled datasets.

In wireless networks, the following techniques may be envisioned:

- Next sample/pulse/multipath prediction, aiming at predicting sequential information.
- Masking which can learn to predict missing samples/pulse/multipath within the signal.
- Denoising learns to remove additive Gaussian noise from signals.
- Sequence order prediction learns to order signals in chronological order.
- GANs learn internal representations of the signal to generate realistic new versions of the signal under different conditions.

Semantic physical-layer wireless representation learning

To generate semantic information, textual descriptions of the wireless network and environment need to be integrated into the foundation model. These descriptions can be obtained from experts or pre-trained LLMs. To fuse text and wireless time series, multimodal fusion foundation models can be designed.

A simple approach is to share the latent space, which enables joint representations of text and time series modalities (e.g., using shared projected layer mapping, cross-attention layers, and the use of contrastive learning). In contrast, separated

latent space representations can be investigated, which are more flexible and allow different architectures or model reuse for both text and time series. The textual description differs from the labels as it represents semantic information that can be provided along the wireless time-series data. For example, labels are the wireless technologies shared in the considered spectrum, while semantic textual descriptions can be the size and type of the environment, similarly to image captions in the computer vision domain.

Learning semantic representations in wireless networks enables the foundation model to produce semantic information about the wireless network and environment in a textual form. In addition, it can generate time series variations based on semantic descriptions in new environments and wireless technologies.

Fine-tuning WPFMs

Although the foundation model may already be able to perform general tasks and applications, fine-tuning allows it to adapt the model to specific downstream tasks or domains.

To enable fine-tuning, the following techniques are vital:

- Supervised transfer learning with a small, labeled dataset (e.g., tens of samples).
- Reinforcement learning with human feedback (e.g., grading the predictions made by the foundation model before and during model deployment).
- Zero- and few-shot learning by providing relevant context and the necessary information using Retrieval Augmented Generation (RAG).
- Example input-output pairs of the expected downstream task. Additionally, GenAI capabilities can enrich small datasets and allow these to be used for fine tuning the model for the considered downstream tasks.

Conclusions

This case study underscores the upcoming transformative work within the field of AI for wireless networks, evolving from task-specific models to the innovative paradigm of adaptable WPFMs. The challenges identified highlight the complexities inherent in developing a unified model for wireless applications. The described WPFM architecture aims to revolutionize AI development in wireless networks by accelerating and sharing AI advances.

The semantic capabilities of a WPFM are a crucial step toward human interactive wireless networks. Such interactivity can bridge the gap between expert-based manual configurations and automatic prompt-based configurations using LLMs. To realize the goal of a unified WPFM, shared across many downstream tasks, upcoming research should focus on refining and standardizing pre-training tasks tailored for the wireless domain. Additionally, real-world implementations are essential to empirically validate the effectiveness of the WPFMs.

3. GenAI in CORE NETWORKS

Overview of Core Network Use Cases

GenAI can be applied across multiple aspects of the core network. It facilitates extensive data analysis on user behavior, preferences, and patterns. This deep analysis allows networks to provide highly personalized services to end users, including customized content recommendations and tailored network settings, thereby boosting user engagement and satisfaction. By leveraging GenAI, networks can securely expose certain capabilities to third-party developers, such as real-time data analytics and network functionalities. This exposure enables the creation of innovative applications and services that can operate within or alongside the network, opening new revenue streams and service models.

GenAI plays a crucial role in network slicing, where it helps create and manage multiple virtual networks on a single physical infrastructure. Each slice can be optimized for specific types of traffic, applications, or services, thus enabling more efficient resource utilization and better overall performance. Through predictive analytics and real-time decision-making capabilities, GenAI effectively manages network traffic. It dynamically allocates bandwidth and prioritizes network resources based on current demand and predicted future load, thus preventing congestion and ensuring smooth service delivery.

Combatting Voice Call Fraud

GenAI can be applied in the core network as a tool to combat voice call fraud. AI analysis can be used in real time to alert subscribers if a phone call is suspicious and to explain the reasons, such as requests for sensitive information like a Personal Identification Number (PIN) — a request that a genuine bank would not make.

The approach to applying AI to telephone calls involves delivering the service from the network itself rather than the handset. This allows operators to control how the service is packaged and delivered to their customers. It also extends access to a wider range of phone users, including older subscribers who may be more vulnerable to telephone fraud.

Two technical components are necessary to apply AI to telephone calls: a gateway to channel the call into the cloud and the AI services themselves. Microsoft has implemented such a service, Azure Operator Call Protection, using a communications gateway that provides advanced Session Initiation Protocol (SIP), Real-time Transport Protocol (RTP), and Hypertext Transfer Protocol (HTTP) interoperability functions to integrate operator voice networks with Azure and cloud communications services.

The use of GenAI in telephone calls involves several steps. First, the call is transcribed to text using Azure services. Then, AI techniques are applied to the call content to detect potentially fraudulent calls. The output from the AI engine is then used to notify subscribers of attempted scams by sending them a text message, which includes a summary of why the call is a scam and mitigation actions the subscriber should take.

4. GenAI in DEVICE

On-Board Device AI — Imagining Next Mobile Experiences

Consider a scenario where smartphones, tablets, or wearable devices not only respond to commands but also anticipate needs, learn from user habits, and deliver real-time, personalized experiences without the need for remote cloud servers. This potential is realized through on-device AI and edge computing, a transformative technology set to redefine user interaction with mobile devices.

With the rapid advancements in LLMs and AI technologies, users now expect their devices to be smarter and more intuitive. On-device AI and edge computing are set to meet these expectations by integrating powerful AI capabilities directly into mobile devices. This integration promises a future where devices are not only faster and more efficient but also capable of providing offline capabilities and enhanced security.

The advantages of on-device AI are significant. By processing data locally, these technologies can greatly reduce latency, resulting in faster and more reliable responses. This real-time processing is essential for applications that require immediate interaction, such as Augmented Reality (AR), gaming, and health monitoring. Additionally, new AI chips that incorporate Central Processing Units (CPUs), Graphics Processing Units (GPUs), Neural Processing Units (NPUs), and Tensor Processing Units (TPUs) are designed to provide powerful processing with lower energy consumption, which is crucial for mobile and wearable devices with limited battery life.

These advancements are supported by various types of AI chips, each tailored to specific tasks:

- **CPU:** The general-purpose processor handles a wide range of tasks with moderate to high power consumption.
- **GPU:** Originally designed for rendering graphics, GPUs are now widely used for AI tasks, offering high parallel processing capability.
- **NPU:** Optimized for AI tasks, particularly neural networks, NPUs are highly efficient and consume significantly less power.
- **TPU:** Custom-designed for AI tasks, TPUs are highly efficient and specifically tailored to accelerate AI workloads.

As these chips are integrated into mobile devices, they will enable a host of advanced applications:

- **Holographic Telepresence:** This technology enables real-time, lifelike holographic calls that let individuals interact with others as if they were physically present, regardless of their location.
- **Personal Health Guardian:** Devices have the capability to continuously monitor an individual's health, providing predictive diagnostics and real-time alerts for potential health issues.
- **Quantum Gaming:** Offers immersive Virtual Reality (VR)/AR gaming experiences featuring dynamic real-time world-building and AI-driven characters that adapt and learn from the player's behavior.
- **Smart Home Orchestrator:** AI-driven home automation is designed to learn the user's preferences, anticipate their needs, and seamlessly integrate with all their smart devices for an intuitive living environment.

The potential of on-device AI is considerable. Advanced AI chips will improve device performance and enable new functionalities. They will support enhanced biometric authentication, real-time data encryption, and federated learning capabilities to ensure the privacy and security of users' data. Quantum computing elements may handle complex problem-solving tasks, while real-time holographic rendering will enable high-fidelity AR and VR applications. Continuous health monitoring with AI algorithms will offer personalized health insights, revolutionizing how individuals approach their well-being.

Looking ahead, the future may bring even more innovative applications:

- **Personal AI Assistants:** Context-aware assistants could provide real-time recommendations and emotional support, tailored to individual needs and preferences.
- **Advanced Language Translation:** Real-time, highly accurate language translation could make communication seamless across different languages.
- **Smart Communication:** Optimized network usage could ensure high-quality communication, even in areas with low signal strength.
- **Personalized Education and Training:** Education could be tailored based on aptitude and interest and integrated with community activities for a more engaging learning experience.

In summary, on-device AI and edge computing are expected to significantly change how users interact with mobile devices. This technological evolution promises to deliver smarter, more responsive, and deeply integrated experiences, paving the way for a future where devices are not just tools but intelligent companions that enhance every aspect of life.

Case Study: Implementing AI-Generated Content on Resource-Constrained Devices

GenAI models have revolutionized content creation, enabling the generation of diverse and high-quality outputs across various domains. Techniques such as stable diffusion have demonstrated remarkable capabilities in producing visually impressive images from textual descriptions. These advancements have made AI-generated content (AIGC) increasingly popular in applications like image inpainting, text-to-image generation, and audio synthesis [20].

However, the implementation of AIGC models on resource-constrained devices, such as mobile phones, presents significant challenges. These challenges include high computational demands and substantial energy consumption, particularly during the denoising steps of the diffusion process. To address these limitations, [21] proposed a collaborative distributed diffusion-based AIGC framework. This framework leverages the collaborative capabilities of devices within wireless networks to optimize edge computation resources, reduce energy consumption, and enhance user privacy. By distributing the computational load and enabling device collaboration, the approach facilitates the efficient execution of AIGC tasks, paving the way for the practical deployment of ubiquitous AIGC services.

Overview of Diffusion Models

Diffusion models function by systematically degrading training data with Gaussian noise and then learning to restore this data through incremental denoising steps [22]. This process enables the models to capture intricate patterns and generate high-fidelity samples. The core principles of diffusion models involve:

- **Systematic Degradation and Restoration:** Introducing Gaussian noise to degrade the data and learning to reverse this process.
- **Modeling Complex Data Distributions:** Iteratively transforming simple distributions into target distributions.
- **Neural Networks as Denoising Functions:** Utilizing neural networks to learn relationships within data for accurate generation.

Collaborative Distributed Diffusion-Based AIGC Framework

To address the computational and energy constraints associated with deploying diffusion models on resource-limited devices, [21] proposed a collaborative distributed diffusion-based AIGC framework. This framework optimizes the use of edge computing resources and facilitates efficient execution of AIGC tasks through device collaboration in wireless networks. Its key components include:

- **Central and Edge Inference:** Combining central server capabilities with edge device resources to balance computational loads.
- **Shared Denoising Steps:** Performing initial denoising steps on a central server or an edge device, and then transmitting intermediate results to other devices for task-specific processing.
- **Privacy and Resource Optimization:** Enhancing privacy by allowing local execution of tasks and optimizing resource usage through collaborative offloading techniques.

Integration with Networking Systems

The integration of distributed diffusion models with networking systems involves various architectural setups to facilitate efficient AIGC task execution:

- **Edge-to-Multiple Devices Architecture:** An edge server performs shared denoising steps for multiple devices with similar task requirements, reducing latency and optimizing resource allocation.
- **Device-to-Device Collaboration:** Two devices directly collaborate by sharing intermediate results and performing task-specific denoising steps, enhancing energy efficiency and privacy.
- **Clustered Device Collaboration:** Forming clusters of devices, either with or without edge server assistance, to collaboratively execute AIGC tasks. This approach is adaptable, scalable, and resource efficient.

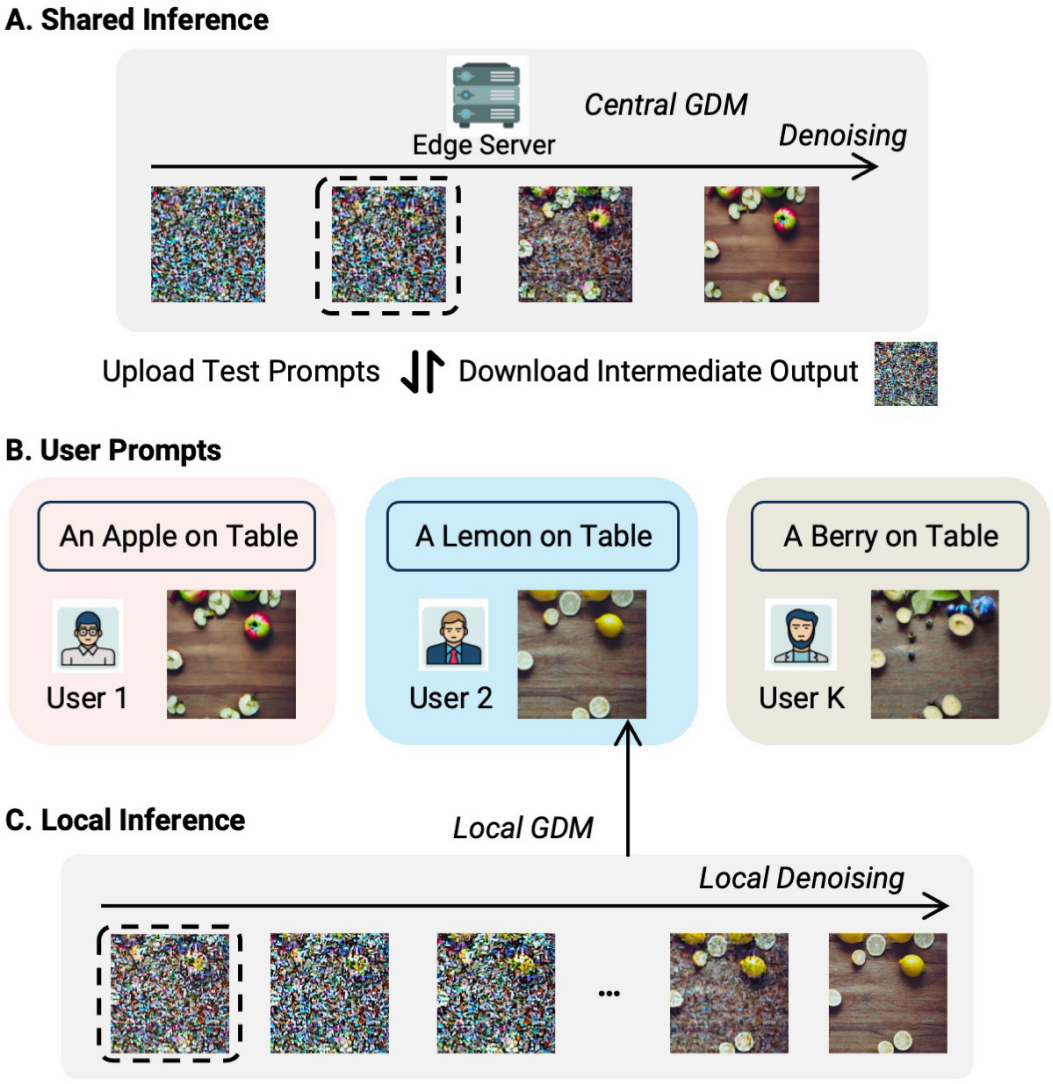


Figure 2: Distributed Diffusion Model in Networking Systems

Figure 2 depicts a scenario where a server serves multiple user terminal devices. The deployment method for distributed generative diffusion model-based text-to-image AIGC services includes several components. Specifically, Part A illustrates the shared inference mechanism, where users with semantically similar prompts collaborate with the server. Part B shows the diverse text-to-image requirements of each of the K concurrent users. Finally, Part C presents the generated images corresponding to each user's specific prompts.

Practical Implementation and Benefits

Implementing this framework has demonstrated the feasibility of executing complex AIGC tasks on mobile devices, such as running the stable diffusion model on a smartphone without an internet connection. The approach ensures high-quality content generation while addressing challenges related to computational limits and privacy concerns.

Future Directions

Future research in this area should focus on designing incentive mechanisms for efficient resource sharing, joint optimization of diffusion models with channel coding for adaptive modulation, and secure schemes to ensure privacy and data integrity during distributed computing processes.

In conclusion, the collaborative distributed diffusion-based AIGC framework offers a promising solution to the challenges of deploying AI-generated content services on resource-constrained devices. This approach paves the way for scalable, efficient, and secure AIGC services across various devices by leveraging the strengths of wireless networking systems and edge computing.

5. GenAI in NETWORK MANAGEMENT

Overview of Network Management Use Cases

GenAI is reshaping network management across several key areas, making operations more proactive and adaptive to meet the demands of modern network environments. GenAI enhances automation and optimization in network operations and design, allowing real-time adjustments that improve overall network performance. GenAI enables the automation of manual network troubleshooting processes, which today is both a time-consuming and an error-prone process. This technology extends its benefits to network planning by using predictive models to simulate different scenarios, helping planners make informed decisions about capacity and deployment strategies.

Additionally, GenAI is instrumental in refining customer support services, utilizing NLP and ML to provide faster and more accurate responses to customer inquiries. Intent-based networking is another area where GenAI plays a pivotal role, translating business objectives into actionable network policies that ensure tasks meet business goals. Furthermore, Gen-AI-as-a-Service is emerging as a transformative model, offering AI-driven capabilities to organizations without needing in-house AI expertise, thereby democratizing access to advanced network management tools. Together, these applications of GenAI are driving significant improvements and advancements in network management, enhancing efficiency, reliability, and customer satisfaction.

Customer Support — Ticket-based Troubleshooting

In the rapidly evolving telecommunications industry, quick and accurate troubleshooting is crucial for maintaining service quality and customer satisfaction. The complexity of modern telecommunication systems poses significant challenges in fault resolution, often resulting in delays and service interruptions. Traditional manual processes for fault resolution are slow and inefficient, leading to service interruptions and reduced customer satisfaction.

AI, particularly NLP (which includes Natural Language Understanding (NLU)), automates the analysis of historical Trouble Reports (TRs), speeding up the identification of solutions. Text-ranking techniques quickly sort through past TR data to find relevant solutions.

Language Model (LM)-based TR Duplicate Identification (TRDI) is an advanced system designed to efficiently and accurately resolve telecommunications issues. The TRDI system leverages AI to refine and prioritize solutions through advanced pre-processing, retrieval, and re-ranking stages.

AI models significantly enhance the accuracy and speed of troubleshooting, reducing response times to milliseconds. The AI-driven system consistently ranks relevant solutions, improving resolution times and maintaining high accuracy.

Case Study: Leveraging AI for Enhanced Ticket-Based Troubleshooting

Problem Statement

The manual troubleshooting process in telecommunications is labor intensive and time consuming. Engineers generate TRs for issues they cannot resolve on site, resulting in extended resolution times. The industry requires innovative solutions to streamline fault detection and resolution to maintain high service standards.

Solution: LM-Based TRDI

To address these challenges, the telecommunications industry has turned to AI, specifically NLP, to automate and enhance the TR process. NLP offers significant potential in automating parts of the TR process. By analyzing historical TR data, NLP can facilitate quicker and more accurate problem-solving approaches. The core of this approach is the text-ranking problem, which is shown in Figure 3. This involves generating an ordered list of texts from a corpus in response to a query. In telecom troubleshooting, this translates to sorting through historical TR data to identify relevant solutions quickly. The process involves three key elements: the query (problem statement), the corpus (collection of past TR solutions), and the output (ranked list of relevant solutions).

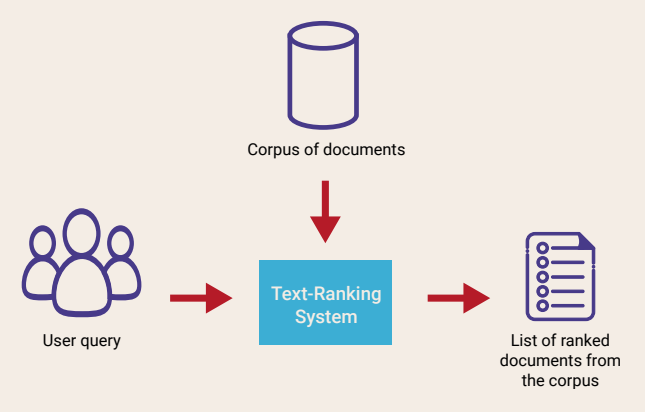


Figure 3: Diagram of the Text-Ranking Problem

There are several techniques for text ranking (Figure 4). They can be grouped into keyword, representation-based, and LLM-based methods. One of the popular language models used is BERT [51], which was developed by Google. BERT represents a paradigm shift in how machines understand human language. Unlike previous models that processed text

in one direction (either left to right or right to left), BERT simultaneously interprets text in both directions. This bidirectional approach allows for a more nuanced understanding of context and meaning in language.

Keyword and Representation Based	**Exact Matching Techniques:** TF-IDF or BM25
	Neural IR: representation- and interaction-based architectures
Language Model Based	**Standard BERT model:** BERT, ELECTRA, RoBERTa
	Multi-Stage Architectures: BERT in the second stage

Figure 4: Techniques for Text Ranking

The cornerstone of BERT is the transformer architecture, introduced in the paper "Attention is All You Need" [44], which revolutionized NLP by focusing on "attention mechanisms." These mechanisms enable the model to weigh the significance of different words in a sentence, thereby capturing the context more effectively than traditional methods. Transformers are known for their ability to handle long-range dependencies in text, making them particularly effective for complex language tasks. Transformers, in turn, provide the structural and functional foundation upon which LLM are built, enabling these models to achieve remarkable levels of language understanding and generation.

The LM-based TRDI system is an advanced solution tailored to efficiently and accurately address telecommunications issues. This system is based on advanced transformer technology, which marks a significant advancement in the field of NLP.

Process Flow of TRDI (see Figure 5):

- **Pre-processing Stage**: In this initial phase, TRs are refined through tokenization, abbreviation expansion, and normalization, preparing them for advanced analysis.
- **Initial Retrieval (IR) Stage**: This stage utilizes Sentence-BERT (a variation of BERT optimized for sentence-level tasks) to retrieve a list of documents that are preliminarily relevant to the problem query.
- **Re-Ranking (RR) Stage**: This stage incorporates monoBERT, another BERT variant, and refines the list from the IR stage. MonoBERT's advanced processing abilities ensure that the most pertinent solutions are accorded the highest priority.

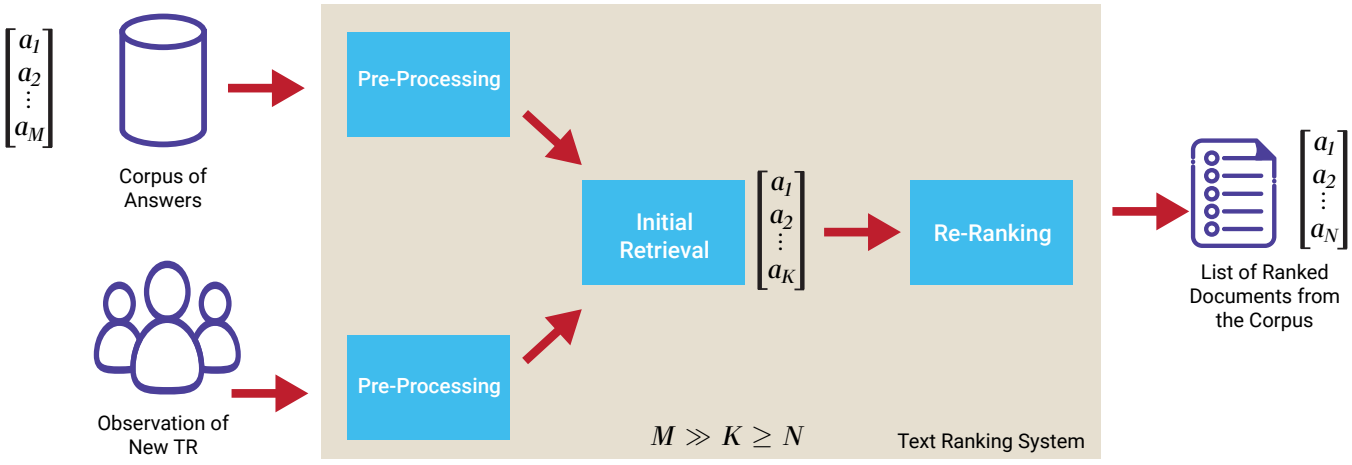


Figure 5: General Overview of the Architecture of TRDI

Incorporating BERT and transformer technology in TRDI allows a deeper understanding of telecommunication troubleshooting data. This advanced comprehension leads to more accurate and efficient problem solving, setting a new standard in automated troubleshooting processes. TRDI represents the culmination of the recent advancements in NLP and ML, specifically tailored to meet the complex demands of the telecommunications sector.

Implementation and Training of TRDI

TRDI utilizes Ericsson's comprehensive troubleshooting data, including 4G and 5G network records. Each TR is dissected into sections (Heading, Observation, Answer, Faulty Area) to form the query for the model, as found in Figure 6.

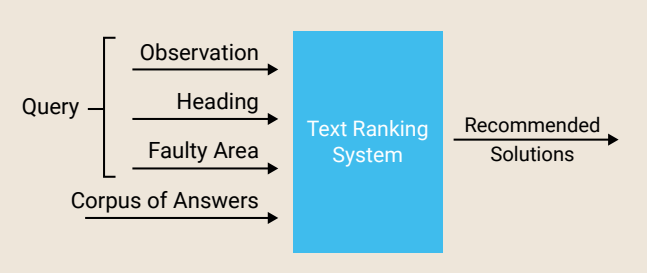


Figure 6: Sections of the TRs that Form the Query and the Corpus of the Document in TRDI

The model undergoes a two-stage training process:

- **IR Stage**: A pre-trained model from the Microsoft Machine Reading Comprehension (MS MARCO) dataset is fine tuned for telecom-specific data.
- **RR Stage**: Similarly, a pre-trained model is fine tuned for the ranking task.

During the inference phase, TRDI processes new TRs, employing the trained IR and RR stages to output a list of ranked solutions.

Evaluation and Impact of TRDI

The model was trained using the dataset consisting of completed 4G and 5G radio network TRs, with testing performed on 15% of all data points. The correct document (answer) for each problem (query) in the dataset was known, allowing for verification of whether this document appeared in a high position in the resulting recommended list. Three metrics were used for evaluation: Recall@K, Mean Reciprocal Rank (MRR), and normalized Discounted Cumulative Gain (nDCG).

Below are performance examples at different stages:

Initial Retrieval Stage Results

A notable comparison involves evaluating the model's performance, designated as IR, against the widely used BM25 retrieval model, an exact matching method. Table 1 shows that Sentence-BERT significantly improved the results over BM25, enhancing Recall@1 by approximately 65%.

Table 1: Comparison of the Results of Sentence-BERT and BM25 in the IR Stage

Initial Retriever	BM25	Sentence-BERT
Recall @1	18.2%	30.2%
Recall @3	23.7%	43.1%
Recall @5	26.6%	49.0%
Recall @10	31.0%	58.2%
Recall @15	33.3%	64.0%

Re-Ranker Stage Results

A key experiment in the RR stage assesses improvement over the IR stage. After receiving a top-K candidate list from the IR stage, the RR stage re-ranks it to elevate the correct document to the top of the list. As shown in Table 2, the RR stage increases the MRR by 12% and nDCG by 9% and boosts Recall@K for smaller values of K.

Table 2: Results of the Experiment of Adding the RR Stage

	After the IR stage	After the RR stage
Recall @1	30.2%	36.6%
Recall @3	43.1%	48.5%
Recall @5	49.0%	53.5%
Recall @10	58.2%	59.8%
Recall @15	64.0%	64.0%
MRR	0.39	0.44
nDCG	0.44	0.48

Latency Results

Latency analysis ensures the system can recommend related TRs quickly. Latency, defined as the time taken to output a list of recommended answers for each query, is crucial. The described solution achieves a final list of ranked documents in just 578 milliseconds on average. In contrast, relying solely on the RR stage (without the IR stage) significantly increases latency, as the RR stage would need to process all documents instead of just the top K.

A comparison between model latency and accuracy shows minimal accuracy improvement for larger K values but a notable increase in latency. Thus, a top-15 document list balances accuracy and latency, as illustrated in Table 3.

Table 3: Accuracy vs. Latency of TRDI

Candidate list length	K = 15	K=50	K = 100
milliseconds/query	578	1940	3820
MRR	0.44	0.455	0.46

Consistency in Ranking Similar TRs

The model's ability to produce consistent ranking lists for similar TRs was also evaluated, even when the TR content is expressed differently. Different customers often report similar underlying problems in varied words and lengths. Analysis shows that the correct answer for similar TRs is ranked similarly in 70% of cases.

Conclusion

Unlike manual TR solving, the TRDI model quickly scans and analyzes thousands of previous answers and suggests the most relevant solved tickets to the new problem. This can provide human experts with a better opportunity to detect the root cause of the issue much faster. By automating the analysis and ranking of TRs, the TRDI system significantly reduces resolution times. It improves accuracy, setting a new standard for efficiency and customer satisfaction in the telecom industry.

This case study illustrates how utilizing transformer-based techniques in real-world industry challenges can simplify complex processes and deliver faster and more reliable solutions.

Case Study: Reinforcement Learning with LLM Interaction for GenAI in Network Management

The advent of GenAI models has significantly enhanced the capabilities of network management systems. One particularly promising approach involves using LLMs to simulate user interactions, thereby providing a means to model and optimize the subjective Quality of Experience (QoE). This technique enables evaluating QoE by leveraging Reinforcement Learning (RL) to dynamically adapt network management strategies. This ensures high user satisfaction while maintaining efficient resource utilization.

Leveraging LLMs to Simulate User QoE

LLMs, such as GPT-4, are proficient in understanding and generating human-readable text, making them ideal candidates for simulating user interactions. Integrating LLMs into network management systems makes it possible to generate subjective QoE feedback without the need for continuous human input. This is particularly beneficial for training models in dynamic environments where real-time feedback is crucial. The LLMs can be programmed to represent various user personalities and preferences, providing diverse and realistic QoE evaluations that reflect the complexity of real-world scenarios.

Interaction Between LLMs and RL in Network Management

RL is well-suited for managing dynamic and complex systems like networks. In this context, RL algorithms are employed to optimize resource allocation and improve QoE based on feedback received from LLMs. The process involves the RL agent making decisions that affect the network's performance, receiving feedback from the LLMs, and adjusting its strategy to maximize cumulative rewards, which, in this case, correspond to the aggregated QoE scores from simulated users.

The interaction between LLMs and RL in network management is centered on using the LLMs to provide immediate and context-aware feedback. This feedback loop allows the RL agent to continuously learn and adapt its policies based on the simulated user experiences. By incorporating a diverse set of user profiles and preferences into the LLMs, the system can account for a wide range of subjective QoE metrics, leading to more robust and user-centric network management strategies. For example, the AIGC-as-a-service concept, as illustrated in Figure 7, leverages the power of edge servers to deploy GenAI models and enable AIGC services.

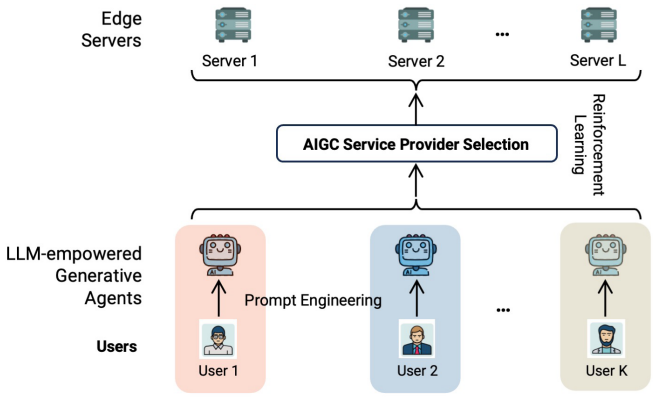


Figure 7: System Model for the AIGC Service Provider Selection Problem [23]

The selection of an AIGC service provider is critical because user preferences and the diversity of image styles generated by different GenAI models, influenced by their unique training datasets, play a significant role. The goal is to maximize the aggregated QoE for users. To gather subjective feedback on image quality as part of the RL network management algorithm training, LLMs can be employed to simulate user perception.

Training Network Management Models with Simulated QoE

Training network management models using LLM-simulated QoE involves several key steps. First, the LLMs are configured to represent different user profiles, each with unique preferences and expectations. These simulated users interact with the network, providing QoE feedback based on their experiences. The RL agent uses this feedback to adjust network parameters, such as bandwidth allocation, latency management, and error correction protocols. Over time, the RL agent learns the optimal strategies that maximize overall user satisfaction, as reflected by the cumulative QoE scores [24].

Using LLMs to simulate user QoE offers several advantages. It reduces the reliance on real-time human feedback, which can be costly and logistically challenging. Additionally, it enables the modeling of diverse user behaviors and preferences, ensuring that the network management strategies are well-rounded and inclusive. This approach also facilitates faster and more efficient training of RL models because the LLMs can provide immediate feedback without delays, accelerating the learning process.

Specifically, Figure 8 illustrates the process of setting initial prompts, which involves both general setup and generative agent-specific settings. Part A shows how to familiarize the generative agents with the big five personality traits during the initial setup. In contrast, Part B demonstrates how varying the number of shared denoising steps can lead to stylistic differences in the final generated images. Part C showcases a user personality trait configuration alongside the corresponding evaluation scores for the generated images. It is important to note that the prompt engineer, a component of the system, automates this process. It eliminates the need for human intervention and simply retrieves the target user's personality traits for the initial prompt setting.

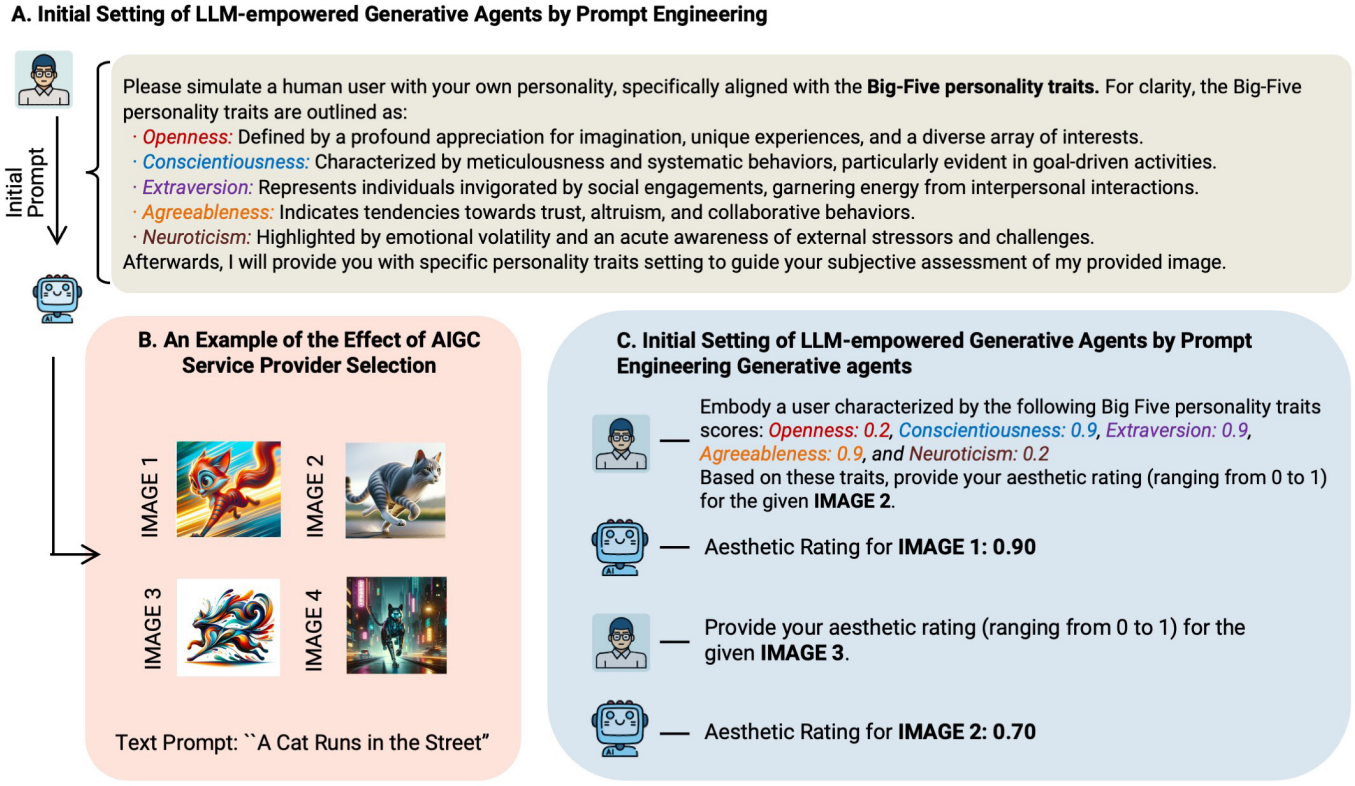


Figure 8: Prompts for LLM-empowered Generative Agent Settings [23]

Conclusion

Incorporating LLMs into the RL framework for network management represents a significant advancement in the field. By simulating user QoE through LLMs, network management systems can be trained to optimize performance and enhance user satisfaction effectively. This approach not only addresses the challenges of obtaining real-time feedback but also ensures that the management strategies are adaptable and responsive to the diverse needs of users, ultimately leading to more efficient and user-centric network operations.

Neuro-Symbolic AI for Enhanced Reasoning and Decision-Making

A novel approach for integrating GenAI into telecommunications involves combining neural networks (such as transformers) with symbolic AI (such as knowledge graphs). This fusion enhances reasoning, decision-making, and explainability in AI systems, addressing key challenges faced by the telecom industry. By combining these two approaches, a neuro-symbolic AI system can learn from data (like neural AI) while also reasoning about abstract concepts and relationships (like symbolic AI). This integration provides more robust, flexible, and interpretable AI systems. A cognitive assistant [26] shows an example of how neuro-symbolic logic can be used in a telecom environment.

Any network is not only inherently heterogeneous but exists in a potentially rapidly changing environment that emits a massive amount of data. The big data analogy to Parkinson's Law is: "*The amount of data you collect will expand to consume your ability to store and process it.*"

The combination of transformers and knowledge graphs is uniquely positioned to help extract value from a big data environment. The following are examples of how such a neuro-symbolic system can help automate network management.

Case Study: Compliance with Upcoming Regulations

Several upcoming regulations are focused on governing AI technologies. The most advanced is the EU's AI Act [27], which is already law. These regulations reflect the global effort to balance innovation with the need for responsible and ethical AI development. Some other efforts include:

- The U.S. is in the early stages of AI regulation, with federal efforts led by Executive Order 14110, issued in October 2023, emphasizing safe and responsible AI development. While comprehensive federal laws have yet to be enacted, several states have introduced their own AI-related regulations, and NIST has published a framework for managing AI risks.
- The UK has adopted a "pro-innovation approach" to AI regulation. Instead of a single overarching regulation, the UK government has proposed an activity-based approach, allowing individual regulatory bodies to govern AI within their respective domains.
- Canada has introduced the Artificial Intelligence and Data Act (AIDA), which aims to regulate high-impact AI systems and ensure they are developed and used responsibly.
- Japan has established the AI Strategy 2022, which outlines the country's approach to AI governance, focusing on promoting innovation while ensuring safety and ethical considerations.
- China has been proactive in regulating AI, ensuring that AI development aligns with national interests. The country has introduced several guidelines and standards for AI ethics and governance.

Problem Statement

The EU's AI Act [27] emphasizes the importance of transparency and explainability [37] in AI systems, especially for high-risk applications. A key priority of the EU's AI Act is transparency, which is considered essential for creating public trust in AI systems and ensuring their responsible deployment. The Act requires AI systems to be sufficiently transparent so that their decisions are explainable to users and affected persons. A cognitive assistant [26] is one way to address transparency.

Solution: Neuro-Symbolic AI

Neuro-symbolic AI can help create systems that comply with the EU's AI Act in several ways. First, the neural component excels at ingesting vast amounts of data and information, keeping the neuro-symbolic system current. The symbolic component (e.g., a knowledge graph or ontology) processes output from the neural component, performing such functions as validating facts and fine tuning context. It also can use rules and logical reasoning to describe why a decision was made, addressing the Act's requirement for explainability. The explainable nature of neuro-symbolic AI also facilitates human oversight, an essential requirement in the Act.

Neuro-symbolic AI can incorporate and encode business rules, policies, and domain knowledge into machine-understandable data. The symbolic reasoning component can enforce regulations related to fundamental rights and ethical considerations by examining proposed actions and mathematically proving whether the proposed actions comply with applicable regulations. These examples of structured knowledge can also help generate more interpretable logs and documentation of decision processes. This enables better documentation of the AI system's knowledge base and decision-making process, supporting compliance with the Act's documentation requirements.

The explainability of neuro-symbolic AI could improve "AI literacy" among users and affected persons, an essential aspect of the EU's AI Act. This enables non-technical people to better understand AI-based decisions.

Neuro-symbolic AI can also significantly mitigate some types of bias in AI systems, particularly in the context of the EU's AI Act using symbolic reasoning. Using knowledge graphs, the system can cross-reference information and detect inconsistencies or biases in the data. Symbolic reasoning allows for the implementation of rules that can adjust or correct biased outputs from the neural network. Techniques like Reduced Implication-Bias Logic Loss can transform biased loss functions into more balanced ones, improving the model's fairness. By leveraging both neural and symbolic methods, the system can better understand the context and semantics, reducing the likelihood of biased interpretations.

In addition, neuro-symbolic AI can incorporate rule-based logic and ethical guidelines directly into the AI system. Ethical guidelines can be encoded as constraints within the system. These constraints ensure that the AI's actions align with ethical principles, such as fairness, transparency, and accountability. Neuro-symbolic AI systems can incorporate explicit rules that govern decision-making processes. By integrating symbolic reasoning, the system can apply logical rules to interpret data and make decisions. This ensures that the AI follows predefined guidelines and behaves predictably. This integration helps ensure that decisions are made according to predefined rules that can be designed to minimize bias.

The symbolic component of neuro-symbolic AI allows continuous auditing and updating of the system's rules and logic. This capability ensures the AI system can adapt to new ethical standards and guidelines, continuously improving its fairness and reducing biases over time.

Finally, neuro-symbolic AI excels in handling ambiguous situations and contextual nuances by combining pattern recognition with logical reasoning. This dual capability allows the AI to consider a broader range of factors and nuances, reducing the likelihood of biased decisions based solely on data patterns.

Case Study: Translating Policies between Different Constituencies

Problem Statement

Different user constituencies use different concepts and terminologies. When business rules are expressed in natural language, their translation is difficult due to the inherent ambiguity in natural language. In addition, business policies may not fully capture the technical requirements needed for network management. Translating these policies requires a deep understanding of both the business objectives and the technical details.

Solution: Neuro-Symbolic Logic

The Policy Continuum [28] in ETSI experiential networked intelligence (ENI) differentiates the needs of various constituencies in defining and expressing policy. These constituencies consist of users with similar business and/or technical needs and terminology, such as business users versus application developers [29].

For example, consider the business rule: "Allocate more bandwidth to customer support applications during peak hours while ensuring executive video conferences always have priority." This requires understanding time-based conditions, application priorities, and network resource management. A neuro-symbolic approach can combine symbolic reasoning with learned network usage patterns to solve this problem.

The reverse is also true. For example, consider the network administration rule: "Configure border gateway protocol (BGP) to prefer paths with the highest local preference, then shortest Autonomous System (AS) path, and apply route maps to prepend AS numbers for specific prefixes to influence inbound traffic." This rule involves complex routing concepts and traffic engineering that are challenging to explain in business terms without oversimplification, which risks losing important details.

In the following examples, neuro-symbolic AI systems leverage knowledge to improve translation accuracy between business and technical domains in several important ways:

- **Contextual Understanding**: Better understanding of the context and relationships between different concepts.
- **Structured Knowledge**: Knowledge graphs represent domain knowledge, relationships, and concepts from both business and technical perspectives. This complements the capabilities of neural networks, enabling more sophisticated translations that account for complex business rules and technical specifications.
- **Handling Incomplete Information**: Managing incomplete information without making false assumptions or hallucinating. This is particularly useful when translating between domains where information may be partial or ambiguous.
- **Improved Interpretability**: Providing more transparent and explainable results, which builds trust in business and technical stakeholders.
- **Dynamic Knowledge Integration**: Continuously updating the system with new information, enabling the system to adapt to changing business rules or technical specifications.
- **Enhanced Generalization**: Combining learned patterns from neural networks with structured knowledge from knowledge graphs, allowing improved generalization to new scenarios.
- **Guided Learning and Inference**: Knowledge graphs can guide the learning and inference processes of neural components, leading to more accurate and contextually appropriate answers.

By leveraging these capabilities, neuro-symbolic AI systems provide more accurate, context-aware, and interpretable translations between business and technical concepts and terminologies compared to using either neural networks or knowledge graphs alone. The cognitive assistant [26] specifically addresses this use case in ETSI ENI [15].

Network Operations Use Cases

Several critical network operations can benefit from the power of neuro-symbolic AI. The common denominator in these is the combination of fast ingestion of data and pattern recognition of neural networks with the logical reasoning and explainability features of symbolic AI.

Anomaly Detection

Transformers can be trained on large datasets of network traffic and activity logs to learn normal patterns. They can then detect anomalies or deviations from these patterns, which could indicate security threats, network issues, or other problems that require attention. A knowledge graph is then queried to understand the normal relationships and behaviors associated with the involved entities. This helps determine whether the detected anomaly is genuine or can be explained by known patterns. When an anomaly is detected, the domain-specific rules encoded in a knowledge graph can be applied to verify if the detected behavior violates any known constraints or business rules. This logical reasoning helps in distinguishing between true anomalies and false positives.

For example, suppose a transformer model detects an unusual spike in data transfer from a specific IP address. This spike is flagged as a potential anomaly. The knowledge graph is queried to retrieve information about the IP address, its usual data transfer patterns, and its relationships with other network entities. It then applies domain-specific rules encoded in the knowledge graph to check if the detected spike violates any known constraints (e.g., data transfer limits for that IP address). It also searches for past data transfer spikes from the same IP address. Finally, the provenance of the data points involved in the anomaly is traced to ensure their authenticity and relevance. The combined reasoning from the neural network and the logical rules in the knowledge graph confirms whether the detected spike is a true anomaly. The system provides an explainable decision, detailing the reasoning process and the contextual information used for verification.

This also applies to similar subjects, such as malware analysis. Neuro-symbolic AI uses the neural network to analyze large datasets to identify patterns and anomalies, which the symbolic component can analyze and provide interpretations and reasoning about the behavior of malware. Similarly, neuro-symbolic AI can enhance vulnerability assessment and threat intelligence activities by understanding the underlying rules and configurations of the network (symbolic reasoning) and predicting potential vulnerabilities based on historical data (neural networks). The symbolic component can enforce security policies and rules, while the neural network can learn from past incidents to make informed decisions. This ensures a more comprehensive assessment of network security.

Predictive Maintenance

Predictive maintenance applications are becoming increasingly complex, involving interactions between many components. A neuro-symbolic system solves two problems in parallel: anomaly detection and explanation of the anomaly. Both systems run online and in parallel. The autoencoder signals an alarm, which is hard for humans to understand because it results from a non-linear combination of sensor data. The rule that triggers that example describes the relationship between the input features and the reconstruction error. The rule explains the failure signal by indicating which sensors contribute to the alarm and allowing the identification of the component involved in the failure. The system can present global explanations for the black box model and local explanations for why the black box model predicts a failure.

By analyzing historical network data, transformers can identify patterns and relationships that predict potential failures or performance degradation of network components, reducing downtime and costs. The knowledge graph can then be queried to retrieve contextual information about the component, apply relevant operational rules and constraints, check historical failure patterns, examine relationships with other network elements, and look at maintenance schedules and past issues. The knowledge graph is used to verify if the predicted anomaly is consistent with known patterns, rules, and expert knowledge encoded in the knowledge graph. If verified, the system can provide a detailed explanation of why the anomaly is considered genuine, including the contextual factors, rules applied, and historical patterns that support the conclusion. If the anomaly is not verified, the system can explain why it is likely a false positive, potentially saving time and resources on unnecessary interventions.

Root Cause Analysis

The neuro-symbolic AI system provides an explainable decision, detailing the reasoning process and the contextual information used for verification during the process of resolving the cause of the problem and verifying that the solution removes the problem. Neuro-symbolic AI can enhance causal inference by combining the data-driven insights from neural networks with the logical reasoning capabilities of symbolic AI.

Transformers can analyze various data sources (logs, metrics, configurations, etc.) to identify the root cause more quickly and accurately than traditional methods. Symbolic AI can interpret and reason about the network's rules, configurations, and policies. This allows for a deeper understanding of the relationships between different network components and their behaviors. The neuro-symbolic system can provide more transparent and explainable results by grounding the transformer's predictions in the symbolic representation of a knowledge graph. This explainability is crucial for network

administrators to understand and trust the reasoning behind the anomaly detection and verification process. The workflow is similar to the above two examples, except after the provenance check, the knowledge graph is updated with new information as it becomes available, ensuring that the system's understanding of normal and anomalous behaviors remains current.

Work on automated troubleshooting is evolving. The goal is to suggest potential fixes based on the identified root causes and the network's rules and configurations. Neuro-symbolic AI systems can continuously learn from new data and adapt their troubleshooting strategies accordingly. This ensures that the system remains effective even as the network evolves.

Conclusions

Neuro-symbolic AI is a promising area that presents a path toward more interpretable AI. The combination of using a transformer for the neural portion and a knowledge graph for the symbolic portion provides a more powerful, flexible, and interpretable AI system that can handle complex reasoning tasks while maintaining the ability to learn from diverse and noisy data sources. The cognitive assistant [26] is one way of achieving these goals. This is because neuro-symbolic AI augments contextual meaning, enabling new relationships to be deduced and context-specific information to be attached to entities. These capabilities allow for more accurate execution of complex reasoning. As highlighted by the network operations use cases above, neuro-symbolic AI helps implement cognitive networks.

6. GenAI Across DOMAINS

Overview of Cross-Domain Use Cases

GenAI significantly enhances telecommunications through its diverse applications across key domains, each contributing unique benefits and innovations. In semantic applications, it aids in object detection, image classification, and data fusion, improving data interpretation for more intelligent network management. It also improves natural language understanding. As a step toward collective intelligence, GenAI fosters a collaborative environment where AI and human expertise merge, enhancing decision-making and operational efficiency. It supports network slicing by creating tailored virtual networks and optimizing performance for varied service requirements.

In specific sectors like vehicular technology and the Internet of Things (IoT), GenAI enables sophisticated functionalities such as better traffic management and smart device automation. It also plays a crucial role in developing immersive VR and AR experiences, adding depth and realism to virtual interactions. For network security, AI-driven predictive models enhance threat detection and response. In the telecommunications industry, LLMs are used for trouble report classification, patent searches, network optimization, predictive maintenance, and virtual customer support assistants, showcasing their diverse and evolving applications.

Network Slicing

GenAI is poised to significantly enhance 5G network slicing by improving network management, optimizing resource allocation and network performance, and personalizing user experiences across various domains for diverse applications.

GenAI plays a crucial role in network slicing, where it helps create and manage multiple virtual networks on one or more physical infrastructures. For example, in a smart city, different IoT devices (e.g., traffic sensors, surveillance cameras, and public Wi-Fi hotspots) are connected. Each of these devices has different requirements for bandwidth, latency, and reliability. GenAI can analyze real-time data from these devices, including their usage patterns, data traffic, and performance metrics. Based on this analysis, GenAI can determine which physical infrastructure (e.g., specific cell towers or data centers) is best suited for virtualizing network slices to meet the specific needs of each type of device.

In general, each slice can be optimized for specific types of traffic, applications, or services, thus enabling more efficient resource utilization and better overall performance. Through predictive analytics and real-time decision-making capabilities, GenAI effectively manages network traffic. It dynamically allocates bandwidth and prioritizes network resources based on current demand and predicted future load, thus preventing congestion and ensuring smooth service delivery. GenAI is context aware and can predict how changing context affects resource needs in different slices. Advanced applications could also use behavioral analytics to anticipate users' future needs.

GenAI can analyze user behaviors, network conditions, and content preferences to optimize content delivery. For example, predictive content caching can anticipate user demand for content and pre-cache it at the network edge, reducing latency for high-demand events like sports broadcasts or live streaming.

GenAI can optimize streaming quality based on network congestion and user device capabilities, ensuring uninterrupted viewing experiences no matter what type of network is being used. It can also adjust content complexity in real time for interactive environments like online gaming or VR, providing an immersive experience without lag. For example, if HD content is either unavailable or will not deliver a good experience, it can suggest alternatives, including adaptive bitrate streaming, more efficient encoding, and frame rate and resolution scaling.

Datasets for Developing Telecom-Specific Language Models

In the telecommunications domain, there are specific technical terminologies, named entities, and their relationships, such as an eNB being a type of base station and User Equipment (UE) referring to a mobile device. Terms like 4G and Long-Term Evolution (LTE) are synonymous, representing a particular technology. The goal is to develop a language model that comprehends sentences with these complex telecommunication terms. Effective training of such a telecom-specific model requires a substantial and domain-specific dataset.

For Ericsson, telecom data was sourced from 3GPP specifications and additional mobile network information from various websites. This was supplemented with Ericsson's Customer Product Information (CPI) and internal community Question-and-Answer (Q&A) data. A specialized benchmark dataset called Telecom Question Answering Dataset (TeleQuAD) was created to evaluate the quality of ML models for Q&A in the telecommunication domain. Benchmark datasets are crucial for tracking model progress, but domain-specific datasets are necessary because models performing well in one domain may not generalize to another. TeleQuAD was developed using a web-based annotation tool, selecting a few hundred high-quality documents from 3GPP specifications and relevant web articles. Annotators formulated questions from presented documents, marking text spans as answers. This process ensured the dataset was representative of the telecom domain and compatible with existing models. As shown in Figure 9, multiple Q&A pairs were created based on the presented paragraph, in which the answer for every question was highlighted in context.

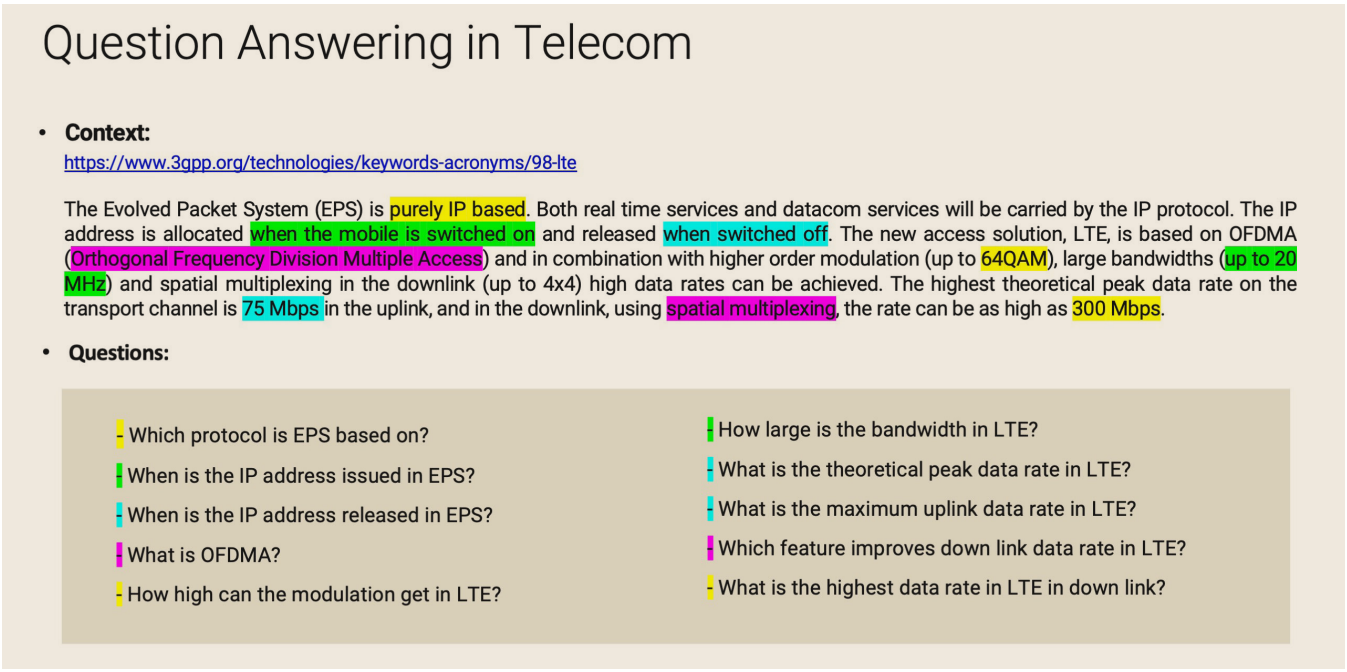


Figure 9: Examples of Annotated Questions and Answers in TeleQuAD

Case Study: Adapting Language Models for Telecom Applications

This case study explores the adaptation and evaluation of language models for the telecommunication industry. The goal is to leverage advanced NLP and NLU techniques to enhance various telecom-specific tasks and demonstrate significant improvements in performance and efficiency. As LLM technology evolves, emerging applications include network optimization, where models analyze performance data to enhance efficiency, predictive maintenance prevents future infrastructure problems, and the deployment of virtual customer support assistants to improve overall customer interactions. This highlights the diverse and evolving contributions of language models in both established and emerging telecom applications (see Figure 10).

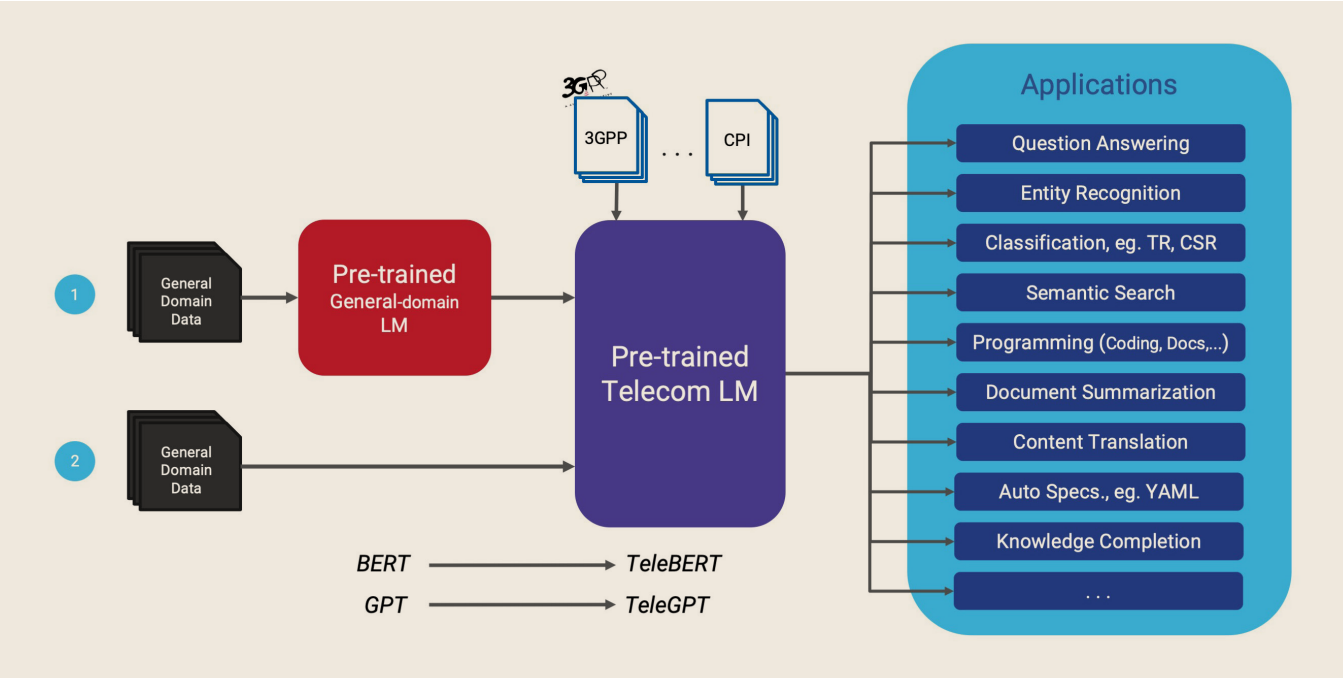


Figure 10: Overview of Building Language Models for the Telecom Domain (1) Adopting a General-domain Model; (2) Pre-training from Scratch

Methodology

Instead of pre-training from scratch, this case study leveraged transfer learning by using general-domain pre-trained models. These models contain valuable general knowledge and were further pre-trained on the telecom-specific domain data. This approach saved time and resources.

The following models were considered in this case study:

- **RoBERTa-Base:** A 12-layer transformer-based model.
- **DistilRoBERTa-Base:** A smaller, distilled version of RoBERTa-Base.
- **ELECTRA-Small:** A smaller model with efficient pre-training techniques.

These models were pre-trained on 21 GB of telecom-domain text data using the PyTorch interface of the Hugging Face Transformers library.

The models were fine-tuned for the Q&A task using the general-domain Stanford Question Answering Dataset (SQuAD) 2.0 dataset and the newly created TeleQuAD.

Figure 11 shows the domain difference between TeleQuAD and SQuAD. It shows word clouds of the most frequent terms in the annotated questions and answers of each dataset. The size of the words in the clouds corresponds to their frequency of occurrence. A greater presence of in-domain terms can be noticed in TeleQuAD compared to SQuAD.

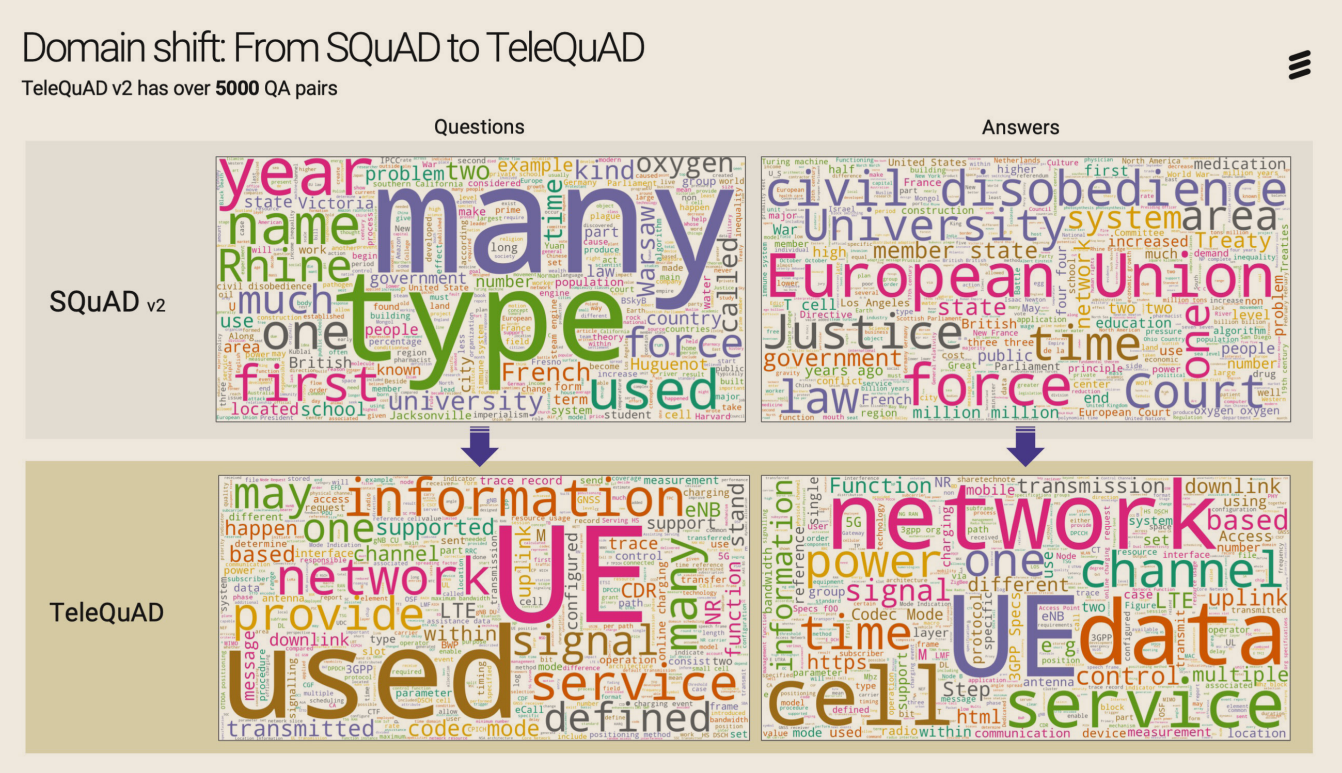


Figure 11: Word Clouds for SQuAD and TeleQuAD

Evaluation Results

Two metrics were used for evaluation:

- Exact match (EM), which measures how often the model's answer exactly matches the ground truth.
- F1 Score, which evaluates the overlap between the model's predicted answer and the ground truth.

As shown in Table 4 below, the telecom-domain adapted models showed significant improvements in Exact Match (EM) and F1 scores on the TeleQuAD test set compared to their general-domain baselines. For instance, TeleDistilRoBERTa's EM score increased by more than 10 points. These results demonstrated the benefits of domain adaptation. Besides, DistilRoBERTa-Base offered a balance of performance and efficiency, with faster inference times and a smaller memory footprint. ELECTRA-Small provided impressive performance for its size, making it suitable for deployment in resource-constrained environments.

Model	Developer	EM	F1	Parameters	GPU time
TeleRoBERTa	Ericsson	64.62	83.34	124M	35s
RoBERTa-Base	HF Community	56.38	77.90	124M	35s
TeleDistilRoBERTa	Ericsson	57.70	77.35	82M	21s
DistilRoBERTa	HF Community	47.60	69.77	82M	21s
TELECTRA	Ericsson	55.89	76.88	14M	24s
ELECTRA-Small	Google	48.14	70.76	14M	24s

Table 4: Evaluation Results for the TeleQuAD Test Set, Model Size, and Inference Time for the Respective Models

Conclusion

Language models adapted to the telecom domain show great potential for various applications within the industry. The creation of the TeleQuAD benchmark enables better adaptation and evaluation of models for telecom-specific tasks. Future work could extend these benchmarks to other tasks such as entity recognition, log analysis, and automated documentation. The adapted language models can contribute significantly to software development, infrastructure configuration, and more within the telecommunication industry.

Retrieval Augmented Generation-Based AI Chatbot for Telecom

LLMs are the evolution of deep learning models applied to NLP and NLU. As models became larger (e.g., using billions of parameters and trained on trillions of tokens), they started to show emerging capabilities such as improved reasoning and planning abilities, improved generalization and decision-making, and advanced interaction and personalization. Because of these new capabilities, LLMs have fundamentally changed how we interact with AI-based technology, providing powerful tools for a variety of tasks.

However, there are challenges when it comes to enterprise-level application using LLMs:

1. **Lack of Proprietary Knowledge:** LLMs cannot access or generate company-specific data unless specific provisions are taken to ensure that proprietary data is not leaked externally.
2. **Outdated Information:** Training datasets can become outdated, leading to decisions based on obsolete data. This is called model drift.
3. **Hallucinations:** LLMs may produce partially correct or entirely fabricated information, even when it contains the knowledge needed to answer correctly.

RAG addresses these issues by integrating LLMs with live enterprise data sources. This approach enhances the accuracy and currency of generated information.

The following explains the components of a RAG pipeline for AI chatbots and presents a real-world use case for telecommunication: a RAG-based AI chatbot for O-RAN. For more information, please see the NVIDIA technical brief on AI chatbot with RAG [19].

RAG consists of two processes:

- Ingestion of documents from document repositories, databases, or Application Programming Interfaces (APIs) that are all outside of the foundation model's knowledge.
- Retrieval of relevant document data and generation of responses during inference.

Figure 12 shows an accelerated RAG pipeline that can be built and deployed in the /NVIDIA/GenerativeAIExamples GitHub repo.

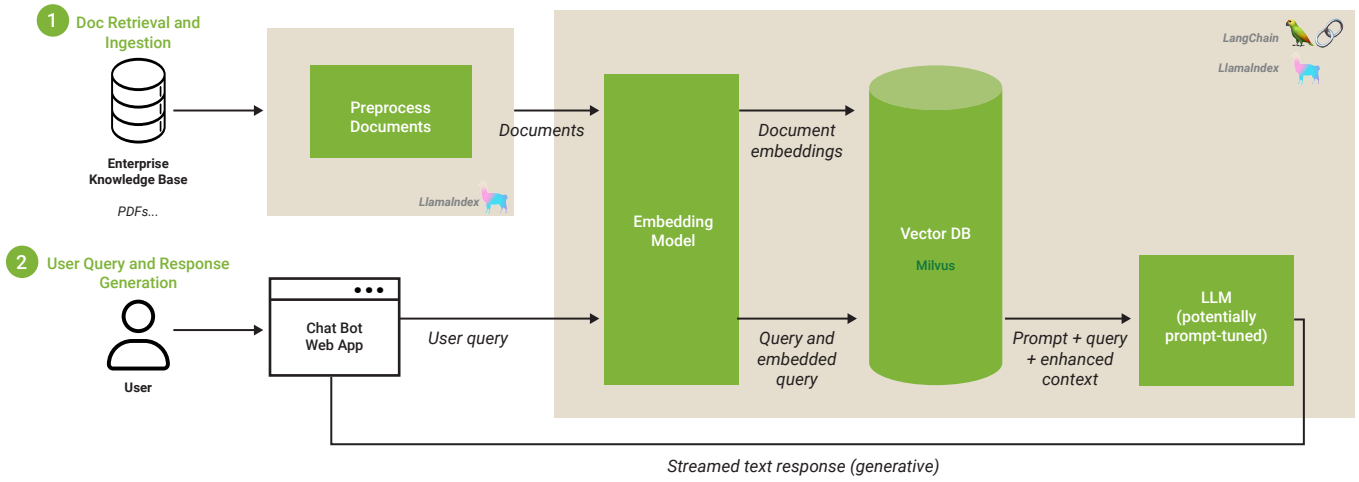


Figure 12: Overview of RAG Pipeline Components [19]

The use case demonstrates a RAG-based chatbot tailored for O-RAN specifications. This chatbot can ingest, comprehend, and provide detailed explanations, examples, and references related to O-RAN, ensuring stakeholders have the necessary knowledge to navigate its complexities. NVIDIA showcased the chatbot at the O-RAN F2F meeting in Athens [52], highlighting its capability to improve user trust and experience and reduce hallucinations by providing up-to-date information.

Cognitive Digital Twins

GenAI enhances digital twins in telecommunications by creating virtual replicas of physical devices, networks, or systems. These digital twins help model, simulate, and analyze operations in a virtual environment, providing insights critical for predictive maintenance, real-time optimizations, and strategic planning.

Cognitive digital twins (CDTs) [18] are digital twins with cognitive abilities. They incorporate AI functions to exploit implicit knowledge from existing systems. They recognize and adapt to changes in their environment, enabling autonomous decision-making through structured reasoning.

Key concepts in CDT decision-making include:

1. **Input Processing and Normalization:**
   - Inspired by the "Orient" phase of the Observe-Orient-Decide-Act (OODA) loop, it involves data processing to handle heterogeneity, cleansing, and normalization for consistency.
2. **Perception:**
   - Creates representations of data and behaviors relevant to the physical twin and its environment, normalizing and classifying data.
3. **Comprehension:**
   - Constructs meaning from data through pattern recognition, anomaly detection, and predictive modeling.
4. **Planning:**
   - Develops action sequences to achieve specific goals, assessing current and desired states, generating and evaluating actions, and monitoring execution.
5. **Action:**
   - Executes the chosen action plan, which could be recommendations or autonomous actions by the experiential networked intelligence (ENI) system.
6. **Denormalization and Output Processing:**
   - Converts internal formats to those recognized by external systems, ensuring compatibility and seamless integration.

The functional architecture involves reactive, deliberative, and reflective processes, enabling CDTs to handle immediate responses, complex planning, and continuous learning from experiences. Reflective learning further refines strategies by analyzing past actions and outcomes. As shown in Figure 14, through a closed control loop (or loops), a CDT can recognize complex and unpredicted behaviors and determine the best course of action. Situational awareness, where achieving or maintaining a set of goals is a primary constraint, can augment this.

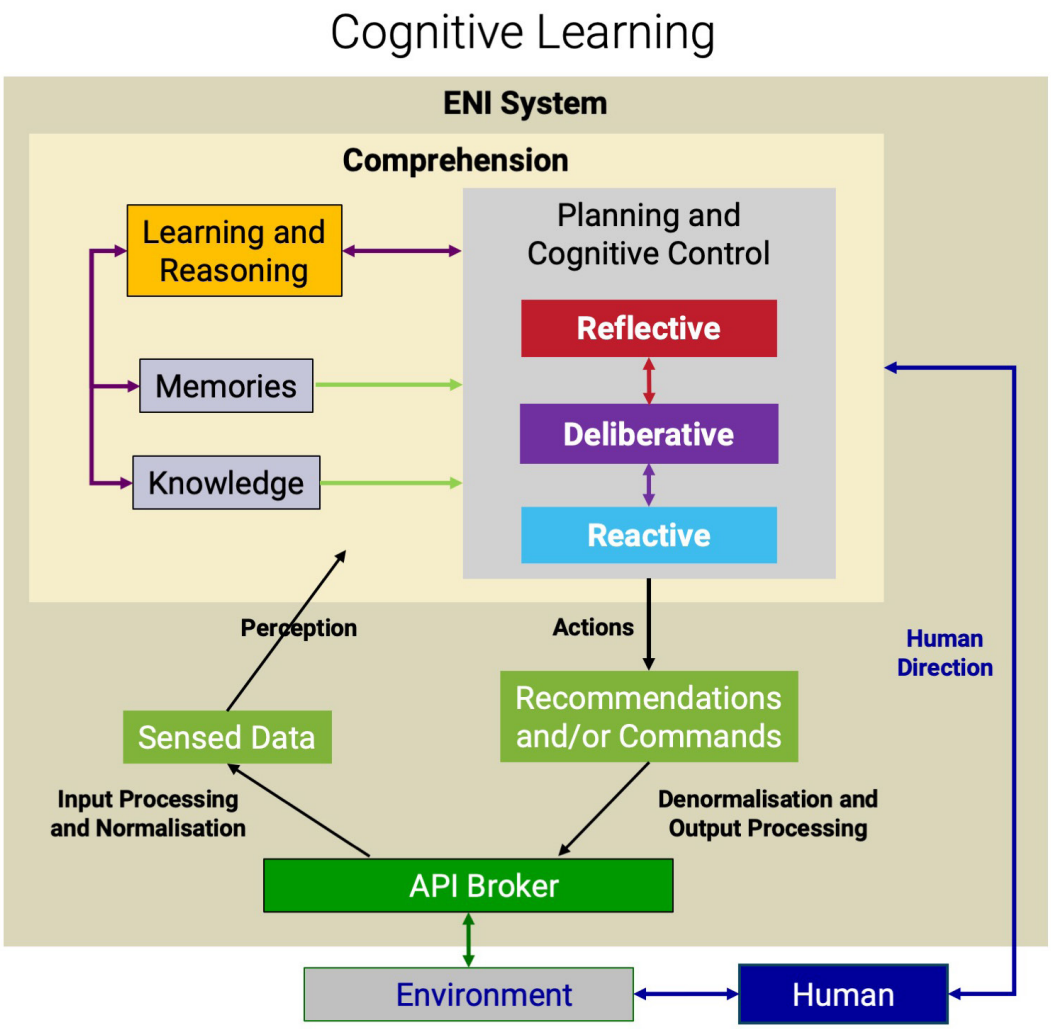


Figure 13: Simplified Functional Architecture of Cognitive Control in an ENI System

CDTs can integrate multiple digital twin systems for complex business environments, representing different lifecycle phases or customer interactions. Future work includes implementation choices (e.g., edge vs. cloud), federated learning for security, and transfer learning for broader ML applications. Use cases range from resource allocation in IoT networks and autonomous vehicle learning to smart manufacturing and improved caching in edge computing.

7. GAP ANALYSIS

The Gap Analysis is organized into four categories of common functionalities and challenges.

Technical and Infrastructure Challenges

This category encompasses the technical limitations and infrastructure needs that hinder the effective development and deployment of GenAI applications in telecom industry.

Limited Development of GenAI Foundation Models for RAN

GenAI foundation models are large, pre-trained models that can be fine-tuned for specific applications. They act as a template for developing different models that can perform a wide range of tasks, from NLP to image generation, by leveraging their extensive knowledge and adaptability. The RAN presents unique challenges and opportunities for GenAI. However, the telecom industry has not yet successfully developed and widely implemented GenAI foundation models specifically optimized for RAN.

RANs are highly dynamic and heterogeneous, with varying site-specific conditions depending on geography, user density, and mobility patterns. RAN operations also often require real-time or near-real-time processing to manage physical layer signal processing, MAC layer scheduling, handovers, and load balancing, all of which are challenging for existing GenAI models not optimized for such tasks. GenAI models must be robust and reliable to handle the variability and unpredictability of RAN environments. Ensuring high performance under diverse conditions is a significant challenge.

Thus, it is critical to invest in research and development programs focused specifically on creating GenAI foundation models for RAN applications, leveraging domain-specific expertise and ensuring that the models are robust and reliable.

Limited Access to High-Performance Accelerated Computing Infrastructure

Although several telcos have started to invest in accelerated computing infrastructure to support GenAI deployment and operation, the telecom industry, in general, has limited access to the necessary high-performance accelerated computing infrastructure.

This infrastructure includes, for example, GPUs and high-speed interconnects, which are essential for training and inference of large-scale GenAI models. Many GenAI applications in telecom, such as real-time customer support and network optimization, require low-latency processing. Limited access to high-performance computing can result in delays that degrade service quality.

Thus, it is essential to invest in cloud-based accelerated computing to enable access to scalable and flexible accelerated computing resources. Also, it is beneficial to deploy GenAI models at the edge to reduce latency and improve real-time processing capabilities, leveraging edge computing infrastructure equipped with high-performance accelerated computing.

Note that there are some initiatives underway that are aimed at helping this problem, but even if they are combined, they cannot solve the problem. One example is the partnership of SK Telecom and Lambda [38], which are offering the use of NVIDIA GPUs in a GPU-as-a-Service. Other companies include Google Cloud, Amazon Web Services, and Microsoft Azure. However, these companies typically offer these services to new customers and, in some cases, to support startups and open source. Note that their offerings are typically not for telecom operators.

The National Science Foundation (NSF) funds various research initiatives that may include access to high-performance computing resources for telecom-related projects. Similarly, the Department of Energy (DoE) has supercomputing resources available for research purposes, which may be accessible through specific grant applications. Again, these are typically not available to established telecom companies unless they partner with the NSF or DoE for specific research initiatives.

Hugging Face provides an inference API that allows users to run inference on pre-trained models hosted on their platform. They offer a free tier with limited usage and paid plans with discounted GPU pricing for higher usage. Hugging Face Spaces is a platform for deploying and sharing demos and applications built with Hugging Face models. Spaces offers free GPU usage for demos and small-scale applications, which could be useful for prototyping ideas. Hugging Face also provides a large collection of datasets that can be used for training ML models. Although they do not offer direct GPU access, using their datasets can help telecom companies reduce the time and resources needed for data collection and preprocessing.

In Europe, the EU's Horizon 2020 program has various projects that allocate free GPU time for research. The European High Performance Computing Joint Undertaking is a joint initiative between the EU, European countries, and private partners to develop a world-class supercomputing ecosystem in Europe [39]. So does the European National Competence Centers (EuroCC) [40], which is part of the European High-Performance Computing (HPC) strategy. EuroCC aims to establish national competence centers in various European countries to promote the use of HPC and provide access to GPU resources. Participants can benefit from training and access to GPU clusters for research

purposes. FZ Jülich Bootcamp, located at the Jülich Supercomputing Centre in Germany, organizes events like the EuroCC Nways to GPU Programming Bootcamp, where participants receive access to GPU clusters during the training sessions. This initiative helps researchers and students learn GPU programming while utilizing high-performance resources.

Measuring the Return on Investment (ROI) of GenAI accurately requires further investigation and is beyond the scope of this white paper.

Lack of Benchmark Datasets for Evaluation

Benchmark datasets are crucial for evaluating the performance and effectiveness of GenAI applications. They provide standardized, comprehensive, and representative data that can be used to assess various aspects of AI models, such as accuracy, reliability, and scalability.

There is no universally accepted set of benchmark datasets specific to telecom use cases for evaluating GenAI applications. This absence makes it difficult to standardize performance evaluation and comparison across different models and solutions. Existing datasets are fragmented or limited to specific aspects of telecom knowledge, without covering the full range of use cases that GenAI applications address.

Thus, it is essential to develop telecom industry-wide benchmark datasets that cover a broad range of use cases and scenarios, as well as the sharing of open benchmark datasets that are representative of various telecom environments and use cases.

Integration with Legacy Infrastructure

Telecom operators often rely on legacy systems that may not be compatible with modern GenAI technologies. The integration of new AI models with these systems can be complex and costly, requiring substantial investment in both time and resources. Some difficult and costly challenges include:

- **Organizational Resistance to Change**: Integrating legacy systems with modern technologies often requires significant changes to organizational processes and workflows. Most legacy systems were built using a "best-of-breed" approach, which often duplicated entire systems (e.g., inventory and configuration management) to work with specific operators. Overcoming resistance to change from employees who are comfortable with legacy systems can be a major challenge, requiring extensive training and change management efforts.
- **Data Format and Protocol Incompatibility**: Legacy systems often use outdated data formats and communication protocols that are incompatible with modern systems. Bridging this gap requires extensive data transformation and protocol conversion, which can be complex and costly. In addition, manual or automated data transformation processes can introduce errors, leading to data integrity issues and impacting service quality.
- **Programming Language and API Concerns:** The variety of programming languages, many of which are no longer mainstream (e.g., COBOL or old and non-supported versions of current languages like C), create interoperability, API compatibility, and skill gap issues. Of particular concern is version compatibility and feature deprecation. This creates the needs for complex middleware and API gateway solutions.
- **Security and Compliance Concerns:** Integrating legacy systems with modern infrastructure raises security and compliance challenges. Legacy systems may lack robust security features, making them vulnerable to cyber threats. Additionally, integrating these systems with new technologies can introduce compliance issues that need to be addressed.
- **Scalability and Performance Limitations:** Legacy systems were often designed for smaller-scale operations and may struggle to handle the increased demands of modern telecom networks. Integrating these systems with new technologies can expose their scalability and performance limitations, requiring costly upgrades or replacements.
- **Lack of Documentation and Expertise:** Many legacy systems were developed a long time ago, and the original developers may no longer be available. This lack of documentation and institutional knowledge makes it difficult to understand the system's inner workings and integrate it with new technologies.

The cognitive assistant [26] can help address some of the challenges, including documentation, expertise, programming language and API incompatibilities, and the data fusion problem that different data formats and protocols cause. In this role, it automatically finds the optimal approach to address each need.

Operational and Strategic Needs

This category focuses on the strategic requirements for operationalizing GenAI, including the need for comprehensive platforms and addressing scalability.

Need for End-to-End GenAI Development and Deployment Platforms

Cloud services powered by general-purpose LLMs provide a quick way to get started with GenAI technology. However, these services are often focused on a broad set of tasks and are not trained on telecom domain-specific data. This leads many telecom organizations to build their own solutions — a difficult task — because they must piece together various open-source tools, ensure compatibility, and provide their own support.

An end-to-end platform for building production-grade GenAI solutions provides a unified environment for every stage of the AI lifecycle, including data management, model

development, training, testing, deployment, and monitoring. Telecom organizations frequently use a collection of separate tools for different phases of AI development, leading to integration challenges and inefficiencies.

Transitioning from a prototype or proof-of-concept model to a production-grade solution involves significant effort. This includes scaling the model, ensuring reliability, and integrating with existing network infrastructure.

Thus, it is essential to adopt comprehensive end-to-end AI platforms designed to streamline GenAI development and deployment for enterprises. These simplify the process of data curation, training, and deployment and facilitate the swift development of customized, production-grade GenAI applications tailored to each organization's specific requirements.

Scalability Challenges

Scaling GenAI solutions from pilot projects to full-scale deployments can be problematic.

Developers of LLMs face several key challenges when scaling their offerings to meet the specific requirements of the telecommunications industry. These challenges stem from the unique demands of telco operations and the complexities involved in integrating AI solutions effectively.

Issues such as model performance under varying loads, infrastructure limitations, and the need for continuous optimization can hinder scalability efforts [42].

Model Performance under Varying Loads

The challenges primarily arise from the increased latency, reduced throughput, and potential rate limiting when the demand on the model exceeds its capacity. In addition, identifying and addressing bottlenecks is crucial for optimizing performance. If the inference speed is lacking, it may be due to memory-bound operations or inefficient resource allocation. Possibly the hardest to account for, especially for engineers who are not used to AI applications, is remembering that LLMs are not deterministic. This means their responses can vary even with the same input (e.g., they could be context sensitive). This variability can complicate performance monitoring and lead to inconsistencies in user experience, especially under varying load conditions.

Model Drift

Model drift (also called concept drift) can significantly impact the performance of ML models in production environments. As the underlying data distribution changes over time, the model's predictions may become less accurate, leading to suboptimal decisions and poor performance. When a model's performance degrades due to drift, its outputs become less trustworthy. Stakeholders may lose confidence in the model's decisions, limiting its practical utility.

Model drift can lead to higher costs, such as increased false positive rates in fraud detection or higher customer churn in marketing applications. These costs can quickly accumulate if the drift goes undetected. More importantly, model drift can have serious consequences depending on the application domain, such as financial losses, regulatory penalties, or reputational damage.

Mitigation of model drift requires the implementation of robust monitoring and drift detection mechanisms. This includes regularly evaluating model performance, deploying drift detection algorithms, and having a plan for retraining or updating the model when drift is detected. By proactively managing model drift, organizations can ensure their machine learning models maintain high performance and reliability over time.

The cognitive assistant [26] can be used to proactively detect model drift and select appropriate mechanisms to avoid the need for retraining.

Complexity of Telecommunications Services

The telecommunications sector is characterized by a diverse range of services and rapid technological advancements. Telcos must manage multiple service streams while ensuring high-quality performance across all platforms. This complexity makes it challenging to implement LLMs that can adapt to various service requirements and customer needs effectively.

Intensive and Special Resource Requirements: Developing and deploying LLMs is resource-intensive, requiring substantial human, infrastructure, and financial investments. The long development times, high costs, and the need for specialized expertise can hinder the scalability of LLM solutions within telco environments. This is particularly pronounced when considering the balance between operational efficiency and innovation because telcos often struggle to allocate resources effectively between routine operations and the development of innovative solutions.

Data Privacy and Customization: Telcos handle sensitive customer data, which raises significant concerns regarding data privacy. When considering whether to build a custom LLM or tune an existing one, telcos must navigate the trade-offs between customization and the potential risks associated with using third-party models. Custom LLMs offer greater control and privacy but are more complex and costly to develop, while tuning existing models can lead to sub-optimal performance due to limited customization options.

Integration with Existing Systems: Integrating LLMs into existing telco infrastructure poses significant challenges. Many telcos operate legacy systems that may be incompatible with modern AI technologies.

Regulatory and Ethical Considerations

The topics in this section address the regulatory frameworks,

ethical implications, and data protection issues associated with the deployment of GenAI technologies.

The cognitive assistant [26] can be used in each of these examples to provide a strong foundation in explainability and transparency. As discussed earlier, these are the basis of existing and future regulatory and ethical considerations because they are able to prove that something is true or false.

Regulatory Compliance

Telecoms must navigate a complex landscape of regulations that govern AI usage, data handling, and telecommunications services.

U.S. Regulations

The U.S. Executive Order (EO) 14110 [47] focuses on the regulation of AI and outlines the government's approach for ensuring responsible AI development and deployment. It establishes a framework for compliance and oversight, particularly concerning the ethical use of AI technologies. Note, however, that EO 14110 is *not* legally enforceable. This is because it is a presidential directive and thus does not have the same legal authority as legislation passed by Congress.

Key aspects of Executive Order 14110 include:

- **Regulatory Framework:** It sets forth guidelines for federal agencies to follow in the development and implementation of AI systems, emphasizing the importance of transparency, accountability, and fairness.
  - *Transparency and Accountability* are promoted by asking organizations to disclose how AI technologies are developed and used. This includes making information available about the data used for training models and the decision-making processes involved.
  - *Explainable AI* is encouraged. For example, it suggests the use of heatmaps to show which input parts the AI system focused on when making a decision.
  - *Mitigation of Discrimination* by calling for anti-discrimination laws to ensure that AI technologies do not perpetuate or exacerbate biases against individuals or groups.
  - *Fairness Testing and Measurement* by establishing frameworks and methodologies for assessing the fairness of AI models, ensuring that they meet specified standards before deployment.
- **Enforcement Mechanisms:** The order discusses the role of compute providers in ensuring compliance with AI regulations, suggesting that they may have legal obligations to safeguard AI systems and critical infrastructure. However, it does not provide actual legal obligations. Rather, it relies on the executive branch's ability to implement and oversee compliance within federal agencies.
- **Impact on AI Development:** It aims to create a balanced approach that fosters innovation while addressing potential risks associated with AI technologies, including bias and privacy concerns.

The Blueprint for an AI Bill of Rights [48] is a set of guidelines developed by the White House Office of Science and Technology Policy (OSTP) to ensure the responsible design and use of AI systems. It outlines five key principles:

1. **Safe and Effective Systems**
2. **Algorithmic Discrimination Protections**
3. **Data Privacy**
4. **Notice and Explanation:** (e.g., informing people when AI is being used and clarifying its impact)
5. **Human Alternatives:** Providing human alternatives and recourse when AI systems fail or cause harm (especially to people)

As such, its scope is different than EO 14110:

- The AI Bill of Rights is more focused on protecting individual rights and ensuring ethical AI use, while EO 14110 has a broader scope, addressing national security, innovation, and international leadership.
- The AI Bill of Rights provides guiding principles, while EO 14110 directs specific actions to take for federal agencies and AI developers.
- The AI Bill of Rights emphasizes civil rights, privacy, and transparency, while EO 14110 includes detailed provisions for AI safety, security, and innovation.

EU Regulations

An example of a legally enforceable regulation is the EU's AI Act [27]. It imposes binding regulations on AI developers and users. It also includes specific enforcement mechanisms, penalties, and compliance requirements for both public and private entities involved in AI.

There are a number of notable differences in the philosophy of the EU's AI Act that distinguish it from EO 14110. These include:

- **Focus:** While EO 14110 primarily focuses on federal agencies, the EU's AI Act applies to all organizations operating within the EU, including private companies. This broader scope of application in the EU's AI Act allows for more comprehensive regulatory oversight.
- **Risk-Based Approach:** The EU's AI Act categorizes AI systems based on their risk level, with stricter regulations for high-risk applications that can significantly impact individuals' lives. This includes AI used in areas such as healthcare, law enforcement, and employment. This is because it is focused on public and private companies throughout Europe rather than just federal agencies.
- **Prohibition of AI Usage:** The EU's AI Act prohibits the use of AI in certain situations, such as social scoring (i.e., evaluating individuals based on their behavior, characteristics, or social interactions, typically by using

algorithms and data analytics) by governments and the use of subliminal techniques to manipulate individuals. It also prohibits AI that exploits vulnerabilities of specific groups.

- **Implementation:** EO 14110 lacks detailed implementation mechanisms and guidelines, which may lead to variability in how agencies interpret and apply its provisions. The EU's AI Act provides detailed requirements for implementation, including conformity assessments and specific documentation practices that organizations must follow to demonstrate compliance.
- **Extraterritorial Usage:** The EU's AI Act applies to AI systems developed or used in the EU, even if the provider is located outside the EU. This ensures that non-EU companies must also comply with the regulations when offering AI services within the EU. This is not present in EO 14110.

The EU AI Act also has specific provisions for General-Purpose AI (GPAI) models, which are often referred to as foundation models. The Act defines a GPAI model as an AI model that displays "significant generality"[1] and is capable of performing a wide range of tasks competently.

In general, the requirements for GPAI models are different from non-GPAI models. The key requirements for GPAI models include:

- **Systemic Risk:** One key aspect of the EU AI Act is the classification of GPAI models with systemic risk. A GPAI model is considered to have systemic risk if it has a significant impact on the EU market or poses reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or society as a whole. This includes potential disruptions in critical economic sectors or negative effects on democratic processes. For GPAI models identified as having systemic risk, the Act imposes stricter requirements, such as enhanced transparency and compliance measures. Providers of these models must notify the EU within two weeks if their model is determined to represent a systemic risk.
- **Technical Documentation:** Providers must develop detailed technical documentation, including information about training data, computational resources, and energy consumption.
- **Training Data Summaries:** Providers are required to make publicly available detailed summaries of the content used for training their models.
- **Compliance with Copyright Laws:** Providers must establish policies to ensure compliance with EU copyright laws.
- **Documentation for Deployers:** Providers need to create documentation for deployers to understand and safely integrate the model.
- **Incident Reporting:** Providers must implement risk-management and incident-reporting frameworks.
- **Transparency and Accountability:** Providers must maintain up-to-date technical documentation and make it available to other AI system providers.

Key differences between GPAI and non-GPAI models are:

- **Scope and Flexibility:** GPAI models have broader applicability and are subject to specific requirements to ensure transparency and safety across various use cases. In contrast, non-GPAI models are regulated based on their risk level and specific use cases.
- **Documentation and Reporting:** GPAI models require extensive documentation and public summaries of training data, while non-GPAI models focus more on risk management and user transparency.
- **Incident Management:** GPAI models have specific incident reporting frameworks, whereas non-GPAI models emphasize risk management throughout their lifecycle.

EO 14110 and the EU's AI Act share several common functionalities, particularly in their emphasis on transparency and accountability in the use of AI. However, there are notable differences in how these functionalities are implemented and enforced. For example:

- **Transparency:** EO 14110 emphasizes the need for transparency in AI systems, *directing* federal agencies to disclose information about the data used and the functioning of AI technologies. The EU's AI Act *mandates* transparency for high-risk AI systems, requiring organizations to inform users when they are interacting with AI and to provide clear explanations of how AI systems operate. Theoretically, the EU's AI Act mandates that chatbots or deepfakes of any modality impersonating humans must be disclosed to end-users.
- **Accountability:** EO 14110 describes accountability measures for federal agencies to ensure that AI systems are developed and used ethically, with mechanisms for oversight and review of AI decision-making processes. In contrast, the EU's AI Act imposes strict accountability requirements for high-risk AI systems, including conformity assessments and the need for organizations to maintain documentation demonstrating compliance with the Act's provisions. Conformity includes market surveillance authorities and the ability to impose fines of up to 6% of a company's global turnover for non-compliance.
- **Fairness and Discrimination:** EO 14110 tries to mitigate discrimination in AI systems by requiring agencies to assess and address potential biases in their AI applications. The EU's AI Act requires risk assessments and measures to be enforced to ensure fairness in AI outputs.

[1]The EU AI Act defines "significant generality" GPAI models as models capable of competently performing a wide range of distinct tasks. Specifically, models with at least a billion parameters and trained with a large amount of data using self-supervision at scale are considered significant generality.

Data Privacy and Security Concerns

Compliance with these regulations can complicate the development and deployment of GenAI applications, requiring additional legal and operational considerations (e.g., GPDR [41] and more importantly the EU's AI Act [27].)

NIST's AI Risk Management Framework (AI RMF) [49] and the EU's AI Act represent two distinct approaches to managing AI risks, reflecting differing regulatory philosophies between the U.S. and the EU. The NIST AI RMF, released in January 2023, is a *voluntary framework* designed to help organizations manage risks associated with AI technologies. Its primary focus is on fostering trust and innovation in AI systems through a structured yet flexible approach. The framework emphasizes the importance of tailoring risk management practices to the specific needs of organizations, promoting ethical and responsible AI development without legal enforcement mechanisms or mandatory compliance requirements. Its main tenets include:

- **Governance:** Establishing roles and responsibilities for overseeing AI systems.
- **Measuring:** Assessing and monitoring AI risks and impacts.
- **Managing:** Implementing strategies to address identified risks.
- **Mapping:** Documenting processes and standards related to AI performance.

In contrast, the EU's AI Act is a comprehensive legal framework aimed at regulating AI within EU member states. It categorizes AI systems based on their risk levels, imposing binding legal requirements and significant penalties for non-compliance. The EU's AI Act is designed to protect citizens' rights and safety, requiring organizations to conduct impact assessments and implement risk management strategies before deploying AI systems.

Hence, there are stark differences between the two frameworks:

- **Scope:** The AI RMF is voluntary and flexible, focusing on *guidance* rather than *enforcement*. The EU's AI Act is *mandatory* and *legally binding*, with specific penalties for non-compliance.
- **Focus:** The AI RMF emphasizes building trust and promoting innovation in AI technologies, while the EU's AI Act prioritizes the protection of citizens' rights and safety through strict regulations.
- **Implementation:** The AI RMF encourages organizations to adopt best practices and tailor risk management to their needs. The EU's AI Act requires formal compliance processes, including risk assessments and the establishment of governance frameworks.

Ethical and Bias Issues

The use of AI raises ethical concerns, particularly regarding bias in AI models. If not properly addressed, biases in training data can lead to unfair treatment of certain user groups, which is particularly sensitive in the telecom industry where customer service and accessibility are critical.

EO 14110 has taken steps to address the issue of bias in AI models by establishing a framework for the safe and responsible development of AI. One of its key principles is "*advancing equity and civil rights*," which emphasizes that AI must not perpetuate discrimination or bias. The order directs federal agencies to ensure AI promotes equity, holds developers accountable for preventing unlawful discrimination, and builds on existing frameworks to protect civil rights.

In addition, the AI RMF provides guidance to organizations on managing AI risks, including bias. Key aspects include:

- Requiring AI vendors to have written policies on model risk management and bias training for employees.
- Encouraging bias assessments to identify sample bias, design bias, and proxy bias in AI models.
- Recommending additional bias mitigation steps like model retraining, human oversight, and ongoing monitoring.

In the telecom industry specifically, the Federal Communications Commission (FCC) has begun addressing AI bias. In July 2024 [50], the FCC approved a rule requiring telecom providers to ensure their AI customer service systems comply with nondiscrimination laws. Providers must establish policies, practices, and controls to minimize bias risks

Toward Responsible AI

Responsible AI is an approach to developing and deploying AI systems that emphasizes ethical, legal, and societal considerations. The primary goal of Responsible AI is to ensure that AI technologies are used in a manner that is safe, trustworthy, and aligned with human values.

The main tenets of Responsible AI include:

- **Fairness:** Ensuring that AI systems treat all individuals and groups equitably, without bias or discrimination. This involves examining the impact on various demographic groups and addressing any biases in the data or algorithms.
- **Transparency:** Making AI systems understandable and explainable. This includes providing clear information about how decisions are made, the data used, and the limitations of the AI system.
- **Accountability:** Holding individuals and organizations responsible for the outcomes of AI systems. This involves ensuring that there are mechanisms in place to audit and review AI systems and that those responsible for deploying AI can be held accountable for its impacts.

- **Privacy:** Protecting personal information and ensuring that data is collected and used with consent. This principle emphasizes the importance of safeguarding individuals' privacy rights in the development and deployment of AI systems.
- **Safety**: Ensuring that AI systems are robust, secure, and do not pose risks to individuals or society. This includes implementing measures to prevent misuse and mitigate potential harms.

Responsible AI incorporates ethical frameworks that guide the development and deployment of AI, ensuring that systems are designed to avoid harm and promote fairness. It has a wide scope of functionality. For this report, interpretability, explainability, and transparency are covered in more detail, as they are critical to both the EO 14110 and the EU's AI Act because they are how hypotheses, facts, and inferences are verified in Responsible AI.

Interpretability

Interpretability defines the ease with which a human can understand and reason about the model's structure and parameters. Interpretable models are often simpler and more intuitive, such as linear models or decision trees, making it easier to comprehend how they work under the hood. Interpretability is also important for building trust and understanding the model's overall behavior.

Interpretability focuses on how a model works rather than its outputs. More formally, a class C_1 of models is more interpretable than another class C_2 if the computational complexity of answering post-hoc queries for models in C_2 is higher than for those in C_1 [43]. Interpretability is a prerequisite for explainability.

Explainability

Explainability in AI systems is the ability of these systems to provide clear and understandable justifications for their decisions and actions to humans. This involves providing insights into how the model arrived at a particular output, such as the key factors or patterns it considered. Explainability helps users trust and accept the model's predictions by making its logic transparent. This concept is crucial as it helps users comprehend how AI models arrive at specific outcomes, thereby fostering trust and facilitating better human-AI interaction. This is an important foundation of the EU's AI Act [26].

Explainability has a number of important uses in modern AI systems, including explaining the behavior of an AI system. For example, the *reasoning paths as rationales* [45] concept involves using structured reasoning processes to explain or justify decisions made by AI models.

Transparency

Transparency refers to the clarity and openness regarding how models are trained, operate, and make decisions. It reflects the degree to which the workings and decisions of models can be understood by users and stakeholders. It involves providing stakeholders with sufficient information to understand the model's capabilities, limitations, and the processes behind its outputs. Transparency is crucial for building trust, ensuring ethical use, and facilitating effective oversight by developers, users, and regulators. For example, this involves model reporting, publishing evaluation results (with a focus on using standard benchmarks), providing explanations for its decisions, reporting any discovered or potential biases inherent in the design or operation of a model, and communicating uncertainty about any operations that it is not sure of. Transparency is about supporting human understanding, tailored to the needs of various stakeholders who may have different goals and contexts for using LLMs [46].

Transparency encompasses both interpretability and explainability but provides a broader understanding of the model's overall operation and behavior. This contrasts with the focus on the internal workings of the model by interpretability and justifying specific decisions by explainability. Transparency builds trust and ensures the ethical use of models and is also a major component of the EU's AI Act.

More specifically, the EU's AI Act mandates that AI systems must be designed to be transparent. Users should be informed when interacting with AI, particularly in high-risk applications. In addition, the EU's AI Act emphasizes the necessity of human oversight in high-risk AI applications to prevent automated decision-making from overriding human judgment. Transparency is also crucial for accountability.

8. RECOMMENDATIONS

Building on the insights from the ATIS TOPS Council study on AI Networks Applications, which identified key use cases of GenAI within the telecom sector, the following recommendations are given:

1. **Coordination across standards development organizations**

It is recommended that industry members coordinate on enabling GenAI support in standards development organizations. This will help bring and align GenAI innovations in evolving standards, ensuring interoperability, scalability, and compliance across global telecom markets.

2. **Facilitate development of GenAI foundation models for RAN**

To advance beyond specific use cases, it is recommended to facilitate the development of GenAI foundation models for RAN applications in the industry, leveraging domain-specific expertise. The models would serve as shared, flexible frameworks to accelerate GenAI-driven optimizations in wireless communications.

3. **Develop a cognitive reasoning assistant**

A cognitive reasoning assistant would enable advanced decision-making, predictive analysis, and contextual reasoning, offering assistance in complex telecom operations and improving the efficiency and accuracy of GenAI-driven telecom solutions. It is recommended that a cognitive reasoning assistant be created by combining transformer models with knowledge graphs.

9. CONCLUSION

Integrating GenAI in telecommunications presents transformative opportunities across various network domains, driving innovation in automation, personalization, and network management. By adopting the recommendations outlined in this report, such as standardizing AI across the industry, developing foundation models, and creating cognitive reasoning systems, the telecommunications sector can overcome current challenges and fully harness the potential of GenAI. The future of telecom networks lies in their ability to seamlessly incorporate AI-driven solutions to enhance efficiency, improve service quality, and enable new use cases that address evolving customer and operational demands. As GenAI continues to advance, its impact will extend beyond current applications, reshaping the telecommunications landscape and creating new opportunities for both service providers and consumers.

10. APPENDIX

A. Background on Semantics and Cognition

Cognition is defined in [14] as the process of acquiring and understanding data and information and producing new data, information, and knowledge. A cognition model defines how cognitive processes, such as comprehension, decision-making, action, and prediction, are performed and influence decisions. This is how, for example, behavior is recognized and managed.

The ENI cognitive management functional block is based on an innovative cognition model [14]. The ENI cognition model draws heavily on how human cognition is performed. More specifically, the perception portion provides the notion of classifying data into pre-defined representations that are understood and relevant to the current situation. Memory is used to increase comprehension of the situation, and actions are judged by how effectively they perform to support the situation.

ENI cognitive management learns from experience to improve its performance. This includes acquiring knowledge from instruction or experience, revising and correcting existing knowledge, and combining existing data and information to infer and deduce new knowledge. A cognitive system must be prepared for data and information to evolve. Hence, a cognitive system may draw on multiple sources of information, including both structured and unstructured digital information, as well as sensory inputs (visual, gestural, auditory, or sensor-provided), in order to comprehend the current situation.

A cognitive system should be able to adapt its governance in accordance with changing information, contexts, and situations. It should also be able to adapt its functionality as its goals and requirements evolve. Cognitive systems may use multiple mechanisms to make decisions, including deterministic, probabilistic, fuzzy logic, neural networks, evolutionary algorithms, and hybrid mechanisms.

The individual functional blocks of a cognitive system, as well as multiple cognitive systems, are able to collaborate on a set of tasks. The specific set of functional blocks assigned to the collective is driven by their suitability to accomplish the tasks of the current set of goals. The ENI system uses registered characteristics of each functional block that describe their capabilities, along with applicable metadata, to make the selection. Once the set of tasks has been completed, the collective may disband.

For instance, the ENI system cognitive architecture [14][15] enables the ENI system to understand ingested data and information, as well as the context that defines how those data were produced. This enables the meaning of the data to be evaluated and prioritized and determines if any actions need to be taken to ensure that the goals and objectives of the system are met. This includes improving or optimizing performance, reliability, and/or availability. Each ENI system functional block can recruit other functional blocks to accomplish tasks in a prioritized manner. This approach enables teams of functional blocks to take on goals that individual functional blocks cannot achieve themselves.

The ENI cognitive system uses hypotheses and reasoning to make decisions. This approach enables explanations to be generated about decisions taken. However, it is important to note that cognitive systems should be used to augment human decision-making and action processes, as opposed to replacing humans. This is why the ENI system operates in two distinct modes: recommendation and command. The former enables an ENI system to function as an assistant that recommends actions to take. The latter enables an ENI system to function as a "super-orchestrator," governing other management components (e.g., orchestrators, management systems, and controllers).

A cognition model enables a cognitive system to reason about what actions to take in a methodological and structured manner. It can learn from its experience to improve its performance. It can also examine its own capabilities and prioritize the use of its services and resources, and if necessary, explain what it did and accept external commands to perform necessary actions. This combines to form a set of closed loops. An exemplary implementation from [14] is shown in Figure 14.

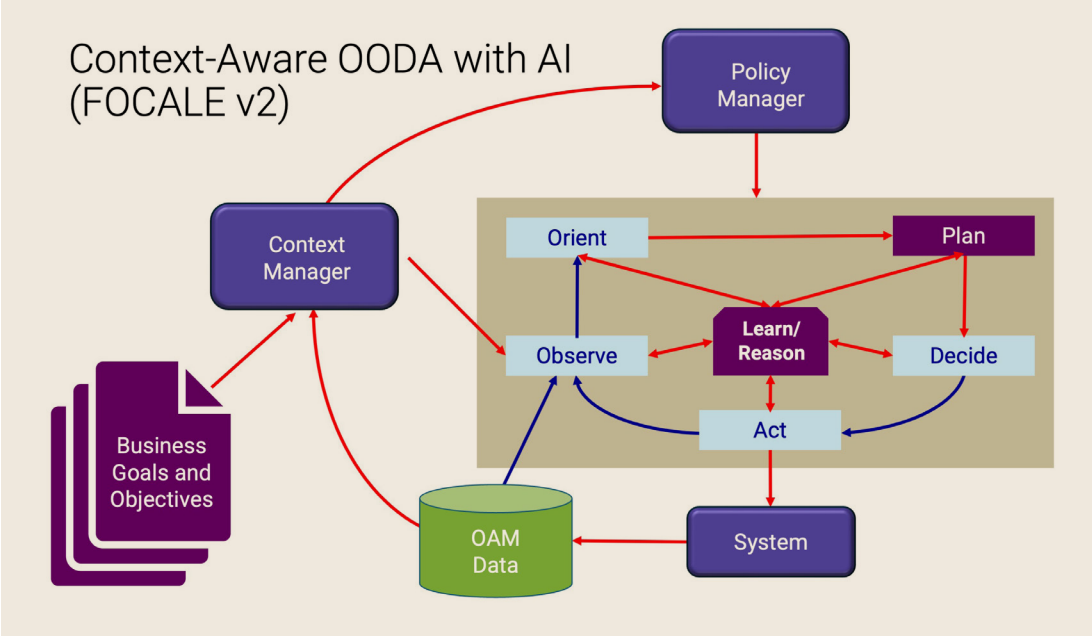


Figure 14: Simplified Cognitive Control Loop used in ENI

This figure is based on the OODA [16] control loop and is enhanced as follows:

- First, OODA was designed to apply to a single decision-maker. ENI's version is designed to accommodate collaborative decision-making.
- Second, a dedicated planning stage was inserted between the Orient and Decide cycles. This was done to accommodate situation awareness.
- Third, based on human psychology [17], cognitive learning (i.e., reactive, deliberative, and reflective) was used.
- Fourth, learning (as a whole) was inserted to monitor all phases of each control loop.
- Finally, policy management is used to make each functional block in the ENI cognition loop configurable using a standardized set of commands.

A Neuro-symbolic AI system is the combination of a neural network (also called a sub-symbolic system) and a symbolic AI system, such as a knowledge graph or an ontology. Symbolic AI focuses on the processing and manipulation of symbols or concepts rather than numerical data.

It uses high-level, human-readable representations of problems, logic, and search. Symbolic AI systems rely on predefined rules and logical reasoning to make inferences and decisions. Examples include expert systems, knowledge-based systems, and automated theorem provers. A neuro-symbolic AI system can learn from data (like neural AI) while also being able to reason about abstract concepts and relationships (like symbolic AI).

11. ACRONYMS AND ABBREVIATIONS

AI	Artificial Intelligence
AIDA	Artificial Intelligence and Data Act
AIGC	AI-Generated Content
API	Application Programming Interface
AR	Augmented Reality
AS	Autonomous System
BERT	Bidirectional Encoder Representations from Transformers
BGP	Border Gateway Protocol
CDT	Cognitive Digital Twin
CIR	Channel Impulse Response
CPI	Customer Product Information
CPU	Central Processing Unit
CSI	Channel State Information
DoE	Department of Energy
EM	Exact Match
eNB	E-UTRAN Node B
ENI	Experiential Networked Intelligence
EO	Executive Order
ETSI	European Telecommunications Standards Institute
FCC	Federal Communications Commission
GAN	Generative Adversarial Network
GPAI	General-Purpose AI
GPT	Generative Pre-Trained Transformer
GPU	Graphics Processing Unit
HPC	High-Performance Computing
HTTP	Hypertext Transfer Protocol
IP	Internet Protocol
IQ	In-Phase and Quadrature
IR	Initial Retrieval
ISAC	Integrated Sensing and Communication
Llama	Large Language Model Meta AI
LLM	Large Language Model
LM	Language Model
LTE	Long-Term Evolution
MAC	Media Access Control Address
MIMO	Multiple-Input Multiple-Output
ML	Machine Learning
MRR	Mean Reciprocal Rank
MS MARCO	Microsoft Machine Reading Comprehension
nDCG	Normalized Discounted Cumulative Gain

NIST ..... National Institute of Standards and Technology
NLP ..... Natural Language Processing
NLU ..... Natural Language Understanding
NPU ..... Neural Processing Unit
NSF ..... National Science Foundation
OODA ..... Observe-Orient-Decide-Act
O-RAN ..... Open Radio Access Network
OSTP ..... Office of Science and Technology Policy
PIN ..... Personal Identification Number
Q&A ..... Question-and-Answer
RAG ..... Retrieval Augmented Generation
RAN ..... Radio Access Networks
RIS ..... Reconfigurable Intelligence Surface
RL ..... Reinforcement Learning
RMF ..... Risk Management Framework
ROI ..... Return on Investment
RR ..... Re-Ranking
RTP ..... Real-Time Transport Protocol
SDO ..... Standards Development Organization
SIP ..... Session Initiation Protocol
SQuAD ..... Stanford Question Answering Dataset
TeleQuAD ..... Telecom Question Answering Dataset
TPU ..... Tensor Processing Unit
TR ..... Trouble Report
TRDI ..... Trouble Report Duplicate Identification
UE ..... User Equipment
VR ..... Virtual Reality
WPFM ..... Wireless Physical Layer Foundation Model

12. REFERENCES

[1] X. Lin, "Artificial Intelligence in 3GPP 5G-Advanced: A Survey," IEEE ComSoc Technology News, Aug. 2023. [Online]. Available: https://www.comsoc.org/publications/ctn/artificial-intelligence-3gpp-5g-advanced-survey

[2] Y. Yang, Y. Li, W. Zhang, F. Qin, P. Zhu, and C.-X. Wang, "Generative-Adversarial-Network-Based Wireless Channel Modeling: Challenges and Opportunities," IEEE Communications Magazine, vol. 57, no. 3, pp. 22-27, Mar. 2019.

[3] S. Chaudhuri, D. Ritchie, J. Wu, K. Xu, and H. Zhang, "Learning generative models of 3D structures," Computer Graphics Forum, vol. 39, no. 2, pp. 643-666, May 2020.

[4] X. Lin, L. Kundu, C. Dick, E. Obiodu, T. Mostak, and M. Flaxman, "6G Digital Twin Networks: From Theory to Practice," IEEE Communications Magazine, vol. 61, no. 11, pp. 72-78, Nov. 2023.

[5] S. Zhang, A. Wijesinghe, and Z. Ding, "RME-GAN: A Learning Framework for Radio Map Estimation Based on Conditional Generative Adversarial Network," IEEE Internet of Things Journal, vol. 10, no. 20, pp. 18016-18027, Oct. 2023.

[6] X. Lin, "An Overview of the 3GPP Study on Artificial Intelligence for 5G New Radio," IEEE ComSoc Technology News, Mar. 2024. [Online]. Available: https://www.comsoc.org/publications/ctn/overview-ai-3gpps-ran-release-18-enhancing-next-generation-connectivity

[7] B. Tolba, M. Elsabrouty, M. G. Abdu-Aguye, H. Gacanin, and H. M. Kasem, "Massive MIMO CSI Feedback Based on Generative Adversarial Network," IEEE Communications Letters, vol. 24, no. 12, pp. 2805-2808, Dec. 2020.

[8] Y. Cui, A. Guo, and C. Song, "TransNet: Full Attention Network for CSI Feedback in FDD Massive MIMO system," IEEE Wireless Communications Letters, vol. 11, no. 5, pp. 903-907, May 2022.

[9] X. Lin et al., "5G New Radio: Unveiling the Essentials of the Next Generation Wireless Access Technology," IEEE Communications Standards Magazine, vol. 3, no. 3, pp. 30-37, Sep. 2019.

[10] E. Balevi and J. G. Andrews, "Unfolded Hybrid Beamforming with GAN Compressed Ultra-Low Feedback Overhead," IEEE Transactions on Wireless Communications, vol. 20, no. 12, pp. 8381-8392, Dec. 2021.

[11] Y. Wang, Z. Gao, D. Zheng, S. Chen, D. Gunduz and H. V. Poor, "Transformer-Empowered 6G Intelligent Networks: From Massive MIMO Processing to Semantic Communication," IEEE Wireless Communications, early access, Nov. 2022.

[12] K. Davaslioglu and Y. E. Sagduyu, "Generative Adversarial Learning for Spectrum Sensing," in Proc. IEEE Int. Conf. Communications (ICC), Kansas City, MO, USA, 2018, pp. 1-6.

[13] A. Toma et al., "AI-Based Abnormality Detection at the PHY-Layer of Cognitive Radio by Learning Generative Models," IEEE Transactions on Cognitive Communications and Networking, vol. 6, no. 1, pp. 21-34, Mar. 2020.

[14] ETSI, "System Architecture," ETSI ENI GS 005, v4.0.2, Nov. 2023.

[15] J. Strassner, "ENI Vision: Understanding the Operator Experience Using Cognitive Management," ETSI White Paper, Dec. 2022.

[16] J. R. Boyd, "The Essence of Winning and Losing," Jun. 1995.

[17] R. Crutzen, G-J. Y. Peters, "Evolutionary Learning Processes as the Foundation for Behaviour," Health Psychology Review, vol. 12, No. 1, pp. 43-57. 2018.

[18] P. Ünal, "Cognitive Digital Twins: Digital Twins That Learn By Themselves, Foresee the Future, and Act Accordingly," Digital Twin Consortium blog, Sep. 2022. [Online]. Available: https://www.digitaltwinconsortium.org/2022/09/cognitive-digital-twins-digital-twins-that-learn-by-themselves-foresee-the-future-and-act-accordingly/

[19] NVIDIA, "AI Chatbot with Retrieval Augmented Generation," Technical Brief. [Online]. Available: https://docs.nvidia.com/ai-enterprise/workflows-generative-ai/0.1.0/technical-brief.html

[20] F. -A. Croitoru, V. Hondru, R. T. Ionescu and M. Shah, "Diffusion Models in Vision: A Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 9, pp. 10850-10869, Sept. 2023.

[21] H. Du et al., "Diffusion-Based Reinforcement Learning for Edge-Enabled AI-Generated Content Services," IEEE Transactions on Mobile Computing, vol. 23, no. 9, pp. 8902-8918, Sept. 2024.

[22] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 6840-6851, 2020.

[23] H. Du, R. Zhang, D. Niyato, J. Kang, Z. Xiong, and D. I. Kim, "Reinforcement Learning With Large Language Models (LLMs) Interaction For Network Services," in Proc. Int. Conf. Computing, Networking and Communications (ICNC), Big Island, HI, USA, Feb. 19-22, 2024.

[24] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 10, pp. 1175-1191, Oct. 2001.

[25] J. Fontaine, A. Shahid, and E. De Poorter, "Towards a Wireless Physical-Layer Foundation Model: Challenges and Strategies," arXiv preprint arXiv:2403, 2024.

[26] J. Strassner, "Cognitive Assistant: A Semantic Knowledge Graph-Enabled Transformer for Reasoning and Decision-Making," contribution to ATIS ANA working group, 2024.

[27] "The Act texts | EU Artificial Intelligence Act." [Online]. Available: https://artificialintelligenceact.eu/the-act/

[28] J. Strassner, "Policy-Based Network Management," Morgan Kaufman, 2003.

[29] "Directory Listing /ISG/ENI/OPEn/ENI030 (Release 4)." [Online]. Available: https://docbox.etsi.org/ISG/ENI/Open/ENI030%20(Release%204)

[30] A. Garcez and L. Lamb, "Neurosymbolic AI: The 3rd wave," Dec. 2020.

[31] A. Piplai et al., "Knowledge-Enhanced Neuro-Symbolic AI for Cybersecurity and Privacy," arXiv preprint arXiv:2308.02031, Jul. 2023. [Online]. Available: https://arxiv.org/pdf/2308.02031

[32] N. Slonim, Y. Alzate et al., "An autonomous debating system," Nature, vol. 591, no. 7850, pp. 379-384, 2021.

[33] D. Kahneman, "Thinking, Fast and Slow," Farrar, Straus and Giroux, 2011.

[34] A. Smirnova et al., "Nessy: A Neuro-Symbolic System for Label Noise Reduction," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 8, pp. 8300-8311, 2023.

[35] V. Belle et al., "Neuro-Symbolic AI+ Agent Systems: A First Reflection on Trends, Opportunities and Challenges," International Conference on Autonomous Agents and Multiagent Systems, 2023.

[36] Daniele et al., "Simple and Effective Transfer Learning for Neuro-Symbolic Integration," in Proc. Int. Conf. Neural Symbolic Learning and Reasoning, 2024.

[37] T. Verma et al., " Defining Explanation in an AI Context," in Proc. 3rd Blackbox NLP Workshop on Analyzing and Interpreting Neural Networks for NLP, 2020, pp. 314-322.

[38] S. Telecom, "SKT Signs Partnership Agreement with Lambda," TelecomTV, Aug. 2024. [Online]. Available: https://www.telecomtv.com/content/telcos-and-ai-channel/skt-signs-partnership-agreement-with-lambda-51086/

[39] "The European High Performance Computing Joint Undertaking," Shaping Europe's Digital Future, Sep. 2024. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/high-performance-computing-joint-undertaking

[40] "EuroCC Workshop HPC and Industry Application at IT2024," EuroCC ACCESS, Feb. 2024. [Online]. Available: https://www.eurocc-access.eu/eurocc-workshop-hpc-and-industry-application-at-it2024/

[41] "Data protection regulation," The European Council. [Online]. Available: https://www.consilium.europa.eu/en/policies/data-protection/data-protection-regulation/

[42] Y. Zhang et al., "Mobile Generative AI: Opportunities and Challenges," IEEE Wireless Communications, vol. 31, pp. 58-64, 2024.

[43] P. Barcelo et al., "Model Interpretability through the Lens of Computational Complexity," Advances in neural information processing systems, 2020, pp. 15487-15498.

[44] A. Vaswani et al., "Attention is all you need," Advances in Neural Information Processing Systems, 2017.

[45] Z. Hu et al., "Empowering Language Models with Knowledge Graph Reasoning for Question Answering," in Proc. Conf. Empirical Methods in Natural Language Processing, 2022, pp. 9562-9581.

[46] Q. V. Liao and J. W. Vaughan, "AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap," Harvard Data Science Review, no. 6, 2024.

[47] "Safe, secure, and trustworthy development and use of artificial intelligence," Federal Register, Nov. 2023. [Online]. Available: https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

[48] The White House, "Blueprint for an AI Bill of Rights | OSTP | The White House," The White House, Nov. 2023. [Online]. Available: https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[49] "AI Risk Management Framework | NIST," NIST, Oct. 2024. [Online]. Available: https://www.nist.gov/itl/ai-risk-management-framework

[50] A. Fellows et al., "FCC issues Notice of proposed Rulemaking regarding the Use of AI-Generated Technologies for Consumer Communications," The Data Advisor, Aug. 2024. [Online]. Available: https://www.wsgrdataadvisor.com/2024/08/fcc-issues-notice-of-proposed-rulemaking-regarding-the-use-of-ai-generated-technologies-for-consumer-communications/

[51] J. Devlin, et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2019. [Online]. Available: https://arxiv.org/pdf/1810.04805

[52] X. Lin et al., "A Primer on Generative AI for Telecom: From Theory to Practice," arXiv preprint arXiv:2408.09031, 2024. [Online]. Available: https://arxiv.org/pdf/2408.09031

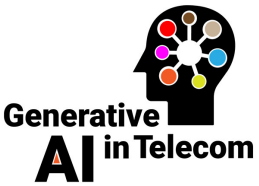
ACKNOWLEDGEMENTS

TOPS Report Leader: Carroll Gray-Preston, ATIS

Industry Contributors: Ivy Kelly, C Spire
Jana Westover, C Spire
Ayodele Damola, Ericsson
John Strassner, Futurewei
Xingqin Lin, NVIDIA

Academic Contributors: Jaron, Fontaine, IMEC - Ghent University
Adnan, Shahid, IMEC - Ghent University
Dusit Niyato, Nanyang Technological University, Singapore
Hongyang Du, The University of Hong Kong

AI NETWORK APPLICATIONS WORKING GROUP



Generative AI applications will improve customer experiences across multiple industries. Despite various telecom standard bodies discussing AI/ML, there is no focused study on Generative AI in telecommunications. Generative AI has the potential to reinvent the telecom industry, enabling customers to use telecom services more easily and efficiently. Concurrently, it enables service providers to offer more secure, reliable, and personalized services. Launched in October 2023, this group has been working to survey generative AI/ML use cases across the network. Use cases may include semantics, context awareness, personalized services, etc. Findings from this work have resulted in this white paper that assesses and prioritizes key use cases for network applications, addresses a critical gap in the industry, and provides recommendations for advancing future AI implementation across networks.

AI Network Applications Working Group Leadership:

Co-Chair, John Strassner, Futurewei
Co-Chair, Xingqin Lin, NVIDIA
Carroll Gray-Preston, ATIS

AI NETWORK APPLICATIONS WORKING GROUP MEMBERSHIP

Analog Devices, Inc.
Apple Inc.
AT&T
C Spire Wireless
Charter Communications
Comtech Telecommunications Corp.
Ericsson
Fujitsu
Futurewei
Hewlett Packard Enterprise
IMEC
InterDigital Communications Corporation
Juniper Networks
Lumen
Microsoft Corporation
Motorola Mobility LLC (A Lenovo Company)
Nokia
Nvidia Corporation
Peraton Labs
Qualcomm Incorporated
T-Mobile USA
TDS
Telnyx
TELUS
Verizon

COPYRIGHT AND DISCLAIMER

ATIS-I-0000102

Published November 2024

Alliance for Telecommunications Industry Solutions
1200 G Street, NW, Suite 500
Washington, DC 20005

www.atis.org
For information, contact ATIS at (202) 628-6380.