الجمهورية الجزائرية الديمقراطية الشعبية
The People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research
جامعة محمد بوضياف بالمسيلة
University Mohamed Boudiaf of M'sila

كلية الرياضيات والإعلام الآلي
Faculty of Mathematics and Informatics

قسم الإعلام الآلي
Department of Computer Science

**Domain:** Mathematics and Computer Science

Thesis Presented to Fulfill the Partial Requirement
for **Master's Degree** in Computer Science

**Specialty:** Networks and Information and Communication and Technologies

**Prepared By:** Fatima Baadji

**Supervised By:**

Hichem Debbi

ENTITLED

# Network Management Assistance through Large Language Models (LLMs)

**Jury Members**

| | |
|---|---|
| Azeddine Attir | President |
| Hichem Debbi | Supervisor |
| Abdessattar Ghemougui | Examiner |

**Academic Year 2024/2025**

# Acknowledgments

First and foremost, I thank Allah the Almighty for giving me the knowledge, courage and patience necessary to carry this work to its conclusion.

I would like to express my deepest gratitude to my advisor, **Dr. Hichem Debbi**, for their invaluable guidance, patience, and continuous support throughout my research journey. Their expertise and encouragement have been instrumental in shaping this work.

A special note of appreciation goes to my colleagues and friends, whose constant support, helpful discussions, and uplifting presence have made this journey both intellectually enriching and personally fulfilling.

I am also deeply grateful to my family for their unwavering love and encouragement. Their belief in me has been my greatest source of strength and perseverance.

**Fatima Baadji**
**June 2025**

**Abstract**

This thesis explores the transformative role of Large Language Models (LLMs) in the domain of network optimization, particularly within the context of 5G communication technologies. It starts by studying deep learning fundamentals and neural network architectures, emphasizing the evolution and impact of Models such as GPT, BERT, and DeepSeek. The study then examines the integration of these models into modern networking workflows, focusing on their applications in network security, task classification as well as answering telecom-domain questions.

A core challenge addressed in this work lies in the sheer volume and complexity of technical data in the telecommunications industry—particularly across 3GPP (3rd Generation Partnership Project) standards, which has overseen the development of universal standards for Mobile Wireless Networks (MWNs). 3GPP continuously publishes a large number of intricate documents, making it difficult for engineers and researchers to stay updated and extract relevant information efficiently. This creates a need for advanced methods to process, analyze, and understand these documents to ensure network reliability and performance.

To address this, the thesis aiming to be a roadmap for researchers and practitioners to leverage LLMs in solving various telecom tasks. Special attention is given to the application of fine-tuning and Retrieval-Augmented Generation (RAG) techniques to improve technical comprehension and automate knowledge extraction from3GPP specifications. A practical evaluation is conducted using the TSpec-LLM dataset, and chatbot are developed to classify telecom tasks and respond to domain-specific queries.

Finally, The research demonstrates how LLMs will become increasingly important for telecom-specific operations by enhancing information retrieval accuracy and efficiency and accessibility in complex technical domains.

---

**Keywords** — Chatbot, 3GPP Documents, Large Language Models (LLMs), Network Optimization, Retrieval-Augmented Generation (RAG), Fine-tuning .

---

**Résumé**

Ce mémoire explore le rôle transformateur des grands modèles de langage (LLM) dans le domaine de l'optimisation des réseaux, notamment dans le contexte des technologies de communication 5G. Elle commence par étudier les fondamentaux de l'apprentissage profond et les architectures de réseaux neuronaux, en mettant l'accent sur l'évolution et l'impact de modèles tels que GPT, BERT et DeepSeek. L'étude examine ensuite l'intégration de ces modèles dans les flux de travail des réseaux modernes, en se concentrant sur leurs applications en matière de sécurité des réseaux, de classification des tâches et de réponse aux questions du domaine des télécommunications.

L'un des principaux défis abordés dans ce travail réside dans le volume et la complexité des données techniques dans le secteur des télécommunications, notamment celles relatives aux normes 3GPP (3rd Generation Partnership Project), qui ont supervisé le développement de normes universelles pour les réseaux sans fil mobiles (MWN). Le 3GPP publie continuellement un grand nombre de documents complexes, ce qui complique la tâche des ingénieurs et des chercheurs pour se tenir informés et extraire efficacement les informations pertinentes. Cela crée un besoin de méthodes avancées pour traiter, analyser et comprendre ces documents afin de garantir la fiabilité et les performances du réseau.

Pour y parvenir, ce mémoire vise à servir de feuille de route aux chercheurs et aux praticiens afin d'exploiter les LLM pour résoudre diverses tâches liées aux télécommunications. Une attention particulière est accordée à l'application de techniques de réglage fin et de génération augmentée de données (RAG) pour améliorer la compréhension technique et automatiser l'extraction des connaissances issues des spécifications 3GPP. Une évaluation pratique est réalisée à l'aide du jeu de données TSpec-LLM, et des agents conversationnels sont développés pour classer les tâches liées aux télécommunications et répondre aux requêtes spécifiques au domaine.

Enfin, la recherche démontre l'importance croissante des LLM pour les opérations spécifiques aux télécommunications, en améliorant la précision, l'efficacité et l'accessibilité de la recherche d'informations dans des domaines techniques complexes.

---

**Mots-clés** — Agent conversationnel, Documents 3GPP, Grands Modèles Linguistiques, Optimisation de réseau, Génération augmentée par récupération, Affinage.

---

## الملخص

تستكشف هذه المذكرة الدور التحويلي الذي تلعبه النماذج اللغوية الكبيرة (LLMs) في مجال تحسين أداء الشبكات، لا سيما في سياق تقنيات الاتصالات من الجيل الخامس .(5G) تبدأ الدراسة بمراجعة أساسيات التعلم العميق وهندسة الشبكات العصبية، مع التركيز على تطور وتأثير نماذج مثل GPT وBERT وDeepSeek. ثم تتناول الدراسة كيفية دمج هذه النماذج في سير عمل الشبكات الحديثة، مع التركيز على تطبيقاتها في أمن الشبكات، وتصنيف المهام، والإجابة على الأسئلة المتعلقة بمجال الاتصالات.

تمثل التحديات الأساسية التي تعالجها هذه المذكرة في الحجم الهائل والتعقيد الكبير للبيانات الفنية في قطاع الاتصالات، لا سيما ضمن معايير 3GPP (مشروع الشراكة للجيل الثالث)، الذي يشرف على تطوير المعايير العالمية لشبكات الاتصالات اللاسلكية المتنقلة .(MWNs) تنشر 3GPP باستمرار عددًا كبيرًا من الوثائق الفنية المعقدة، مما يجعل من الصعب على المهندسين والباحثين مواكبتها واستخلاص المعلومات ذات الصلة بكفاءة. ويؤدي ذلك إلى ضرورة تطوير أساليب متقدمة لمعالجة هذه الوثائق وتحليلها وفهمها لضمان موثوقية وأداء الشبكة.

ولمواجهة هذه التحديات، تهدف هذه المذكرة إلى أن تكون بمثابة خارطة طريق للباحثين والممارسين للاستفادة من النماذج اللغوية الكبيرة في حل مختلف المهام المتعلقة بمجال الاتصالات. ويتم إيلاء اهتمام خاص لتقنيات التخصيص (Fine-tuning) والتوليد المعزز بالاسترجاع (RAG) لتحسين فهم المحتوى التقني وأتمتة استخراج المعرفة من مواصفات 3GPP. كما يتم إجراء تقييم عملي باستخدام مجموعة بيانات ،TSpec-LLM وتطوير روبوتات محادثة (Chatbots) لتصنيف المهام المتعلقة بالاتصالات والرد على الاستفسارات الخاصة بالمجال.

وأخيرًا، تُظهر هذه الدراسة كيف ستصبح النماذج اللغوية الكبيرة ذات أهمية متزايدة في العمليات المتخصصة في مجال الاتصالات، من خلال تعزيز دقة وكفاءة واستيعاب استرجاع المعلومات في المجالات الفنية المعقدة.


---


**الكلمات المفتاحية** — روبوت الدردشة، مستنداتGPP3 ، نماذج اللغة الكبيرة، تحسين الشبكة، التوليد المعزز بالاسترجاع ، الضبط الدقيق.


---

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| ADN | **A**utonomous **D**riving **N**etwork |
| AI | **A**rtificial **I**ntelligence |
| ANNs | **A**rtificial **N**eural **N**etworks |
| BERT | **B**idirectional **E**ncoder **R**epresentations from **T**ransformers |
| CLM | **C**ausal **L**anguage **M**odeling |
| CNNs | **C**onvolutional **N**eural **N**etworks |
| CPS | **C**yber-**P**hysical **S**ystems |
| CV | **C**omputer **V**ision |
| DAI | **D**istributed **A**rtificial **I**ntelligence |
| DCN | **D**ata **C**enter **N**etworks |
| DTNs | **D**igital **T**win **N**etworks |
| EMBB | **E**nhanced **M**obile **B**roadband |
| GANs | **G**enerative **A**dversarial **N**etworks |
| GPT | **G**enerative **P**re-trained **T**ransformers |
| IBN | **I**ntent-**B**ased **N**etworking |
| ICL | **I**n-**C**ontext **L**earning |
| IoT | **I**nternet of **T**hings |
| LLMs | **L**arge **L**anguage **M**odels |
| MEC | **M**ultiaccess **E**dge **C**omputing |
| MIMO | **M**ultiple **I**nput **M**ultiple **O**utput |
| MLM | **M**asked **L**anguage **M**odeling |

| | |
|---|---|
| MLPs | **M**ulti-**L**ayer **P**erceptrons |
| MMTC | **M**assive **M**achine-**T**ype **C**ommunication |
| NAS | **N**eural **A**rchitecture **S**earch |
| NFV | **N**etwork **F**unction **V**irtualization |
| NLP | **N**atural **L**anguage **P**rocessing |
| NoC | **N**etwork-on-**C**hip |
| PQC | **P**ost-**Q**uantum **C**ryptography |
| RAG | **R**etrieval-**A**ugmented **G**eneration |
| ReLU | **R**ectified **L**inear **U**nit |
| RNNs | **R**ecurrent **N**eural **N**etworks |
| SDN | **S**oftware-**D**efined **N**etworking |
| SFT | **S**upervised **F**ine-**T**uning |
| URLLC | **U**ltra-**R**eliable **L**ow-**L**atency **C**ommunication |
| VNFs | **V**irtualized **N**etwork **F**unctions |
| WBAN | **W**ireless **B**ody **A**rea **N**etwork |
| WDMA | **W**avelength **D**ivision **M**ultiplexing **A**ccess |

# General Introduction

The fast development of wireless communication technologies has reshaped modern connectivity through 5G network deployment and ongoing 6G system development which advances speed capabilities and scalability and intelligence features. The increasing complexity and data intensity of networks requires more than traditional optimization methods to support real-time decision-making and anomaly detection and answering-questions as well as adaptive resource management. The development of Large Language Models (LLMs) within artificial intelligence (AI) has established them as essential tools for enhancing network intelligence.

LLMs, such as GPT, Llama, DeepSeek, and their fine-tuned derivatives, have demonstrated remarkable success in understanding and generating human-like text. Their capabilities extend beyond language tasks to domains such as cybersecurity, intelligent document parsing, anomaly detection, and automated reasoning.

This thesis explores how the capabilities of LLMs can be strategically applied to optimize telecommunication networks, with a particular focus on their integration in 5G architectures, 3GPP (3rd Generation Partnership Project) standards, and network automation workflows.

This thesis will be includes the following chapters:

- **Chapter 01:** Large language Model (LLMs): introduces the foundational concepts of deep learning, the evolution of neural networks, and the architectural innovations that have enabled the rise of LLMs, such as the transformer model and self-attention mechanisms.
- **Chapter 02:** Networking technologies and optimization: The chapter positions these models in networking technologies by analyzing 5G current status and 6G potential while discussing network slicing and intelligent edge computing and AI-based anomaly detection and real-time optimization.
- **Chapter 03:** Application of LLMs in 3GPP Specifications: The third chapter connects theoretical concepts to practical uses by studying LLM implementations in

real-world operations at Cisco ,Microsoft as well as Huawei. Hence reviews research on datasets like SPEC5G ,TSpec-LLM and frameworks like Retrieval-Augmented Generation (RAG), demonstrating how LLMs can improve the comprehension and automation of dense technical documentation such as 3GPP specifications.

- **Chapter 04:** Implementation and results analysis: presents the practical implementation of LLM-powered chatbots trained on telecom-specific datasets. It evaluates the performance of different RAG architectures and discusses development tools such as LangChain, HuggingFace, and Streamlit.

# Chapter 1

# Large language Model (LLMs)

## 1.1 Introduction

This chapter explores the fundamental aspects of **L**arge **L**anguage **M**odels (LLMs), tracing their evolution and significance in artificial intelligence. It begins with an overview of deep learning fundamentals, including key methodologies and neural network architectures, before delving into LLMs such as GPT, BERT, and DeepSeek. The chapter also examines the critical role of transformers, highlighting their self-attention mechanisms and widespread applications in **N**atural **L**anguage **P**rocessing (NLP) and **C**omputer **V**ision (CV).

## 1.2 Deep learning fundamentals

Deep Learning, a subset of Machine Learning, is inspired by the information processing patterns found in the human brain. DL does not require any human-designed rules to operate, rather, it uses a large amount of data to map the given input to specific labels. DL is designed using numerous layers of algorithms (artificial neural networks), each of which provides a different interpretation of the data that has been fed to them. Achieving the classification task using conventional ML techniques requires several sequential steps, specifically pre-processing, feature extraction, wise feature selection, learning, and classification.

### 1.2.1 Introduction to deep learning and its significance in modern AI

Deep learning has significantly advanced modern AI systems, particularly in NLP and CV. In NLP, deep learning techniques, such as **R**ecurrent **N**eural **N**etworks (RNNs) and transformers, have enhanced tasks like machine translation, sentiment analysis, and text generation, enabling machines to better understand and produce human language [79,

41, 53]. For instance, models like BERT and GPT have revolutionized language comprehension and generation, facilitating applications in chatbots and virtual assistants [17, 41]. In the realm of CV, deep learning has driven innovations in object detection, image recognition, and semantic segmentation, impacting industries such as healthcare and autonomous vehicles [15, 17]. The integration of these technologies not only improves efficiency but also addresses complex real-world challenges, highlighting the transformative potential of deep learning across various sectors [15, 17].

### 1.2.2 Key techniques and methodologies

Key techniques and methodologies in deep learning encompass a variety of approaches that leverage artificial neural networks to achieve state-of-the-art results across numerous applications. Deep learning encompasses a variety of techniques and methodologies that have significantly advanced the field of artificial intelligence.
Key techniques in deep learning

- Neural Networks: Fundamental structures that mimic human brain functions, allowing for complex data analysis [73].
- **C**onvolutional **N**eural **N**etworks (CNNs): Specialized for processing visual data, CNNs excel in identifying patterns and features in images[73, 72].
- RNNs: Designed for sequential data, RNNs are effective in tasks like language modeling and time series prediction [59].
- **G**enerative **A**dversarial **N**etworks (GANs): These networks generate synthetic data by pitting two neural networks against each other, enhancing data generation capabilities [59, 72].

Advanced methodologies

- Transfer Learning: This technique allows models to leverage knowledge from pretrained networks, improving efficiency and performance in new tasks [59].
- Attention Mechanisms: Enhancing model focus on relevant parts of the input data, attention mechanisms have revolutionized tasks in natural language processing [72].
- Federated Learning: A method that enables decentralized model training, enhancing privacy and security in data handling [59].

## 1.3 Neural network architectures

Neural network architectures are complex systems inspired by the human brain, designed to process information through interconnected nodes known as artificial neurons. These

architectures vary significantly based on their intended applications, ranging from simple models like perceptrons to advanced structures such as CNNs and RNNs. The following sections outline the key components and functionalities of neural network architectures.

*Key components*

- Neurons and Layers: Neural networks consist of layers of neurons, including input, hidden, and output layers, where each neuron processes input and passes it to the next layer.

In a regular neural network, there are three types of layers:

- Input Layers: It's the layer in which we give input to our model. The number of neurons in this layer is equal to the total number of features in our data (number of pixels in the case of an image).
- Hidden Layer: The input from the Input layer is then fed into the hidden layer. There can be many hidden layers depending on our model and data size. Each hidden layer can have different numbers of neurons which are generally greater than the number of features. The output from each layer is computed by matrix multiplication of the output of the previous layer with learnable weights of that layer and then by the addition of learnable biases followed by activation function which makes the network nonlinear.
- Output Layer: The output from the hidden layer is then fed into a logistic function like sigmoid or softmax, which converts the output of each class into the probability score of each class.

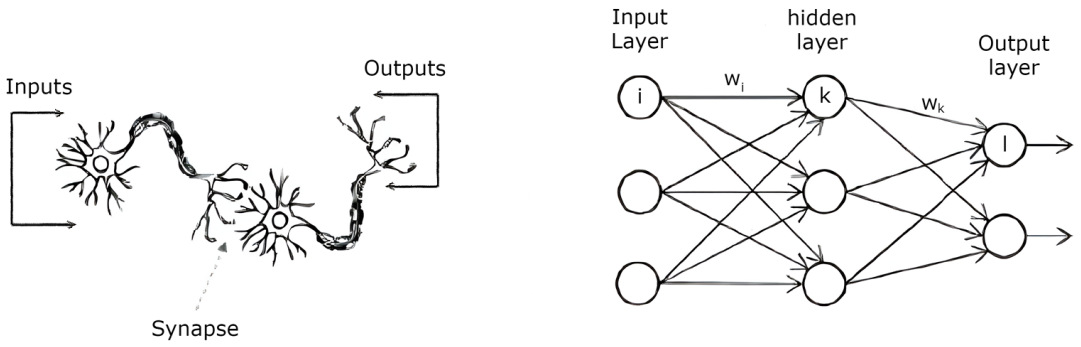


Figure 1.1: Artificial neural network architecture

- Activation Functions: These functions introduce non-linearity into the model, allowing it to learn complex patterns. Common examples include **R**ectified **L**inear **U**nit (ReLU) and sigmoid functions.
- Connections: Neurons are interconnected through weighted connections, which are adjusted during training to minimize.

*Functionalities*

- Feed-Forward and Feedback Mechanisms: Feed-forward networks allow unidirectional signal flow, while feedback networks enable bidirectional communication, enhancing learning capabilities.
- Learning Algorithms: Techniques like back-propagation are employed to optimize weights and biases, ensuring the network learns effectively from data.
- Specialized Architectures: Variants like CNNs excel in image processing, while RNNs are suited for sequential data tasks.

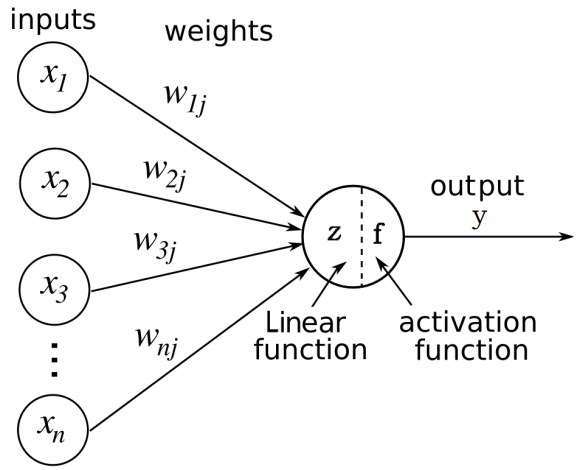


Figure 1.2: Computational Process of an Artificial Neuron in Neural Networks

### 1.3.1 Overview of neural network architectures (ANNs, CNNs, RNNs)

Neural network architectures, including **A**rtificial **N**eural **N**etworks (ANNs), CNNs, and RNNs, each serve distinct purposes and exhibit unique characteristics.

- ANNs are foundational models that consist of interconnected nodes, suitable for various tasks, including classification and regression.
- CNNs excel in processing grid-like data, particularly images, by utilizing convolutional layers to capture spatial hierarchies.
- RNNs, on the other hand, are designed for sequential data, making them ideal for tasks such as time series prediction and natural language processing.

*The specifics of each architecture:*

**Artificial neural networks (ANNs):**
ANNs are a type of machine learning model that are inspired by the structure and function of the human brain. Composed of input, hidden, and output layers. Suitable for tasks like classification and regression. Can be applied in various domains, including finance and healthcare [4].

**Convolutional neural networks (CNNs):**
These networks are designed to process data with a grid-like topology, such as images. The layers consist of convolutional layers, which learn to detect specific features in the data, and pooling layers, which reduce the spatial dimensions of the data.
1. Convolutional Layer
This layer is the first layer that is used to extract the various features from the input images. In this layer, the mathematical operation of convolution is performed between the input image and a filter of a particular size MxM. By sliding the filter over the input image, the dot product is taken between the filter and the parts of the input image with respect to the size of the filter (MxM).
2. Pooling Layer
In most cases, a Convolutional Layer is followed by a Pooling Layer. The primary aim of this layer is to decrease the size of the convolved feature map to reduce computational costs. This is performed by decreasing the connections between layers and independently operating on each feature map. Depending upon the method used, there are several types of Pooling operations.
3. Fully Connected Layer
The Fully Connected (FC) layer consists of the weights and biases along with the neurons

and is used to connect the neurons between two different layers. These layers are usually placed before the output layer and form the last few layers of a CNN Architecture.In this, the input image from the previous layers is flattened and fed to the FC layer. The flattened vector then undergoes a few more FC layers where the operations of the mathematical function usually take place. In this stage, the classification process begins to take place.
Utilize convolutional layers to detect patterns in spatial data. Commonly used in image recognition and CV tasks (cv is a field of Artificial Intelligence that enables a computer to understand and interpret the image or visual data)Demonstrated superior performance in SQL injection detection compared to ANNs and RNNs [4].

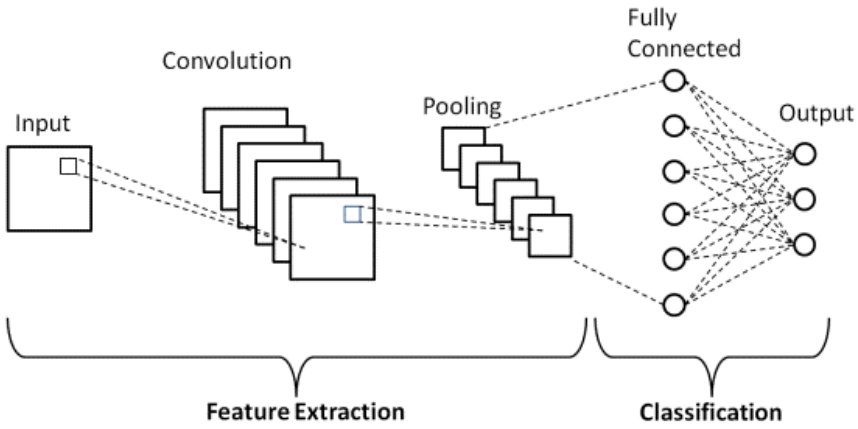


Figure 1.3: CNNs Architecture for Feature Extraction and Classification

**Recurrent neural networks (RNNs):**
These networks have a "memory" component, where information can flow in cycles through the network. This allows the network to process sequences of data, such as time series or speech. Designed to handle sequential data by maintaining a memory of previous inputs. Effective in applications such as language modeling and time series analysis. Showed competitive results in SQL injection detection, following CNNs [4].
The defining feature of RNNs is their hidden state, which preserves essential information from previous inputs in the sequence. By using the same parameters across all steps, RNNs perform consistently across inputs, reducing parameter complexity compared to traditional neural networks. This capability makes RNNs highly effective for sequential tasks.

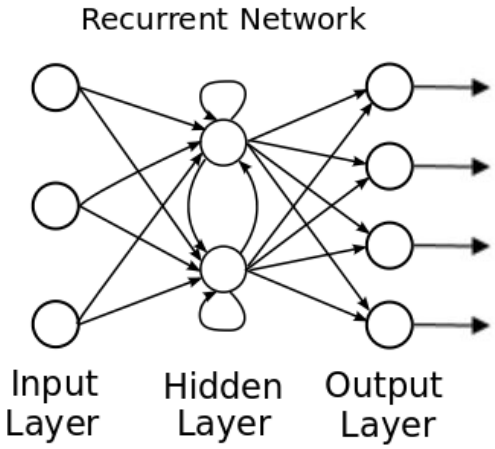


Figure 1.4: RNNs Architecture for Sequential Data Processing

### 1.3.2 Role of architecture in model performance

The architecture of neural networks plays a crucial role in determining their performance, impacting aspects such as efficiency, generalization, and computational resource requirements. The design and optimization of neural network structures are essential for achieving optimal performance across various tasks and datasets. This involves considerations of network depth, width, and dynamic adjustments, as well as the use of **N**eural **A**rchitecture **S**earch (NAS) to automate the discovery of effective architectures.

*Neural architecture search (NAS)*

- NAS is a method used to automatically find optimal neural network architectures, which is particularly useful in complex environments like 6G networks where performance, energy consumption, and data availability are critical factors [55].
- It reduces the manual effort in designing architectures and can adapt to multi-objective scenarios, such as balancing performance with resource constraints[75].

*Network properties and performance*

- The performance of neural networks is closely linked to their structural properties, such as size and configuration. Smaller networks can be used to optimize hyperparameters, which can then be applied to larger networks [26].
- A mathematical framework has been developed to predict model performance based on early training results, using a neural capacitance metric that captures generalization capabilities [34].

*Structural optimization strategies*

- Optimization strategies include adjusting network depth and width, modular design, and model compression techniques like weight pruning and quantization to enhance performance and efficiency [75].
- Different architectures, such as CNNs, RNNs, and Transformers, have specific advantages depending on the task, and their performance can be further improved through structural optimization [75].

*Diversity and ensembling*

- Diverse network architectures can achieve similar performance levels, which is beneficial for ensembling. Varying network sizes and activation functions can promote diversity, leading to better ensemble performance [29].

## 1.4 Models

### 1.4.1 Definition and evolution of LLMs (e.g.,DeepSeek ,GPT, BERT, Llama)

*Large language models:*
Large language models are sophisticated deep learning architectures trained on large amounts of unlabeled and self-supervised data, enabling them to generate human-like text and respond to complex queries. They predict what word comes next Instead of predicting one word with certainty, though, what it does is assign a probability to all possible next words. In contrast to traditional language models, which often rely on simpler statistical methods and smaller datasets, LLMs leverage extensive training to understand and produce language that mimics human communication. This distinction highlights their transformative potential across various fields.

*The evolution of Large Language Models (LLMs):*
Significant milestones that reflect advancements in artificial intelligence and natural language processing have marked the evolution of LLMs. From their onset, LLMs have transitioned from simple rule-based systems to complex architectures capable of comprehending and generating human-like text. This evolution is defined by key developments in model architecture, training methodologies, and application domains.

- Historical milestones
  - Early AI Foundations: The journey began with the Dartmouth Conference in 1956, which laid the groundwork for AI research [16].
  - Shift to Deep Learning: The resurgence of AI in the 21st century, particularly through deep learning, enabled the development of sophisticated models like BERT and GPT [16, 57].
  - GPT-3 Launch: The introduction of GPT-3, with 175 billion parameters, exemplified the capabilities of LLMs in various applications, including healthcare and customer service [16].
- Architectural advancements
  - From RNNs to Transformers: The transition from recurrent neural networks (RNNs) to transformer models marked a significant leap in processing capabilities.
  - These methodologies have allowed LLMs to adapt across diverse natural language processing tasks, enhancing their performance and versatility [16, 74].

- Applications and impact
  - Diverse Use Cases: LLMs have transformed sectors such as finance, healthcare, and education, showcasing their ability to revolutionize traditional tasks [74].
  - Ethical Considerations: As LLMs advance, issues like bias and interpretability have emerged, necessitating ongoing research to address these challenges [69, 57].
- Evolution of LLMs:
  The transition from rule-based systems to transformer models marked a significant milestone in LLM development. Notable models include:
  - **Llama**: Large Language Model Meta AI is Meta's LLM developed by [70] represents a major advancement in language models, offering sizes from 7B to 65B parameters and trained on large-scale public datasets. Later versions, such as LLaMA-2 and the 405B-parameter LLaMA-3, have improved multilingual, coding, and reasoning abilities, achieving performance on par with GPT-4. LLaMA models also demonstrate strong results in specialized tasks, such as medical informatics, showcasing their versatility across domains.
  - **BERT**: stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (BERT). Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. Recent advancements integrate BERT with CNNs to enhance sentence representation and classification, leading to superior accuracy in sentiment analysis [23].
  - **GPT**: **G**enerative **P**re-trained **T**ransformers (GPT) models have been marked by significant advancements in their architecture, capabilities, and applications. The transition from GPT-2 to GPT-3 and subsequently to ChatGPT introduced more sophisticated dialog capabilities, allowing for nuanced responses that better emulate human understanding [46]. Fine-tuned GPT-like models can predict user evolution and network dynamics, addressing challenges in recommendation systems [31].


Figure 1.5: OpenAI model logo

– **DeepSeek**: The model is fine-tuned with a diverse dataset to improve user experience in practical applications, achieving competitive performance across visual-language benchmarks [44]. Focused on advancing theorem proving, DeepSeek-Prover utilizes large-scale synthetic data to enhance mathematical reasoning in LLMs. The model demonstrates superior performance in generating proofs, surpassing GPT-4 and other baseline methods in formal theorem proving tasks [76].


Figure 1.6: Deepseek model logo

### 1.4.2 Pre-training and fine-tuning of LLMs

*Pre-Training:*

Pre-training of LLMs is a critical phase that significantly influences their performance and capabilities. This process involves training models on vast datasets to learn language patterns, structures, and knowledge before fine-tuning them for specific tasks. The pre-training phase is essential for equipping LLMs with a foundational understanding of language, which can be further refined through fine-tuning.

*Key Aspects of Pre-Training:*

1. **Data collection and Preparation**
   The scale and quality of pre-training data are crucial for the success of LLMs. High-quality, diverse datasets enable models to generalize better and perform a wide range of tasks [36]. Pre-training requires a dataset collected from the variants resources such as articles, websites, etc. Moreover preprocessing of this obtain data include:

   - Tokenization: This step involves breaking the text into smaller units.
   - Standardization: Converts all characters to a consistent format.
   - Stop-word Removal: Eliminates common, frequently occurring words that typically add little value to the analysis, such as articles and prepositions.
   - Stemming and Lemmatization: reducing words to their root forms.

2. **Model architecture**

- Most LLMs use the transformer architecture, which relies on self-attention mechanisms to capture relationships between words in a sentence. This is designed to scale with more parameters, enabling the model to learn more patterns that are complex.

3. **Training Objectives**

   - **M**asked **L**anguage **M**odeling (MLM): Used in models like BERT, where random tokens in the input are masked, and the model predicts the masked tokens based on context.
   - **C**ausal **L**anguage **M**odeling (CLM): Used in models like GPT, where the model predicts the next token in a sequence, learning to generate coherent text.

4. **Optimization Techniques**

   - Gradient Descent: The model is trained using variants of gradient descent, such as Adam or AdamW, to minimize the loss function.
   - Learning Rate Scheduling: Techniques like warm-up and decay are intended for adjust the learning rate during training for better convergence.
   - Batch Size and Parallelism: Large batch sizes and distributed training across multiple GPUs/TPUs are intended for speed up training.

5. **Transfer Learning**
   After pre-training, the model is fine-tuned on specific tasks or domains, leveraging the general knowledge learned during pre-training to achieve high performance with less task-specific data.

*Fine-tuning:*

Fine-tuning refers to taking a pre-trained model and training at least one model parameter (the internal weights or biases inside the neural network). Fine-tuning a model involves adjusting the parameters of a pre-trained model in order to make it better suited for a given task. There are generally three steps to fine-tuning a model:

1. Select a base model: select a pre-trained deep learning model that has been trained on a large dataset.

2. Adjust parameters: adjust parameters of the pre-trained model to better suit the desired task. This may include changing the number of layers adjusting learning rate, adding regularization, or tweaking the optimizer.

3. Train the model: train the new model on the desired dataset. The amount of data and the amount of training required will depend on the task and the model.

Fine-tuning Large Language Models (LLMs) involves several key factors that significantly influence their performance. These factors include the choice of fine-tuning techniques, the nature and amount of data used, and the optimization strategies employed. Understanding these elements is crucial for enhancing the effectiveness of LLMs in various applications.

### *Fine-tuning techniques*

- Low-Rank Adaptation: fine-tunes model by adding new trainable parameters, this method reduces the number of parameters needed for fine-tuning, but it can be sensitive to hyperparameter settings. Techniques like MonteCLoRA improve stability and accuracy by employing Bayesian reparameterization[65].

- Semantic Knowledge Tuning: This approach utilizes meaningful tokens instead of random ones, leading to faster training and better performance on specific tasks [58].

### *Data considerations*

- **S**upervised **F**ine-**T**uning (SFT): The amount and type of data significantly affect performance. Research indicates that as few as 60 data points can activate pre-trained knowledge for question-answering tasks [80].

    1. Choose fine-tuning task: this could be text summarization, text classification, and binary classification.
    2. Prepare training dataset.
    3. Choose a base model.
    4. Fine-tune model via supervised learning.
    5. Evaluate model performance.

- Knowledge Retention: Fine-tuning on partially mastered knowledge can enhance learning while minimizing the risk of forgetting previously acquired information [43].

### *Optimization strategies*

Linear Chain Transformation: This method introduces multiple linear transformations during fine-tuning, improving the model's ability to learn complex representations and enhancing generalization [74].

### 1.4.3 Applications of LLMs in various fields (NLP, computer vision)

Large Language Models (LLMs) have emerged as powerful tools in both NLP and CV, showcasing their versatility across various applications. Their ability to process and generate human-like text, as well as their integration with visual data, opens up numerous possibilities in these fields.

*Applications in natural language processing:*

- Text Generation and Translation: LLMs like GPT-3 excel in generating coherent text and translating languages, leveraging vast datasets to understand context and nuances [69].
- Question Answering: They can provide accurate answers to complex queries by synthesizing information from diverse sources [69].

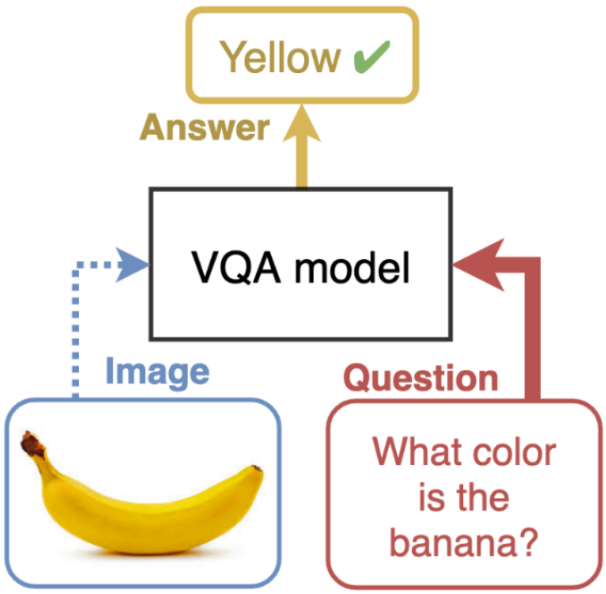


Figure 1.7: How LLMs Assist in Answering User Queries

- Neuroscience Insights: Research indicates that larger LLMs can predict neural activity related to language processing, enhancing our understanding of human cognition [30].

*Applications in computer vision:*

- Vision-Language Models: Large Language and Vision Models combine visual and textual data, enabling tasks such as image captioning and visual question answering [85].

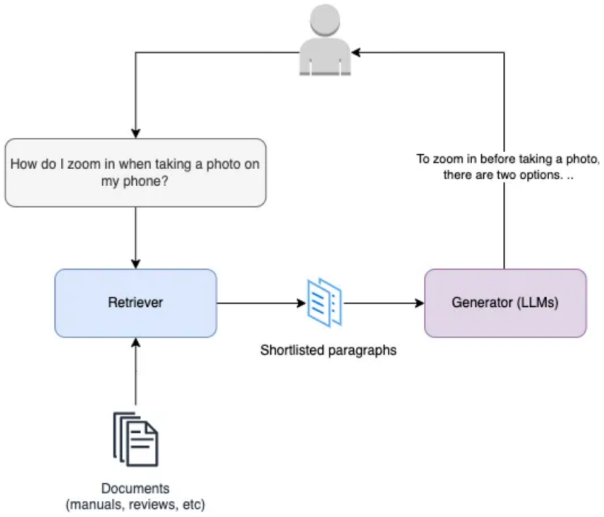


Figure 1.8: A VQA example where the model answers

- Perceptual Judgments: LLMs can correlate perceptual data with language, revealing insights into how language influences sensory perception [47].
- Cross-Modal Learning: Co-training on vision and language allows models to perform tasks that require understanding both modalities, although performance can vary based on the task complexity [47].

## 1.5 transformers

### 1.5.1 Introduction to the transformer architecture

The development of the Transformer architecture in NLP is marked by several key innovations that have significantly enhanced its performance and applicability. Central to this architecture is the Attention Mechanism, which allows for the effective extraction of relevant context from long sequences, addressing a critical challenge in NLP tasks. This mechanism is complemented by the use of **M**ulti-**L**ayer **P**erceptrons (MLPs), although recent studies suggest that the attention mechanism alone may suffice for many applications, potentially reducing model complexity [13].

- Key characteristics of transformers
  - Self-Attention Mechanism: This allows the model to weigh the importance of different input elements, enabling it to process complex patterns efficiently[35].
  - Layer Normalization and Adaptive Optimizers: These features enhance the training stability and performance of transformers compared to traditional architectures like MLPs and CNNs [54].

  - Distributed Training: Transformers benefit from parallelism in multi-node and multi-GPU settings, although communication bottlenecks remain a challenge [6].

- Application of transformers

  - Natural Language Processing: Transformers are foundational to LLMs such as GPT and BERT, which have set new benchmarks in various NLP tasks [35].
  - Molecular Science: Their ability to handle intricate data has made transformers popular in molecular modeling and analysis [35].
  - Neuroimaging: Recent studies highlight the effectiveness of transformers in classification and regression tasks within neuroimaging, showcasing their versatility [86].

### 1.5.2 Self-attention mechanism and its importance in LLMs

The self-attention mechanism is a pivotal component of large language models (LLMs), this mechanism allows models to weigh the importance of different words in a sentence relative to each other, facilitating the capture of complex dependencies and relationships. The sense of self-attention can be discovered through its impact on memory capacity, efficiency, and the evolution of LLM architectures.

- Memory capacity

  - Self-attention influences the working memory capacity of LLMs, as evidenced by performance drops in N-back tasks when the number of positions increases[28].
  - The attention scores tend to aggregate towards specific positions, indicating a learned strategy to manage memory limits [28].

- Efficiency

  - The self-attention mechanism is a major source of latency in processing long sequences. Innovations like Chunk Attention optimize this by sharing key/value tensors across requests, significantly improving memory utilization and processing speed [81].
  - Efficient attention mechanisms are crucial for handling long contexts, as traditional methods face challenges due to their quadratic complexity [48].

- Evolution of LLMs

  - The self-attention mechanism has driven the evolution of transformer architectures, allowing for the modeling of intricate relationships and dependencies over extended distances, which is essential for modern AI applications [45].

### 1.5.3 Applications of transformers in LLMs

Transformers have revolutionized the application of LLMs across various domains, enhancing their capabilities in NLP and beyond. Their architecture, particularly the attention mechanism, allows for efficient handling of complex language tasks, long-context inputs, and In-Context Learning (ICL).

*Long-Context Handling*

- Transformers are increasingly adapted to manage long-context scenarios, addressing the quadratic complexity of traditional attention mechanisms [48].
- Innovations in transfer architectures have been developed to enhance long-context capabilities, enabling LLMs to process extensive text inputs effectively [32].

*In-Context Learning*

- Transformers exhibit strong ICL abilities, allowing them to perform new tasks using only prompts without additional fine-tuning [14].
- The multi-concept semantic representation in transformers supports their innovative problem-solving capabilities, enhancing their performance on unseen tasks [14].

*Efficiency and Scalability*

- The transformer architecture has led to significant advancements in machine translation, text summarization, and sentiment analysis, outperforming baseline models [77].
- New architectures, such as Tandem transformers, improve inference speed and accuracy by combining small and large models, demonstrating enhanced performance in real-time applications [3].

## 1.6 Conclusion

LLMs have transformed AI-driven applications, particularly in NLP and computer vision, by leveraging deep learning architectures and self-attention mechanisms. The advancements in pre-training, fine-tuning, and transformer models have significantly enhanced language understanding and generation. As LLMs continue to evolve, they offer promising potential for improving AI applications across various domains while raising ethical and computational challenges that require further research.

# Chapter 2

# Networking technologies and optimization

## 2.1 Introduction

This chapter provides an in-depth discussion of networking technologies, focusing on 5G and emerging 6G networks. It covers the key features, challenges, and applications of AI and machine learning in optimizing network performance. Additionally, security concerns in next-generation networks are addressed, along with the role of AI and LLMs in anomaly detection.

## 2.2 5G networks

### 2.2.1 Overview of 5G technology

5G technology represents a significant leap in mobile communication, offering transformative capabilities that extend beyond traditional network improvements. It is characterized by high-speed connectivity, low latency, and the ability to support a massive number of devices, which collectively enable a wide range of applications across various sectors. The integration of advanced technologies such as **N**etwork **F**unction **V**irtualization (NFV), **S**oftware-**D**efined **N**etworking (SDN), and edge computing further enhances the flexibility and efficiency of 5G networks. These advancements are crucial for supporting applications like **E**nhanced **M**obile **B**roadband (EMBB), **U**ltra-**R**eliable **L**ow-**L**atency **C**ommunication (URLLC), **M**assive **M**achine-**T**ype **C**ommunication (MMTC) [63, 33].

*Key Technological Advancements*

- NFV and SDN: These technologies enable more flexible and efficient network management, allowing for dynamic resource allocation and improved network performance [63].

- MIMO and Beamforming: **M**ultiple **I**nput **M**ultiple **O**utput (MIMO) and beamforming technologies enhance signal quality and network capacity, crucial for supporting high data rates and connectivity [67].

- Millimeter Wave Spectrum and Small Cells: Utilization of higher frequency bands and deployment of small cells increase network capacity and coverage, essential for urban environments [67, 60].

### *Applications and Use Cases*

- Smart Cities and IoT: 5G facilitates the development of smart cities and the **I**nternet of **T**hings (IoT) by providing the necessary infrastructure for real-time data exchange and device connectivity [62, 60].

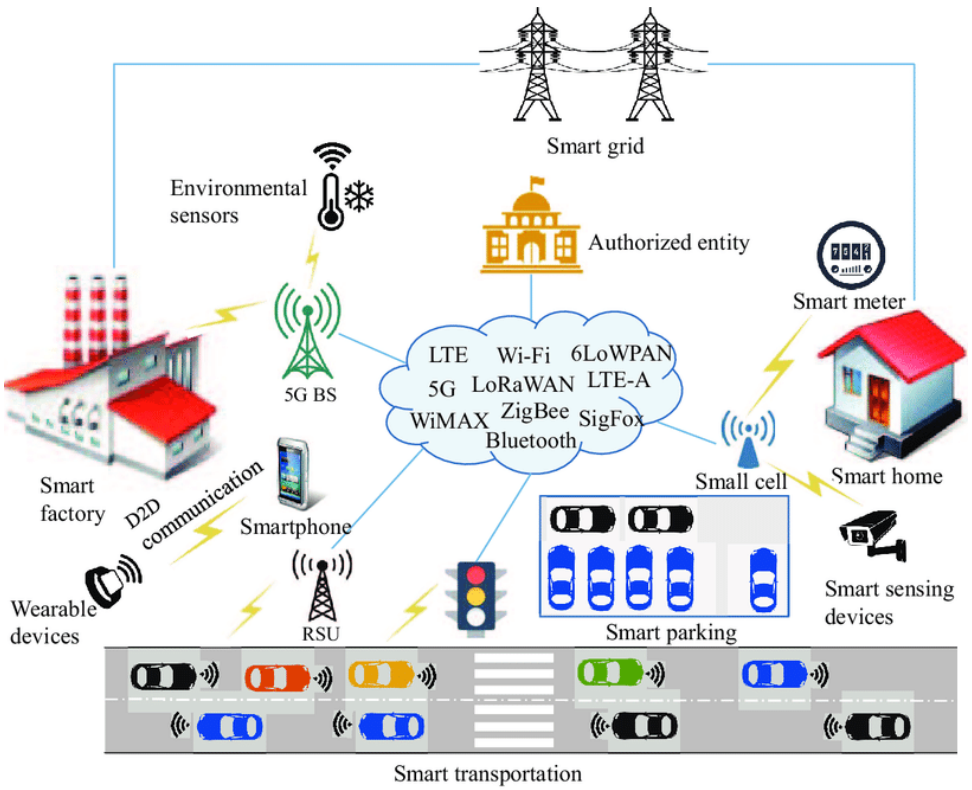


Figure 2.1: Smart city ecosystem diagram showing 5G-connected devices (sensors, cameras, vehicles).

- Healthcare and Remote Surgery: The low latency and high reliability of 5G networks enable applications such as remote surgery and telemedicine, improving healthcare accessibility and outcomes [60].

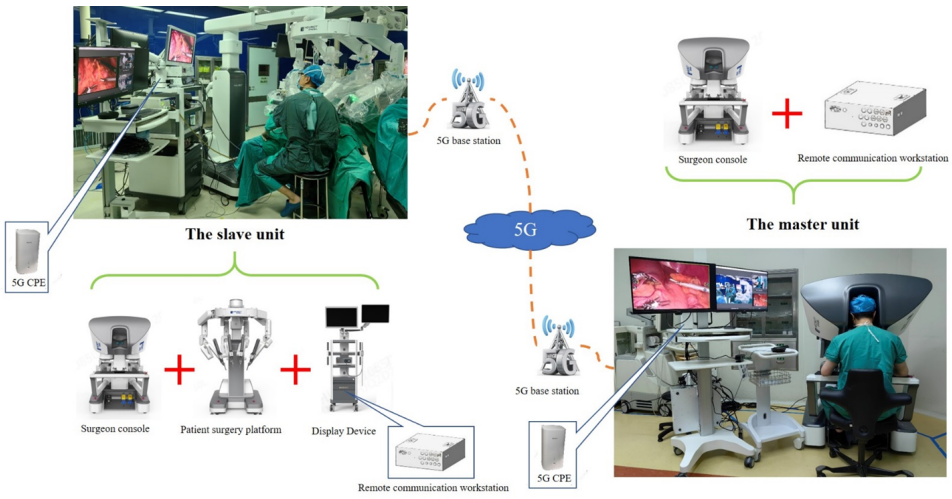


Figure 2.2: Robot-assisted remote radical distal gastrectomy with 5G technology

- Industry 4.0 and Automation: 5G supports industrial automation and smart manufacturing, driving efficiency and innovation in production processes [62].

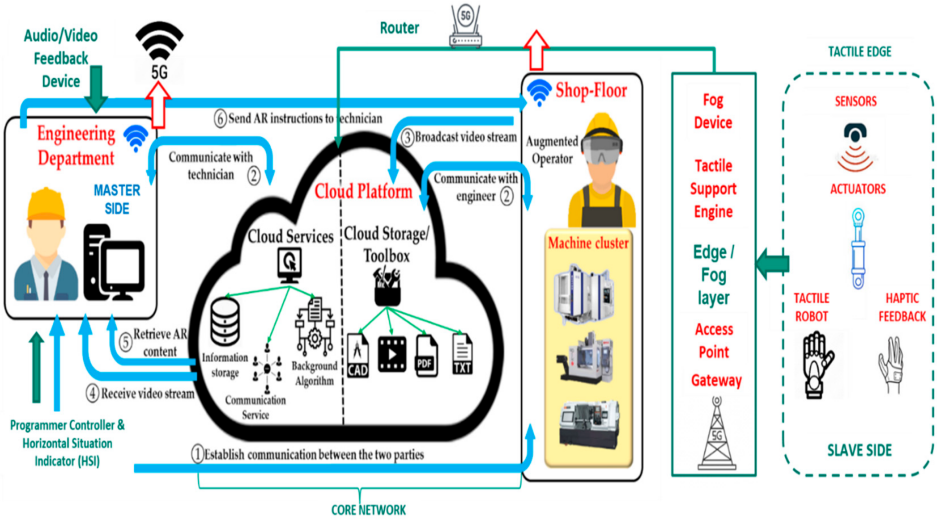


Figure 2.3: Industrial IoT (IIoT) Ecosystem Powered by 5G and Edge Computing

### 2.2.2 Key features and challenges

5G technology introduces a transformative era in communication, offering high-speed connectivity, low latency, and support for massive machine-type communications. These features enable a wide array of applications across various domains, including IoT, smart cities, and industrial automation. However, the deployment and integration of 5G networks present several challenges that need to be resolved to fully realize it is potential.

*Key Features of 5G*

- High-Speed Connectivity: 5G offers unprecedented throughput, which is crucial for applications requiring high-speed data transmission, such as IoT and real-time data processing in smart cities (traffic cameras, environmental sensors) [68, 71].

| | 3G | 4G | 5G |
|---|---|---|---|
| **Deployment** | 2004-05 | 2006-10 | 2020 |
| **Bandwidth** | 2 mb per second | 200 mb per second | >1 gb per second |
| **Latency** | 100-500 milliseconds | 20-30 milliseconds | <10 milliseconds |
| **Average speed** | 144 kb per second | 25 mb per second | 200-400 mb per second |

Table 2.1: Comparing 5G Network Performance to Earlier Wireless Generations

- Low Latency: The ultra-low latency of 5G networks supports real-time applications, enhancing fields like healthcare (surgeons control robots in real time), autonomous vehicles (V2X communication for collision avoidance), and industrial automation (precise synchronisation in factories)[56].

- Massive Connectivity: 5G can support a vast number of simultaneous connections, making it ideal for IoT environments (predictive maintenance, asset tracking) and smart city applications such as smart agriculture (soil sensors, drone monitoring).[71].

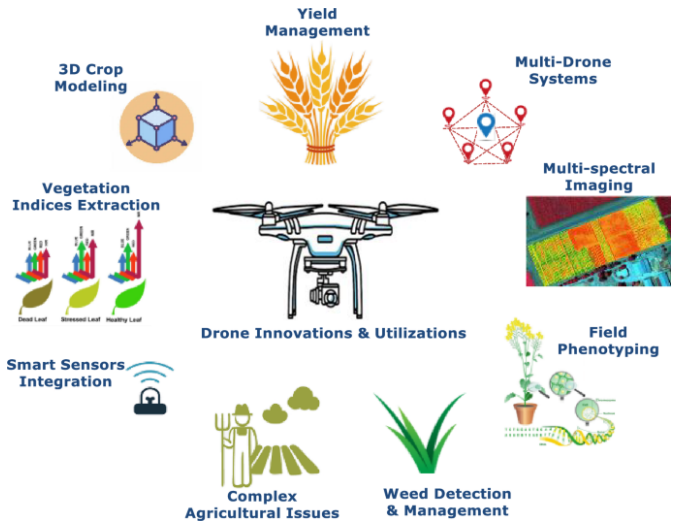


Figure 2.4: Smart Agriculture Ecosystem: AI-Driven Crop Monitoring and Precision Farming

- Network Slicing: This feature allows for the creation of multiple virtual networks within a single physical 5G network, optimizing resource allocation for different applications [20].

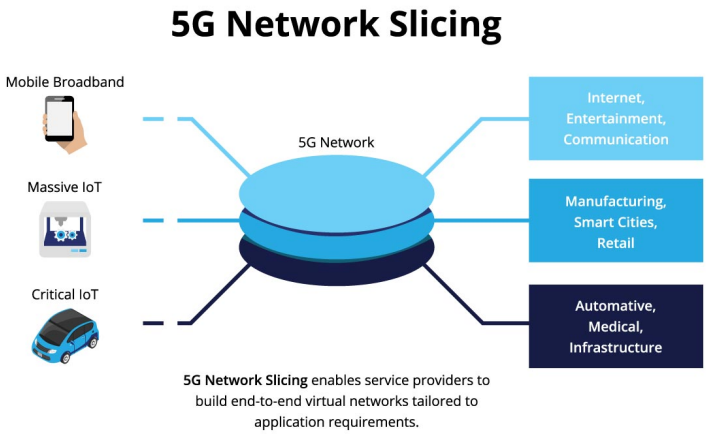


Figure 2.5: 5G Network Slicing

*Challenges of 5G*

- Security Vulnerabilities: Despite enhanced security features, 5G networks face threats such as DDoS attacks, man-in-the-middle (MITM) exploits, and side-channel attacks, necessitating robust security measures [68, 20].
- Infrastructure Requirements: The deployment of 5G requires significant infrastructure changes, including the availability of high-frequency spectrum and high deployment costs [62].
- Interoperability Issues: Integrating 5G with existing networks and IoT devices presents interoperability challenges, requiring standardized protocols for seamless operation [56].
- Network Management: Managing network congestion and resource allocation becomes increasingly complex with the growing number of connected devices [56].

## 2.3 6G networks: the future of wireless communication

### 2.3.1 Introduction to 6G and its expected advancements over 5G

The transition from 5G to 6G represents a significant leap in wireless communication technology, promising enhancements in speed, latency, and connectivity. 6G aims to address the limitations of 5G by integrating advanced technologies such as Artificial Intelligence (AI), IoT, and edge computing to create a more intelligent and efficient network ecosystem. This evolution is expected to support new applications like immersive

mixed-reality experiences, holographic communications, and smart city infrastructures. However, the shift to 6G also introduces challenges, particularly in terms of security and privacy, due to the increased integration of IoT devices and AI-driven analytics. Below are the key advancements expected in 6G over 5G:

*Enhanced Network Capabilities*

- Faster Data Rates and Lower Latency: 6G is anticipated to offer significantly higher data rates and near-zero latency, enabling seamless connectivity and real-time applications [78, 40].
- Higher Capacity: The network will support a larger number of connected devices, facilitating the growth of IoT and smart city applications [78, 22].

*Integration of Advanced Technologies*

- Artificial Intelligence: AI will play a crucial role in managing networks, predicting maintenance needs, and optimizing resource allocation, leading to more intelligent and responsive network operations [84, 40].
- Edge Computing: Mobile edge computing will provide computing, storage, and networking resources closer to end-users, ensuring efficient operations for latency-sensitive applications [5].

*New Applications and Use Cases*

- Immersive Experiences: 6G will enable new applications such as virtual and augmented reality, holographic communications, and smart healthcare, enhancing user experiences and service quality [22].
- Industry 5.0: The integration of 6G with Industry 5.0 will drive advancements in smart farming, drones, and smart grids, leveraging high data rates and low latency for improved industrial automation [22].

### 2.3.2 Vision and requirements for 6G

The vision for 6G networks is characterized by a transformative leap in mobile communication, driven by the need for extreme connectivity, inclusivity, and flexibility. This vision is underpinned by emerging use cases and key value indicators that demand a comprehensive end-to-end architecture. The 6G vision also emphasizes the integration of digital twins and ubiquitous intelligence, which are expected to revolutionize various sectors through scenarios like intelligence-shared living and intelligence-empowered production. These advancements necessitate a robust set of requirements to ensure the successful deployment and operation of 6G networks.

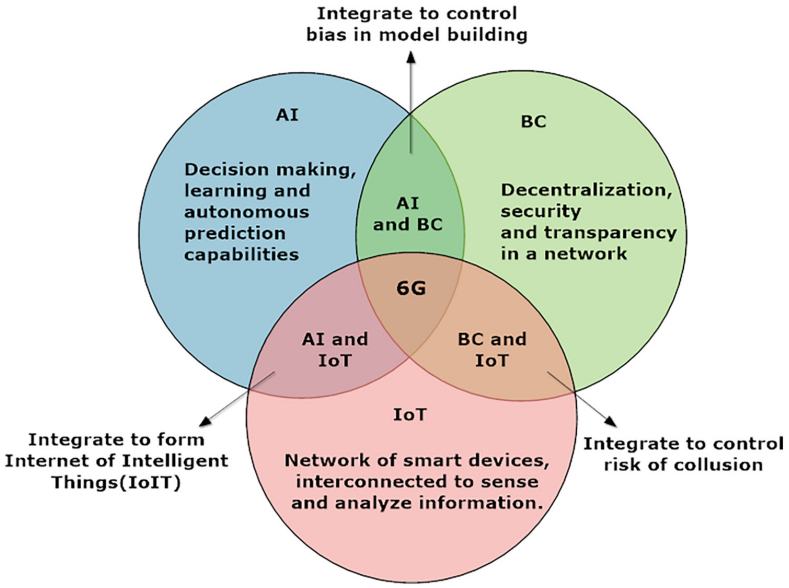


Figure 2.6: 6G Vision: How AI, Blockchain and IoT Create Intelligent, Secure Networks

- Key Requirements for 6G
  - Extreme Connectivity and Flexibility: 6G networks must support a wide range of applications, including holographic telepresence and immersive communication, which require high performance, reliability, and ubiquity across the edge-cloud continuum [39, 19].
  - Security and Trustworthiness: With the increase in security threats and complex use cases, 6G architecture must enhance trustworthiness and security measures compared to previous generations[39].
  - Sustainability: Sustainability is a critical design consideration, ensuring that 6G networks are environmentally friendly and resource-efficient [39].
  - AI/ML Integration: AI and machine learning are pivotal for optimizing network performance and enabling automation, particularly through **D**istributed **A**rtificial **I**ntelligence (DAI) [19].
  - Spectrum Management: The exploration of new spectrum and refarming of underutilized spectrum are essential to accommodate the expansive range of 6G applications[50].
- Vision for 6G
  - Intelligence-Driven Scenarios: 6G envisions scenarios such as intelligence-shared living and intelligence-enlightened societies, facilitated by digital twins and ubiquitous intelligence [83].

– Innovative Applications: Promising applications include AI, sensing, autonomous vehicles, urban air mobility, extended reality, and digital twins, which are expected to redefine customer experiences[50].

## 2.4 Security and anomaly detection

### 2.4.1 Importance of security in 5G/6G networks

Security in 5G and 6G networks is of paramount importance due to the critical role these networks play in modern communication infrastructures. As these networks evolve, they introduce new services and technologies that, while beneficial increase the potential attack surface for cyber threats. The transition from hardware-based to software-based systems in 5G, and the anticipated advancements in 6G, necessitate robust security measures to protect against increasingly sophisticated cyberattacks. This is especially crucial as these networks underpin essential services and critical infrastructures globally. The following sections delve into the key aspects of security in 5G and 6G networks.

*Cybersecurity Challenges and Solutions*

The shift to software-based **V**irtualized **N**etwork **F**unctions (VNFs) in 5G increases cybersecurity risks, necessitating enhanced security protocols like **P**ost-**Q**uantum **C**ryptography (PQC) to protect against quantum attacks [64].

The DMRN Protocol in 6G aims to provide Perfect Forward Secrecy and PQC, addressing vulnerabilities in mutual authentication and key exchange, which are critical for secure communications [82].

*Security Tools and Protocols*

MECHATRON, a security tool enabled by **M**ultiaccess **E**dge **C**omputing (MEC), offers comprehensive security for network assets and services, emphasizing continuous monitoring[12].

Blockchain-based security management in 6G networks, combined with machine learning techniques, enhances network optimization and security, achieving high throughput and energy efficiency [18].

*Regulatory and Collaborative Efforts*

The SAND5G project highlights the importance of operational collaboration among stakeholders and alignment with European cybersecurity policies to secure 5G and future 6G networks [9].

### 2.4.2 Role of AI and LLMs in anomaly detection

AI and LLMs play a significant role in anomaly detection across various domains, leveraging their capabilities in understanding complex data patterns and enhancing detection methodologies. These models are particularly useful in environments where traditional methods struggle due to data constraints or the need for domain-specific expertise. The integration of LLMs into anomaly detection frameworks offers promising advancements in scalability, adaptability, and accuracy. Below are key aspects of their role in anomaly detection:

- Automation and Scalability: LLMs can automate the extraction of physical invariants from **C**yber-**P**hysical **S**ystems (CPS) documentation, reducing the need for manual, domain-specific expertise. This approach enhances scalability and cost-effectiveness in anomaly detection by integrating these invariants into training datasets, thereby improving model reliability and precision [2].

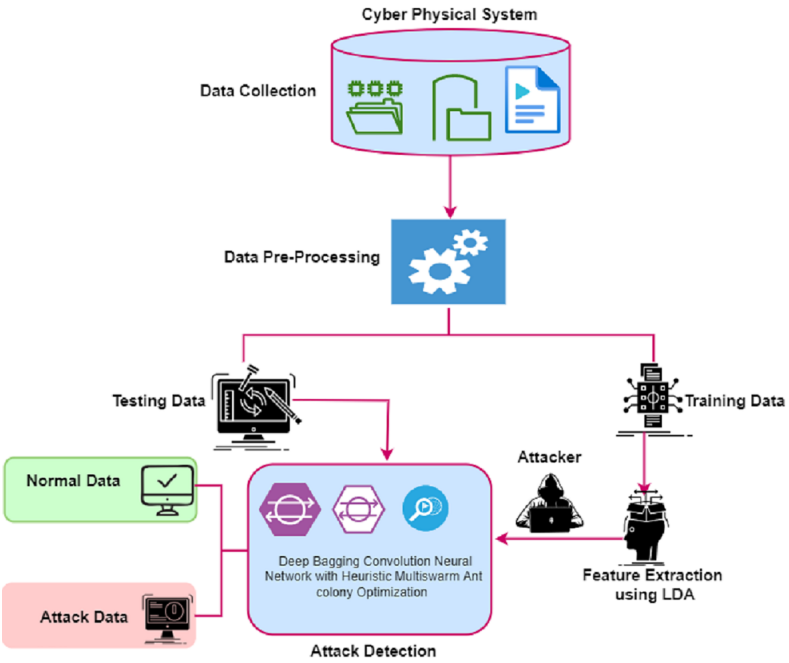


Figure 2.7: LLM automates invariant extraction from CPS documents to enhance anomaly detection models.

- Transferability and Multimodal Capabilities: In complex industrial environments, LLMs enhance the transferability of anomaly detection models. They enable the conversion of anomaly detection into a ”language” task, allowing for context-aware detection in data-sparse applications. This adaptability is crucial for addressing concept drift in dynamic settings such as LLM analyzing audio + text logs from a CNC machine to detect tool wear. [61].

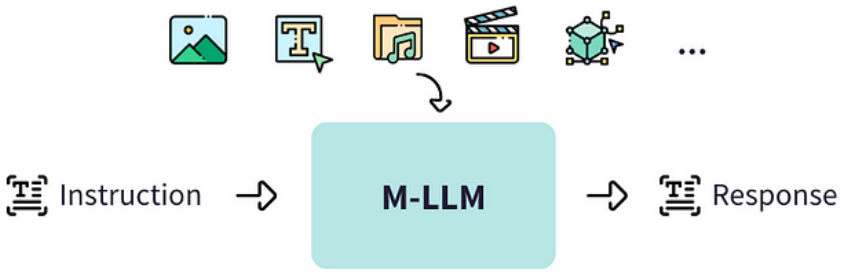


Figure 2.8: Multimodal LLMs

- Time Series and Tabular Data: LLMs, through prompt strategies like in-context learning, can detect anomalies in time series data, showing competitive results with baseline methods. Fine-tuning on synthesized datasets further improves their performance in these tasks [24]. Additionally, LLMs serve as zero-shot batch-level anomaly detectors for tabular data, identifying outliers without specific model fitting [42].

- Financial Data Encoding: In financial anomaly detection, LLM embeddings enhance the analysis of non-semantic categorical data, improving the performance of machine learning models in identifying irregular journal entries. This approach addresses feature sparsity and complexity in financial records [8].

## 2.5 Network optimization

### 2.5.1 Overview of network optimization techniques

Network optimization techniques are crucial for enhancing the performance, efficiency, and reliability of various network systems. These techniques are applied across different domains, including deep learning models, network-on-chip architectures, wireless body area networks, and general network settings. Each domain employs specific strategies tailored to its unique challenges and requirements. Below is an overview of key network optimization techniques derived from the provided papers.

- Deep Learning Model Optimization

  - Pruning and Quantization: These techniques reduce the size and complexity of deep neural networks by removing redundant parameters and reducing precision, respectively, without significantly affecting performance[21].

  - Model Distillation: This involves training a smaller model to mimic a larger model, thereby achieving similar performance with reduced computational resources [21].

  - Layer Fusion and Parallelization: These methods enhance computational efficiency by combining layers and executing operations concurrently [21].

- Network-on-Chip (NoC) Optimization
  - Fault-Tolerant Routing Algorithms: These algorithms improve the resilience of NoC systems by ensuring reliable data transmission even in the presence of faults [38].
  - Dynamic Voltage Scaling: This technique adjusts the voltage and frequency of NoC components to optimize energy efficiency[38].
  - Wireless NoC Congestion Control: Strategies are employed to manage data traffic and reduce congestion, thereby improving network latency and throughput [38].
- Wireless Body Area Network (WBAN) Optimization
  - Energy Efficiency and Network Lifetime: Techniques focus on minimizing power consumption and extending the operational life of the network [27].
  - Routing Protocol Optimization: Various methods are used to enhance data transmission reliability and scalability in WBANs [27].
- General Network Optimization Techniques
  - Convex Optimization: A mathematical approach used to find the optimal solution for network resource allocation problems [11].
  - Drift and Mean-Field Methods: These are used to analyze and optimize network performance by modeling and predicting network behavior [11].

### 2.5.2 Role of AI and machine learning in optimizing network performance

AI and machine learning play a crucial role in optimizing network performance by enhancing traffic management, reducing latency, and improving fault detection across various network types. These technologies leverage advanced algorithms to analyze real-time data, predict network conditions, and dynamically adjust configurations to maintain optimal performance. The integration of AI and ML in network management is transforming traditional approaches, leading to more resilient and efficient networks.

- AI-Driven Network Traffic Optimization
  - AI-driven approaches in enterprise WANs utilize knowledge graphs and reinforcement learning to optimize traffic routing and detect faults proactively.

This method ensures low latency and high availability for critical applications by dynamically adjusting routes based on real-time telemetry and historical patterns [51].

  - In cloud networking, AI-based optimization techniques significantly reduce latency and bandwidth variation in LTE networks. Deep learning models predict network quality and enable real-time configuration changes, achieving up to 40% improvement in latency reduction [1].

- Enhancements in **D**ata **C**enter **N**etworks (DCN)
  Optical data center networks face challenges due to the high traffic demands of AI/ML applications. An optical intra-rack DCN architecture with **W**avelength **D**ivision **M**ultiplexing **A**ccess (WDMA) efficiently handles diverse traffic classes, achieving high bandwidth utilization and low latency, thus optimizing performance for AI/ML traffic [10].

- **D**igital **T**win **N**etworks (DTNs) and IoT

  - DTNs employ AI tools like ML, DL, and RL to enhance network performance, optimize latency, and improve energy efficiency. These virtual representations of physical networks provide refined recommendations for real-world challenges [66].

  - In IoT networks, AI/ML algorithms enhance real-time decision-making, energy efficiency, and data management, contributing to the overall performance and security of IoT applications [49].

## 2.6 Conclusion

The integration of AI and machine learning in networking has led to significant improvements in network performance, security, and optimization. While 5G has enabled faster and more reliable communication, the anticipated advancements of 6G promise even greater connectivity and intelligence. However, challenges such as security threats and infrastructure costs must be addressed to fully realize the potential of future networks.

# Chapter 3

# Application of LLMs in 3GPP Specifications

## 3.1 Introduction

The chapter investigates the implementation of Large Language Models (LLMs) to enhance network management and optimization and security within 3GPP standards for the telecom sector. The chapter examines major company deployments by Cisco, Microsoft, and Huawei while summarizing recent research on telecom-specific datasets and models and demonstrating hands-on experiments to evaluate LLMs for 5G protocol tasks.

## 3.2 Practical Deployments of LLMs in Network Management and Optimization

### 3.2.1 Cisco: AI Assistants and LLMs for Network Optimization

Cisco integrates AI assistants and generative AI, including LLMs, across its portfolio to enhance network management, automation, and security. The company focuses on **I**ntent-**B**ased **N**etworking (IBN), predictive analytics, leveraging its AI-powered Cisco AI Assistant and frameworks like **R**etrieval-**A**ugmented **G**eneration (RAG).

- Cisco AI Assistant: This tool uses generative AI to simplify network operations, assist with policy management, and provide actionable insights. For example, the Cisco AI Assistant for Security streamlines complex policy settings and supports IT decision-making.

- Network Automation with RAG: Cisco employs RAG to integrate structured network data (e.g., JSON) with LLMs, reducing hallucination risks and grounding responses in real-time network telemetry.cisco.com

- IBN: Cisco DNA Assurance uses AI/ML, including NLP and machine reasoning, to translate human intent into network policies, optimize performance. LLMs assist in processing natural language inputs for policy creation and analyzing network behavior for self-optimization.cisco.com

- Webex Contact Center: Cisco's WxAI team integrates LLMs with Amazon SageMaker Inference to optimize contact center operations. LLMs analyze call transcripts, perform sentiment analysis, and power intelligent virtual assistants, improving agent productivity and customer experience. aws.amazon.com

### 3.2.2 Microsoft: AI Assistants and LLMs for Network Optimization

Microsoft leverages its Azure AI platform and Copilot AI assistants, powered by LLMs, to optimize cloud and enterprise networks. The focus is on predictive maintenance, traffic management, and security, integrating LLMs with Azure's network telemetry and analytics.

- Azure Network Analytics: Microsoft uses LLMs within Azure Monitor and Azure Network Insights to analyze network telemetry, predict congestion, and optimize routing. azure.microsoft.com

- Security Optimization: Microsoft Sentinel, a cloud-native SIEM, uses LLMs to analyze network security logs, detect threats, and automate incident response. azure.microsoft.com

- Project Vegea: While primarily focused on datacenter efficiency, Microsoft's Project Vegea employs LLMs to optimize network resource allocation, reducing latency and energy consumption in Azure datacenters. This indirectly improves network performance for customers.

### 3.2.3 Huawei: AI Assistants and LLMs for Network Optimization

Huawei pioneers AI-driven network optimization through its **A**utonomous **D**riving **N**etwork (ADN) initiative and Pangu LLMs. The company integrates AI assistants like the RAN Intelligent Agent and Telecom Foundation Model to achieve self-optimizing, zero-touch networks.

- Digital Twins with LLMs: Huawei combines LLMs with digital twins to simulate network behavior, validate configurations, and optimize operations. This en-

sures zero-wait, zero-interruption networks for telecom and enterprise use cases. huawei.com

- Pangu Models: Huawei's Pangu LLMs, ranging from 1 billion to 100 billion parameters, are tailored for telecom tasks like parameter prediction and fault diagnosis. These models are deployed in ADN for Enterprises to guarantee zero service delays and network disruptions. huawei.com

## 3.3 Fine-Tuning Techniques for LLMs in Telecom Context

This section examines the application of fine-tuning large language models to 3GPP-related tasks. The following three main initiatives are discussed:

- **GOOD-SPEC5G: A 5G Protocol Dataset for LLM Benchmarking**
  - **Objective**: SPEC5G is a dataset for the analysis of natural language specification of 5G Cellular network protocol specification. The authors in [37] have used this dataset for security-related text classification and summarization. Security-related text classification can be used to extract relevant security-related properties for protocol testing. On the other hand, summarization can help developers and practitioners understand the high level of the protocol, which is itself a daunting task.
  - **Model/LLM was used**: The pretrained models uses in this article are: BERT5G, RoBERTa5G, XLNET5G.
  - **Dataset**: The dataset introduced is called Good-SPEC5G which includes:
    - Our original 134M Word training corpus.
    - 5GSum - Summarization Dataset.
    - 5GSC - Classification Dataset.
    - 5GSC Annotator Reasoning - Annotator Explanation for Subset of 5GSC.
- **Fine-Tune Your Own Llama 2 Model in a Colab Notebook**
  This article Fine-Tune Your Own Llama 2 Model discusses how to fine-tune a Llama 2 7b model using a Colab notebook, including necessary background on LLM training and fine-tuning, and custom parameters.
  - It provides a step-by-step guide on fine-tuning a Llama 2 7b model using a Colab notebook.
  - The article concludes by emphasizing the importance of high-quality datasets for fine-tuned models and their integration into LangChain and other architectures.

## 3.4 Retrieval-Augmented Generation (RAG) for Technical QA

The section presents RAG-based architectures that have been specifically tuned for 3GPP comprehension:

- **TSpec-LLM: Processing and Generating 3GPP Specifications**
  - **Objective**: In this paper[52], Nikbakht introduces TSpec-LLM, an open-source dataset covering all 3GPP releases (1999-2023), maintaining all original tables and formulas to preserve technical fidelity. Hence this dataset was designed to facilitate the training and fine-tuning of LLMs to provide engineers and researchers with an assistant model capable of comprehending and organizing 3GPP technical documents.
    To evaluate its efficacy, the authors first select a representative sample of 3GPP documents,create corresponding technical questions, and assess the baseline performance of various LLMs. We then incorporate a retrieval augmented generation (RAG) framework to enhance LLM capabilities by retrieving relevant context from the TSpec-LLM dataset.

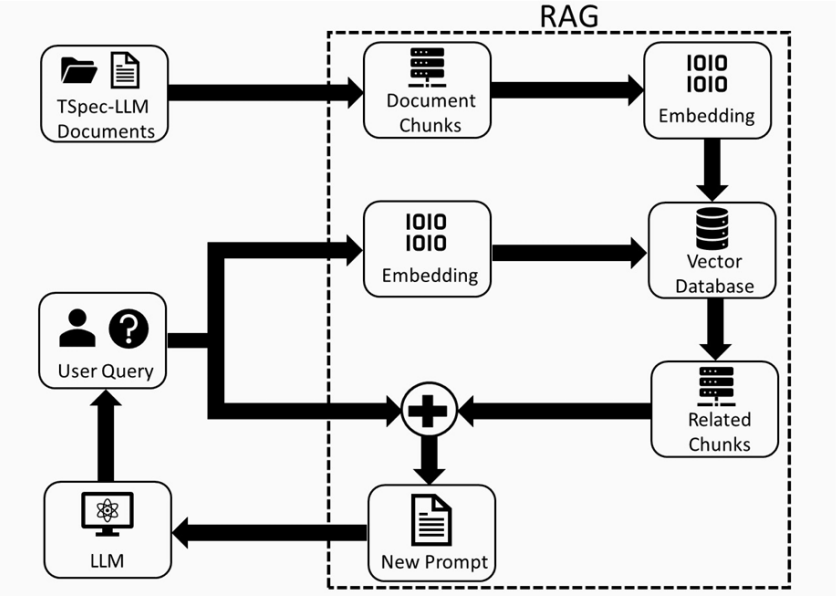


Figure 3.1: RAG-Based Workflow for Enhancing LLM Understanding of 3GPP Documents Using TSpec-LLM.

  - **Model/LLM was used**: GPT-3.5 ,GPT-4 and Gemini Pro 1.0 the researchers fine-tuned them with their proposed TSpec dataset to evaluate improvements in handling techenical spec tasks.
  - **Dataset**: the TSpec-LLM dataset, publicly available via Hugging Face at TSpec-LLM on Hugging Face, is structured into samples consisting of:

  - instruction: a prompt or question based on technical documentation.
  - input: optional context or supporting content.
  - output: the expected model response or answer.

- **Leveraging Fine-Tuned RAG with Long-Context Support for 3GPP Standards**
  - **Objective**: Fine-tuned retrieval-augmented generation RAG system based on the Phi-2 small language model (SLM) to serve as an oracle for communication networks using TeleQnA dataset.[25]
  - **Model used**: In this work they used a SLM Phi-2.
  - **Dataset**: TeleQnA dataset was introduced as a domain-specific benchmark focused on question ansewering over 3GPP and telecom-related documents available via GitHub at TeleQnA on GitHub.

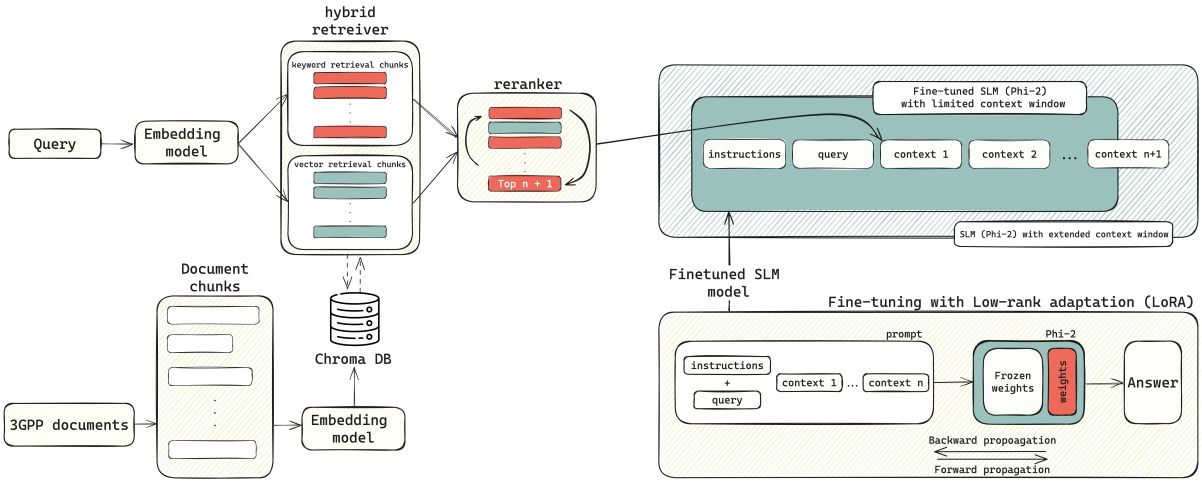


Figure 3.2: Overview of proposed RAG architecture with semantic chunking, extended context support, and fine-tuned Phi-2 SLM integration for 3GPP document processing.

- **Efficiency of LLMs for Understanding 3GPP Specifications Using RAG, Fine Tuning, and RAGAS**
  - **Objective:** evaluate the performance of both small and large language models (LLMs) in answering telecommunications-related questions [7]. The study aimed to:
    - Compare the performance of different LLMs using a domain-specific dataset (TeleQnA).
    - Implement and test a Retrieval-Augmented Generation (RAG) system using 3GPP technical documents (TSpec-LLM) to enhance context and accuracy.

    - Assess the impact of fine-tuning a smaller model to see if domain adaptation significantly improves performance

- **Model used**:Three models were evaluated:
    - GPT-4o-mini from OpenAI.
    - LLaMA 3.2 3B from Meta, tested in two configurations:
        - Base (pretrained only).
        - Fine-tuned version (referred to as LLaMA 3.2 3B LoRA), trained using the TeleQnA dataset.
- **Dataset**:Two datasets were used:
    - TeleQnA: A curated set of 10,000 telecommunications-related QnA items (multiple choice), with 733 focused on 3GPP Release 17. 100 questions were used for evaluation.
    - TSpec-LLM: A comprehensive corpus of over 30,000 3GPP technical documents (1999–2023), used for retrieval in the RAG setup. Documents were split into 780,000 chunks for embedding and indexing.

## 3.5 Comparison of experimental results and analysis

A comparative summary of the results obtained across four state-of-the-art studies that explored the application of Large Language Models (LLMs) to the understanding and processing of 3GPP standard documents. The analysis covers different model architectures while performing question answering (QA), summarization, and classification tasks using the used including SPEC5G, TSpec-LLM as well as TeleQnA datasets.

The evaluation metrics consist of Accuracy and F1 Score alongside ROUGE-L for summarization and RAGAS metrics which include Factual Correctness and Semantic Similarity. The models compared span small fine-tuned LLMs (e.g., Phi-2, LLaMA 3.2B with LoRA) and general-purpose models like GPT-4o-mini ,also the finetuned models XLNet5G, BERT5G and RoBERTa5G.

Below a table that includes the results of each related works:

| **Related Works** | **Task Type** | **Model(s) Evaluated** | **Fine-Tuned** | **Dataset(s) Used** | **Accuracy (%)** | **F1 Score** | **R-L** | **Factual Correctness** | **Semantic Similarity** |
|---|---|---|---|---|---|---|---|---|---|
| Leveraging FT RAG | multiple-choice questions | Phi-2 | No | TeleQnA | 71.35% | – | – | – | – |
| | | GPT-4o | No | | 80.90% | – | – | – | – |
| | | Phi-2 (LoRA) | Yes | | 80.30% | – | – | – | – |
| Efficiency of LLMs | question-answering (MCQ+RAG) | RAG LLaMA 3.2B LoRA | Yes | TeleQnA + TSpec-LLM | 77% | – | – | 0.180 | 0.410 |
| | | RAG GPT-4o-mini | No | | 73% | – | – | 0.181 | 0.421 |
| | | RAG LLaMA 3.2B | No | | 72% | – | – | 0.141 | 0.415 |
| SPEC5G Dataset | Summarization | BERT5G | Yes | SPEC5G + 5GSum | – | – | 0.472 | – | – |
| | | RoBERTa5G | Yes | | – | – | 0.469 | – | – |
| | | XLNet5G | Yes | | – | – | 0.453 | – | – |
| | Classification | BERT5G | Yes | SPEC5G + 5GSC | 76.55% | 0.6856 | – | – | – |
| | | ROBERTa5G | Yes | | 69.66% | 0.5785 | – | – | – |
| | | XLNet5G | Yes | | 71.03% | 0.6619 | – | – | – |
| | | GPT-2 | No | | 79.73% | 0.5767 | – | – | – |
| TSpec-LLM | Query Response (RAG) | RAG+GPT-3.5 | No | TSpec-LLM | 71% | – | – | – | – |
| | | RAG+GPT-4 | No | | 72% | – | – | – | – |
| | | RAG+Gemini 1.0 | No | | 75% | – | – | – | – |

Table 3.1: Comparison of Experimental Results Across related works on LLMs for 3GPP Understanding

- Also a table that includes a notes about each related works:

| **Study / Related works** | **Notes** |
|---|---|
| Leveraging Fine-Tuned RAG | - They improve RAG setup with techniques like semantic chunking (splitting text meaningfully), SelfExtend (to handle longer texts).<br>- fine-tuned small language model (SLM) called Phi-2 which outperforms GPT-4 qualitatively in telecom-specific QA tasks.<br>- The work shows that a small, fine-tuned model can rival large LLMs like GPT-4 in a specialized domain such as telecom. |
| Efficiency of LLMs for Understanding 3GPP Specifications | - evaluates how well small and large language models (like GPT-4o-mini and LLaMA 3B) can answer telecom-specific questions using the TeleQnA and TSpec-LLM datasets, leveraging fine-tuning and (RAG).<br>- Small models, when fine-tuned and supported by RAG, can perform as well or better than larger models for domain-specific tasks.<br>- RAG LLaMA 3.2B LoRA: strong fine-tuned small model performance which provides best accuracy among all tested models. |
| SPEC5G:A dataset for 5G Cellular Network Protocol Analysis | - First open-source dataset for 5G protocol includes annotated data for classification and summarization.<br>- SPEC5G aims to make 5G protocol analysis easier using LLMs.<br>- Highest summarization performance.<br>- SPEC5G proves powerful for 5G-specific tasks:<br>BERT5G: best F1 score among all classification models,<br>XLNet5G: best recall (0,6829),<br>GPT-2: best accuracy but poor F1 due to imbalance. |
| TSpec-LLM:Open-source dataset for LLM understanding of 3GPP specifications | - It aims to enhance LLM understanding of telecom standards by using TSpec-LLM ,that is includes tables and formulas which are important for understanding telecom standards.<br>- Test how well popular language models like GPT-3.5, GPT-4, and Gemini perform on telecom questions with and without help from this dataset.<br>- TSpec-LLM proves to be a powerful resource for improving LLM performance in the telecom field, when used with a simple RAG setup. |

Table 3.2: Notes of Related Work on LLMs for 3GPP Specifications.

## 3.6 Conclusion

The chapter demonstrates how LLMs optimize telecom networks through actual implementations and testing. The combination of specific datasets with optimized models enhances LLM performance for 3GPP-related tasks yet the high resource requirements continue to be a challenge. The research demonstrates that LLMs have significant potential to automate and improve telecom operations but additional work is required to enhance efficiency and expand datasets.

# Chapter 4

# Implementation and results analysis

## 4.1 Introduction

The chapter demonstrates how to apply Large Language Models (LLMs) for understanding 3GPP specifications in practical terms. The chapter presents chatbot systems developed from Naive-RAG and Graph-RAG architectures through LangChain and Hugging Face tools. The evaluation assesses the ability of these models to handle telecom-related queries. The chapter describes the datasets together with architectures and tools and testing methods which evaluate their performance in actual telecom applications.

## 4.2 Datasets overview

The open-source dataset TSpec-LLM exists to help language models understand technical telecommunications documents that stem from the 3rd Generation Partnership Project (3GPP). The dataset contains all 3GPP specification documents spanning from Release 8 to Release 19 (1999–2023) with their original content including tables and formulas which are crucial for technical understanding.

The dataset contains more than 30,000 documents with approximately 535 million words which enables Large Language Models (LLMs) to process complex telecommunications information effectively for standard specification interpretation.

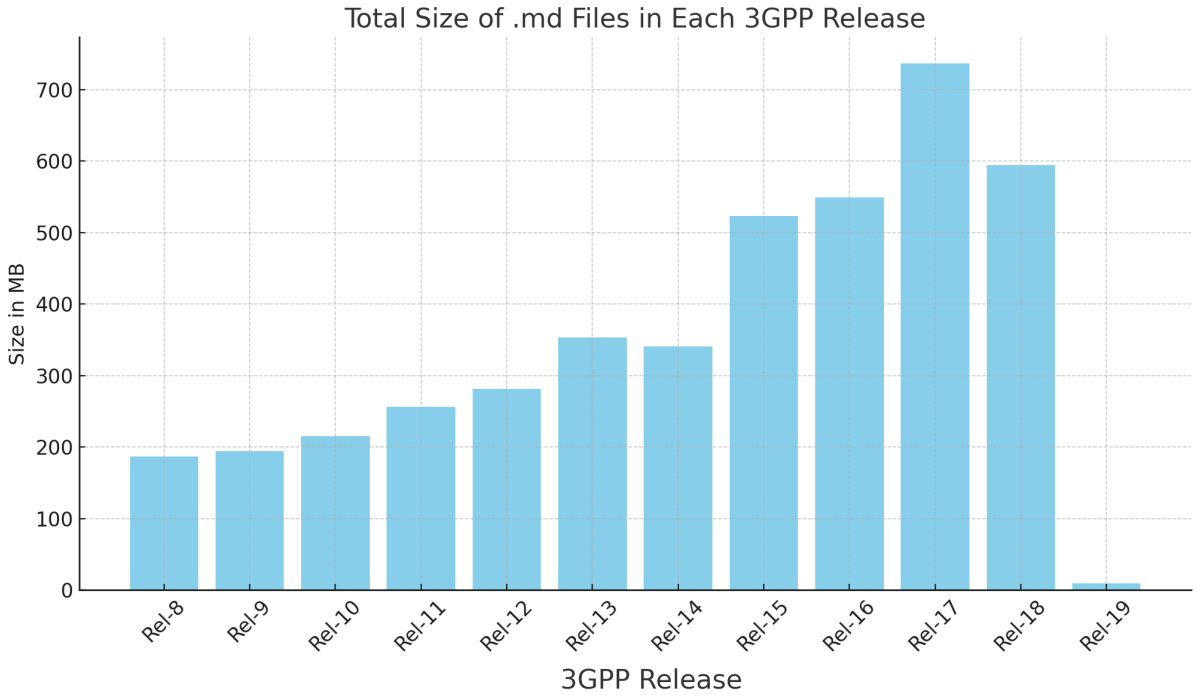


Figure 4.1: Total Markdown File Volume Across 3GPP Releases

The TSpec-LLM serves as a fundamental research instrument for AI telecommunications applications because it enables LLMs to function as assistants for reading and analyzing industry technical standards.

The TSpec-LLM dataset will be used to assist in information retrieval within the Retrieval-Augmented Generation (RAG) framework. In this context, the dataset will serve as the database of 3GPP technical documents, enabling the retrieval system to locate and provide relevant excerpts from these documents to complement the responses generated by LLMs.

## 4.3 Model

Two different versions of Llama model was used to create both Naive-RAG and Graph-RAG:

- **In the case of Naive-RAG:** Three different models were tested for comparison using the TeleQnA dataset:
  the GPT-4o-mini model from OpenAI, the Llama 3.2 model with 3 billion parameters from Meta, utilized through the Unsloth library, and the same Llama 3.2 model with 3 billion parameters but fine-tuned and trained using the TeleQnA dataset.
- **In the case of Graph-RAG:** the Llama-2-7b-chat-hf model comes in a range of parameter sizes — 7B, 13B, and 70B — as well as pretrained and fine-tuned

variations.IT is optimized for dialogue use cases. Llama-2-Chat models outperform closed-source chat models like ChatGPT and PaLM.

## 4.4 Results and performance evaluation

To validate the performance and applicability of the implemented systems, we conducted three categories of experiments mirroring real-world telecom LLM applications text classification (SPEC5G's), fine-tune a Llama 2 7b with Good-SPEC5G dataset as well as Retrieval-Augmented Generation for TeleQnA.

| **Model/Dataset** | **Task** | **Precision** | **Recall** | **F1 Score** | **Accuracy** | **Notes** |
|---|---|---|---|---|---|---|
| XLNet5G (GOOD-SPEC5G) | 5G Security Classification | 0.5925 | 0.5494 | 0.5644 | 61.50% | Effective but slightly lower performance. |
| BERT5G (GOOD-SPEC5G) | 5G Security Classification | 0.5927 | 0.5756 | 0.5828 | 62.88% | Moderate improvement over XLNet5G. |
| RoBERTa5G (GOOD-SPEC5G) | 5G Security Classification | 0.6164 | 0.6028 | 0.6078 | 64.82% | Best-performing among tested models. |
| LLaMA 2 7B (Fine-Tuned) | General 5G Text Tasks | - | - | - | - | Training limited by hardware constraints. |
| RAG (TeleQnA) | QA on 3GPP Documents | - | - | - | - | Delivered high-quality answers via chunking. |

Table 4.1: Results of Fine-Tuning and RAG Approaches

Furthermore, to evaluate the performance of Llama 3.2 model from Meta with 3 billion parameters but fine-tuned and trained using the TeleQnA dataset and GPT-4o-mini model from OpenAI in answering questions related to telecommunications, leveraging two main datasets: TeleQnA and TSpec-LLM.
A Retrieval-Augmented Generation (RAG) system was implemented to enhance response accuracy by incorporating relevant context from TSpec-LLM into the TeleQnA queries.the obtained results:

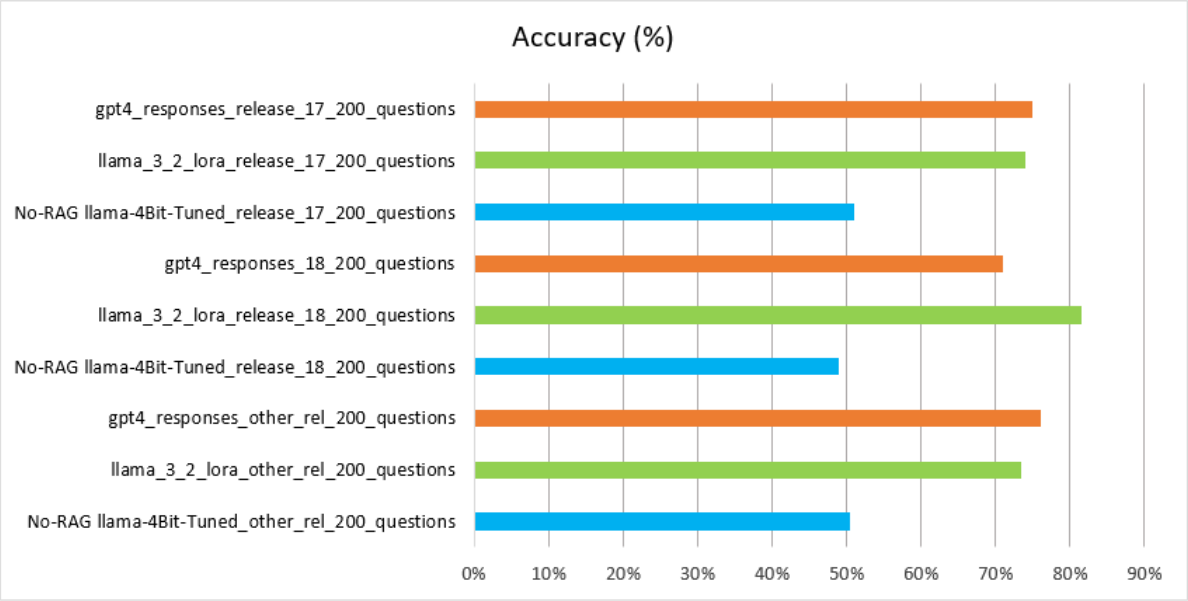


Figure 4.2: Average Accuracy Across Fine-Tuned LLaMA 3.2 LoRA Models and GPT-4o-mini on 200-Question Evaluation Sets

- RAG offers significant performance gains over both base models and fine-tuned models without retrieval augmentation.

- Even fine-tuned models alone (LoRA) do not match the performance of RAG-based approaches.

- The combination of semantic retrieval (RAG) and powerful language modeling enables more accurate answers by grounding outputs in relevant source material.

## 4.5 Application

### 4.5.1 Classification of Chatbots:

Chatbots could be classified into various categories based on several criteria e.g. mode of interaction, knowledge domain, their usage and the design techniques (response generation method) that are typically employed in building these chatbots.
Also the need of stored and considered in understanding the conversation or the type and purpose of the conversation for which the chatbot needs to be designed.
Open domain chatbots can talk about general topics and respond appropriately, while closed domain chatbots are focused on a particular knowledge domain such as TSpec-LLM datastes and might fail to respond to other questions.

| Classification Criteria | Interact Mode | | Knowledge Domain | | Goals | | Design Approach | | |
|---|---|---|---|---|---|---|---|---|---|
| | Text-Based | Voice-Based | Open Domain | Closed Domain | Task-Oriented | Non-Task-Oriented | Rule-Based | Retrieval-Based | Generative-Based |
| 3GPP Spec Chatbot | ✓ | | | ✓ | ✓ | | | ✓ | |

Table 4.2: Classification of 3GPP specifications Chatbot.

### 4.5.2 Chatbot Creation Using Naive-RAG

- **Designing the User Interface (UI):**

  – This is the main interface of the TSpec Chatbot where users can input their queries. The interface includes:
  - A query input field for entering questions related to 3GPP specifications.
  - A slider to adjust the number of results to retrieve (from 1 to 10).
  - A search button to submit the query.
  - On the left sidebar: A "Clear Chat History" button to reset the chat session.An About section that explains the chatbot's technical architecture.

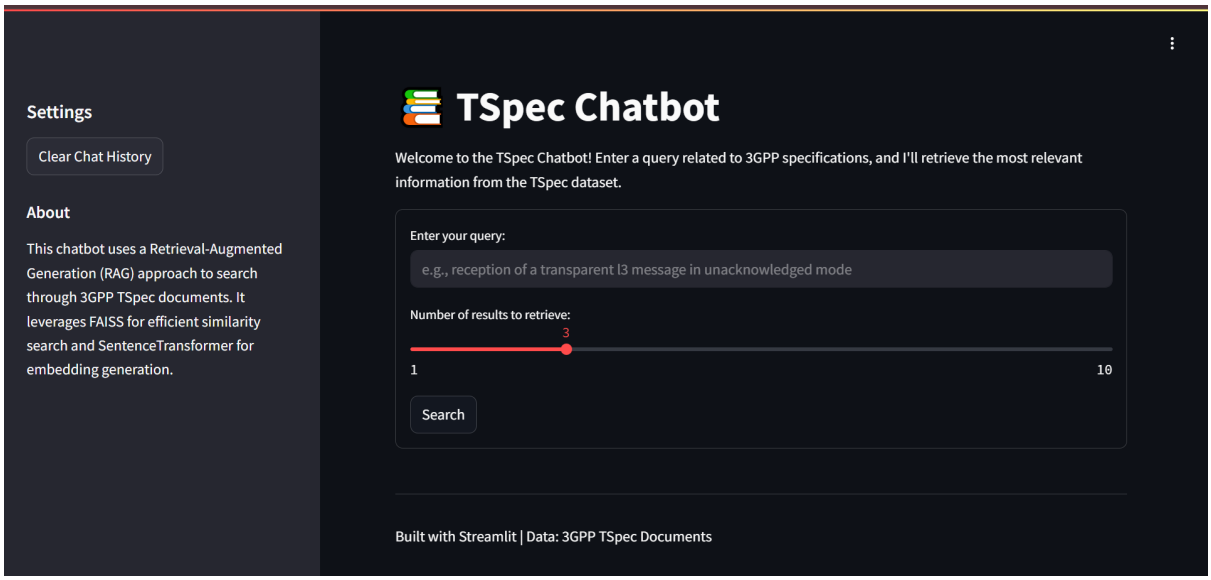


Figure 4.3: TSpec Chatbot Home Interface

  – This image shows the response output after submitting a query (in this case, about the reception of a transparent L3 message in unacknowledged mode).

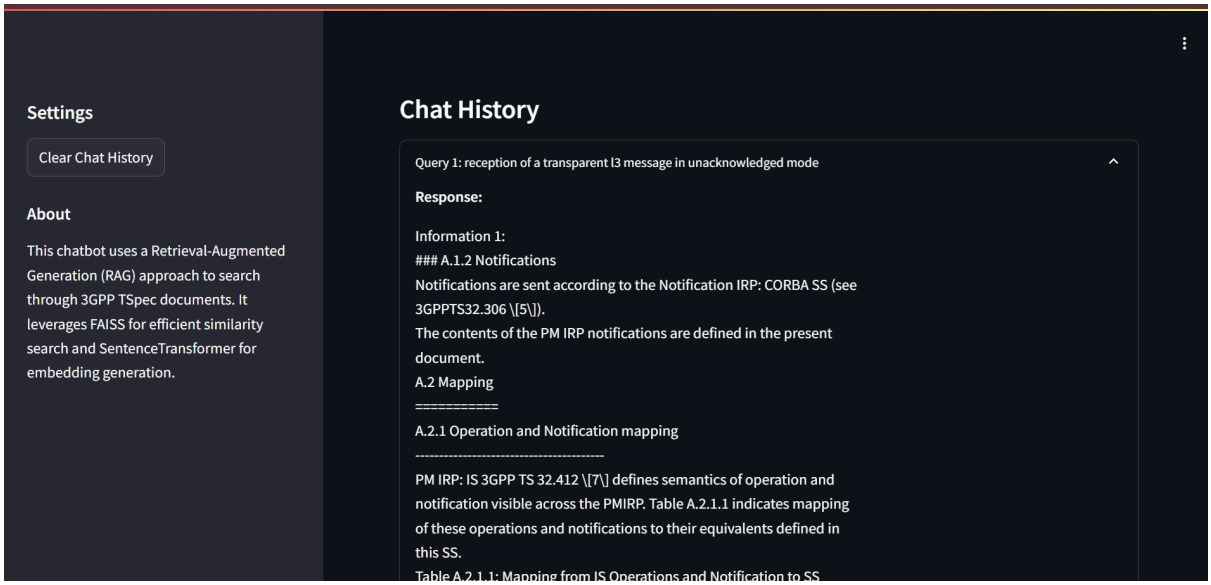


Figure 4.4: TSpec Chatbot Response Interface

– The query field filled with a previously used query, now with the number of results increased to 5.
The chat history helps users track and review past searches without needing to retype them, supporting iterative learning and exploration.

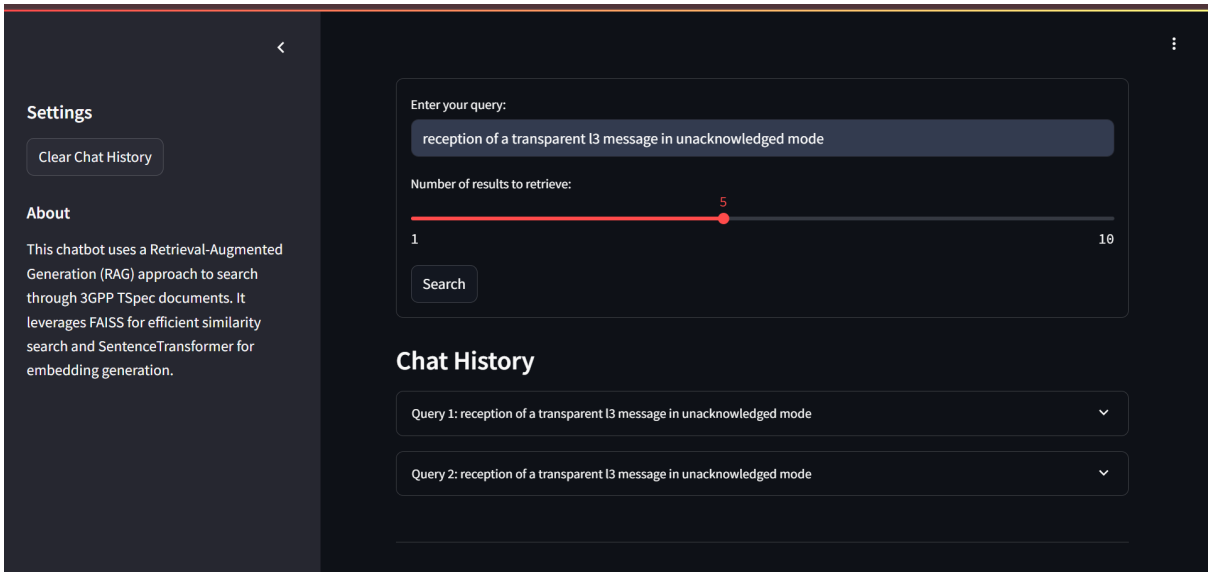


Figure 4.5: Query Refinement and Chat History

### 4.5.3 Chatbot Creation Using Graph-RAG

A chatbot system based on Graph-RAG (Graph-based Retrieval-Augmented Generation) technology was created to improve user interaction with technical documents. It leverages a structured knowledge graph to retrieve and reason over complex domain-specific

content, such as 3GPP standards. The chatbot system involves both backend processing (graph construction and traversal) and frontend interaction (answer generation and delivery).

- **Implementation Details:**
  - Building the model: loading a quantized LLaMA model via Hugging Face, setting up the tokenizer and pipeline, and wrapping it for inference. Additionally, it includes NLTK setup for tokenization and word semantics.

```python
# Set HuggingFace token
os.environ["HUGGINGFACEHUB_API_TOKEN"] = "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"
# Define quantization configuration
quantization_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_compute_dtype=torch.float16
)
# Load tokenizer and model
model_id = "meta-llama/Llama-2-7b-chat-hf"
tokenizer = AutoTokenizer.from_pretrained(model_id, token=os.environ["HUGGINGFACEHUB_API_TOKEN"])
model = AutoModelForCausalLM.from_pretrained(
    model_id,
    quantization_config=quantization_config,
    device_map="auto",
    token=os.environ["HUGGINGFACEHUB_API_TOKEN"]
)
# Create HuggingFace pipeline
hf_pipeline = pipeline(
    "text-generation",
    model=model,
    tokenizer=tokenizer,
    max_new_tokens=512,
    device_map="auto",
    return_full_text=False
)
# Wrap pipeline in HuggingFacePipeline
llm = HuggingFacePipeline(pipeline=hf_pipeline)
# Initialize ChatHuggingFace
chat_llm = ChatHuggingFace(
    llm=llm,
    model_id=model_id,
    temperature=0.0
)
# NLTK setup
nltk.download('punkt', quiet=True)
nltk.download('wordnet', quiet=True)
```

Figure 4.6: Model Pipeline Initialization Code

  - Core Components of the Graph-RAG Framework
    - DocumentProcessor: Handles the initial processing of input documents, creating text chunks and embeddings.
    - KnowledgeGraph: Constructs a graph representation of the processed documents, where nodes represent text chunks and edges represent relationships between them.
    - QueryEngine: Manages the process of answering user queries by leveraging the knowledge graph and vector store.

* Visualizer: Creates a visual representation of the graph and the traversal path taken to answer a query.

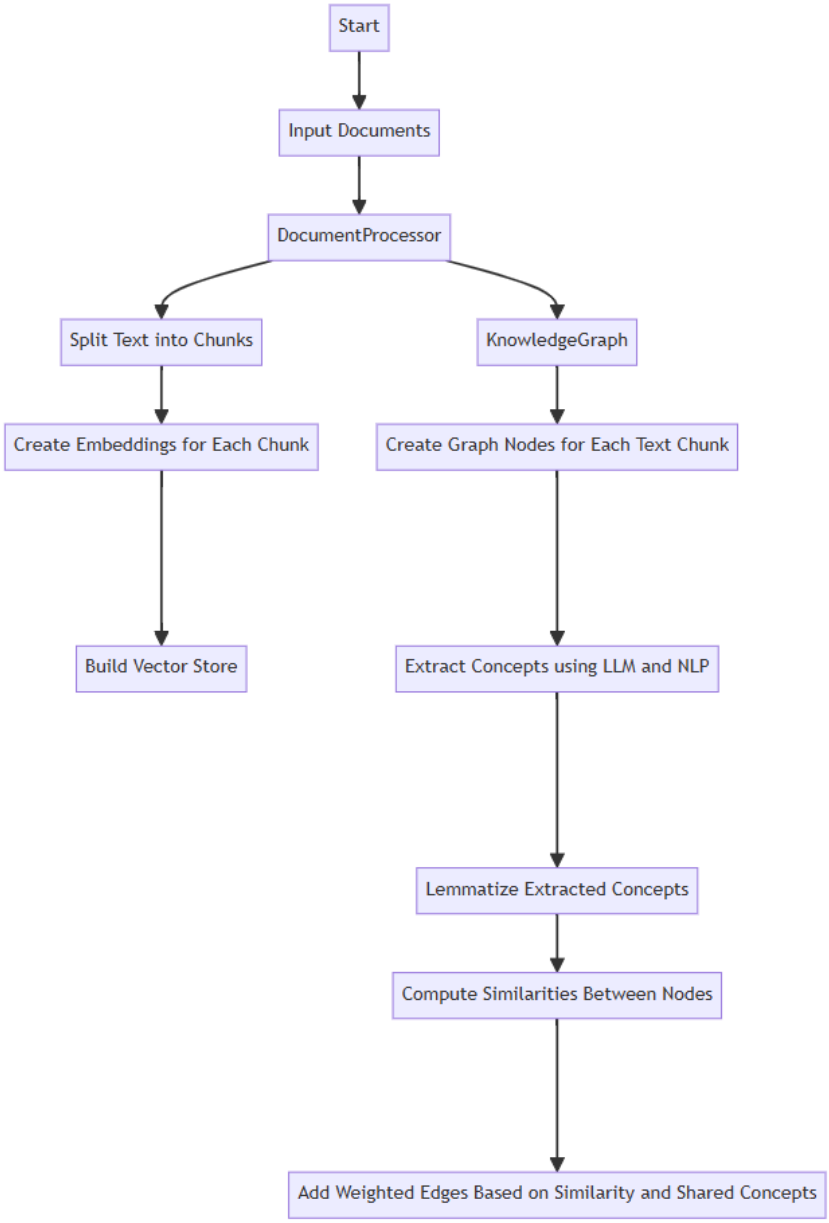


Figure 4.7: Semantic Document Representation via Vector Stores and Knowledge Graphs

- **Visualization of traversal graph**

  This figure illustrates the graph traversal process used by the Graph-RAG system to extract relevant context for answering the query. Also transparency into the decision-making path taken by the model.

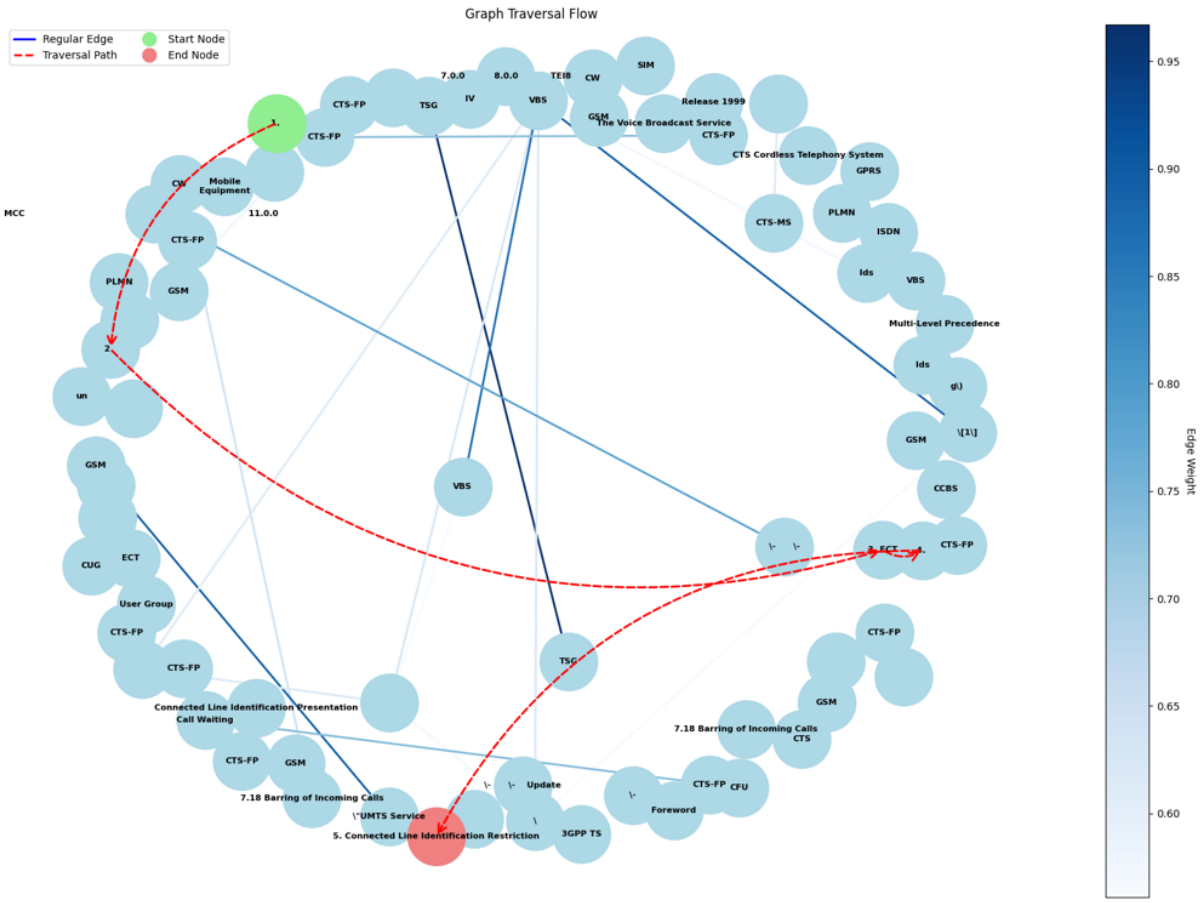


Figure 4.8: Graph Traversal Flow

- **Generating final answer of query**

# Input a query and get the retrieved information from the graph RAG

```
# Input a query and get the retrieved information from the graph RAG
query = "reception of a transparent 13 message in unacknowledged mode"
response = graph_rag.query(query)
print(f"\nQuery Response: {response}")
```

Figure 4.9: Graph-RAG Query Example for 3GPP Specification Retrieval

```
Generating final answer...

Final Answer:   Based on the provided context, the query is related to the reception of a transp
arent L3 message in unacknowledged mode.

From the context, we can infer that the query is asking about the behavior of a mobile station
(MS) in receiving a transparent L3 message in unacknowledged mode. The context mentions that the
MS is in unacknowledged mode, which means that the MS is not expecting any acknowledgement for t
he message it is receiving.

According to the 3GPP TS 23.040 specification, when a MS is in unacknowledged mode, it will not
send an acknowledgement for any message it receives. However, the MS may still be able to detect
and process certain types of messages, such as transparent L3 messages, even if it is not expect
ing an acknowledgement for them.

Therefore, the answer to the query is that the MS will receive the transparent L3 message in una
cknowledged mode, but it will not send an acknowledgement for the message.
```

Figure 4.10: Final Answer Generated by the Graph-RAG Chatbot

which is provide

- A natural-language explanation synthesized from technical specifications.
- Standards-compliant reasoning (from 3GPP) for user queries.
- An interpretable and context-relevant final answer.

## 4.6 Development tools and libraries

This project incorporates multiple tools, frameworks and libraries. Below is a concise overview of each.

- **Used libraries:**
  - **Pytorch:** Is a machine learning library based on the Torch library, used for applications such as computer vision and natural language processing. It is one of the most popular deep learning frameworks, alongside others such as TensorFlow offering free and open-source software.


Figure 4.11: Pytorch logo

  - **Pickle:** Python pickle module is used for serializing and de-serializing a Python object structure. Any object in Python can be pickled (Pickling is a way to convert a Python object (list, dictionary, etc.) into a character stream) so that it can be saved on disk. Understanding Python Pickling

– **LangChain:** is a framework for developing applications powered by large language models (LLMs).It implements a standard interface for large language models and related technologies, such as embedding models and vector stores, and integrates with hundreds of providers. langchain.com


Figure 4.12: LangChain logo

– **Transformers:** Is a library of pretrained natural language processing, computer vision, audio, and multimodal models for inference and training. Use Transformers to train models on your data, build inference applications, and generate text with large language models. huggingface/transformers


Figure 4.13: Transformers (Hugging Face) logo

– **Faiss:** Is a library for efficient similarity search and clustering of dense vectors. It contains algorithms that search in sets of vectors of any size.
So, given a set of vectors, we can index them using Faiss then using another vector (the query vector), we search for the most similar vectors within the index.

– **NumPy:** Is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.


Figure 4.14: NumPy logo

– **Streamlit:** Is an open-source app framework that allows you to create interactive web apps for machine learning and data science.


Figure 4.15: Streamlit logo

- **Unsloth:** Is a Python framework designed for fast fine-tuning and accessing large language models. It offers a simple API and performance that is 2x faster compared to Transformers.


Figure 4.16: Unsloth logo

- **NetworkX:** Is a Python library for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. It is used to study large complex networks represented in form of graphs with nodes and edges.


Figure 4.17: NetworkX logo

- **Matplotlib:** Is a comprehensive library for creating static, animated, and interactive visualizations in Python.

- **Working environment:**
  - **Kaggle:** Is a data science competition platform and online community for data scientists and machine learning practitioners under Google LLC.
    Where users can discover datasets for building AI models, share their own datasets and participate in competitions aimed at solving data science challenges and Served as the development environment for executing the code and conducting experiments.


Figure 4.18: Kaggle logo

– **Google Colab:** Or colaboratory is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education.
Google Colab's major differentiator from Jupyter Notebook is that it is cloud-based and Jupyter is not. This means that if you work in Google Colab, you do not have to worry about downloading and installing anything to your hardware.


Figure 4.19: Colab logo

## 4.7 Conclusion

The research showed that LLMs can be used in real telecom applications to understand complex 3GPP documents. The results of Naive-RAG and Graph-RAG comparison showed that better document structuring leads to more accurate and useful results. The work confirmed that LLMs—when fine-tuned and well-integrated—can serve as practical, intelligent tools in network automation and technical support.

# General Conclusion

In this thesis, we have investigated the potential and practical implementation of Large Language Models in enhancing network optimization, particularly in the context of 3GPP (3rd Generation Partnership Project) standards. From foundational neural architectures to fine-tuning techniques, we demonstrated how LLMs can be adapted for technical use cases such as understanding 3GPP specifications, improving network automation, and enabling intelligent anomaly detection.

The main achievement of this research involved using LLMs to process 3GPP specifications which represent complex technical documents needed for cellular network design and operation. By leveraging Retrieval-Augmented Generation (RAG) techniques and fine-tuning pretrained models on domain-specific datasets such as TSpec-LLM and TeleQnA, we demonstrated that LLMs can significantly improve document comprehension, technical QA, and the automation of support tasks in telecom environments.

Moreover, through the implementation of chatbots using Naive-RAG and Graph-RAG architectures, we demonstrated how LLMs can be employed in practical tools to assist engineers and researchers in querying 3GPP documentation. The tools enhance operational efficiency and reduce the cognitive load associated with navigating complex technical content.

# Bibliography


[1] Akram AbdelBaqi AbdelRahman, Noor Kadhim Meftin, and Eman Khalil Alak. "AI-Driven Cloud Networking Optimizations for Seamless LTE Connectivity." In: (Oct. 2024).

[2] Danial Abshari, Chenglong Fu, and Muralikrishna Sridhar. "LLM-assisted Physical Invariant Extraction for Cyber-Physical Systems Anomaly Detection." In: (Nov. 2024).

[3] A. P. Aishwarya, K. Das, and L. N. Rao. "Tandem Transformers: Improving Inference Speed and Accuracy in LLMs." In: *arXiv preprint arXiv:2407.65432* (2024).

[4] Majid Alshammari. "Deep Learning Approaches to SQL Injection Detection: Evaluating ANNs, CNNs, and RNNs." In: *Cybersecurity and AI Research* (Dec. 2023).

[5] Ammar Haziq Annas, A. A. Zainuddin, and Afnan Wajdi Ramlee. "Analyses of 6G-Network and Blockchain-Network Application Security: Future Research Prospect." In: *International Journal on Perceptive and Cognitive Computing* (July 2024).

[6] D. Anthony, J. Smith, and M. Rodriguez. "Scaling Transformers: Addressing Communication Bottlenecks in Distributed Training." In: *arXiv preprint arXiv:2404.98765* (2024).

[7] José de Arimatéa Passos Lopes Júnior. *Efficiency of LLMs for Understanding 3GPP Specifications Using Approaches such as RAG and Fine Tuning, and Evaluation with Accuracy and RAGAS*. https://github.com/josearimatea/3gpp_llm_evaluation. Accessed: Nov. 2024. 2024.

[8] Alexander Bakumenko, Kateřina Hlaváčková-Schindler, and Claudia Plant. "Advancing Anomaly Detection: Non-Semantic Financial Data Encoding with LLMs." In: (June 2024).

[9] Aimilia Bantouna, K Λ , and Omar Qaise. "SAND5G - Security Assessments for Networks and Services in 5G Networks." In: (Aug. 2024).

[10] Peristera A. Baziana, G. Drainakis, and David Georgantas. "AI and ML Applications Traffic: Designing Challenges for Performance Optimization of Optical Data Center Networks." In: (Sept. 2024).

[11] Ruben Van Belle. "Learning for Decision and Control in Stochastic Networks." In: *Network Optimization Techniques*. Synthesis Lectures on Learning, Networks, and Algorithms. Jan. 2023.

[12] Davide Berardi and Barbara Martini. "MECHATRON – Security Analysis of 6G and 5G Networks Using Multiaccess Edge Computing." In: (Oct. 2024).

[13] B. Bermeitinger et al. "Reevaluating the Necessity of MLPs in Transformer Architectures." In: *arXiv preprint arXiv:2402.12345* (2024).

[14] F. Bu, Y. Wang, and M. Li. "In-Context Learning in Transformers: Mechanisms and Applications." In: *Proceedings of the 2024 AI Summit*. 2024, pp. 67–79.

[15] T. Chai and L. Zhou. "Deep Learning in Computer Vision: Object Detection and Beyond." In: *Computer Vision Research Journal* 15.4 (2021), pp. 220–237.

[16] K. Chaitanya and Krishna Jayanth Rolla. "The Evolution and Impact of Large Language Models in Artificial Intelligence." In: *Book Chapter*. June 2024.

[17] R. Chinnaiyan and A. Gupta. "BERT, GPT, and the Future of NLP." In: *International Journal of Machine Learning* 19.2 (2024), pp. 99–118.

[18] P. Chinnasamy, G. Charles Babu, and Ramesh Kumar Ayyasamy. "Blockchain 6G-Based Wireless Network Security Management with Optimization Using Machine Learning Techniques." In: *Sensors* (Sept. 2024).

[19] Chrıstophoros Christophorou, Iacovos Ioannou, and Vasos Vassiliou. "ADROIT6G DAI-driven Open and Programmable Architecture for 6G Networks." In: (Mar. 2024).

[20] Francesco D'Alterio et al. "Navigating 5G Security: Challenges and Progresses on 5G Security Assurance and Risk Assessment." In: (Sept. 2024).

[21] Geoffrey Daniel, M. Trupthi, and Gopal Reddy. "Model Optimization Methods for Efficient and Edge AI: Federated Learning Architectures, Frameworks and Applications." In: *AI Model Optimization Techniques*. Nov. 2024.

[22] Priyanka Das et al. "Secure and Smart Cyber-Physical Systems Chapter of 6G Communication Technology for Industry 5.0." In: (June 2024).

[23] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *arXiv preprint arXiv:1810.04805* (2018). URL: https://arxiv.org/abs/1810.04805.

[24] Manqing Dong, Hao Huang, and Longbing Cao. "Can LLMs Serve As Time Series Anomaly Detectors?" In: (Aug. 2024).

[25] Omar Erak et al. "Leveraging Fine-Tuned Retrieval-Augmented Generation with Long-Context Support: For 3GPP Standards." In: *arXiv preprint arXiv:2408.11775* (2025). Available at: https://arxiv.org/abs/2408.11775.

[26] Victor V. Erokhin and Elena V. Eliseeva. "Influence of Neural Network Architecture on Its Performance." In: *Artificial Intelligence and Machine Learning Review* (Jan. 2022).

[27] Swati Goel, Kalpna Guleria, and Surya Narayan Panda. "Applied Data Science and Smart Systems." In: *Optimization Techniques for Wireless Body Area Network Routing Protocols: Analysis and Comparison*. June 2024.

[28] L. Gong and W. Zhang. "Memory Constraints in Large Language Models: Analyzing Self-Attention's Role." In: *arXiv preprint arXiv:2405.45678* (2024).

[29] John Brandon Graham-Knight et al. "Predicting and Explaining Performance and Diversity of Neural Network Architecture for Semantic Segmentation." In: *Machine Learning and Vision Journal* (Oct. 2022).

[30] Yuanning Hong et al. "Larger Language Models Better Predict Neural Activity During Naturalistic Language Comprehension." In: *Proceedings of the Conference on Cognitive Computational Neuroscience (CCN)*. 2024. URL: https://2024.ccneuro.org/pdf/260_Paper_authored_CCN_2024-%282%29.pdf.

[31] Ismail Hossain, M. J. Alam, and Sai Puppala. "EVOLVE: Predicting User Evolution and Network Dynamics in Social Media Using Fine-Tuned GPT-like Model." In: *arXiv preprint arXiv:2407.12345* (July 2024).

[32] R. Huang, C. Zhao, and W. Xu. "Advancements in Transfer Architectures for Long-Context Processing." In: *arXiv preprint arXiv:2309.11234* (2023).

[33] G. Jeon et al. "Guest Editorial: Unfolding the Potential of 5G Technologies for Future Wireless Networks." In: *IEEE Communications Magazine* (July 2024).

[34] Chunheng Jiang et al. "Network Properties Determine Neural Network Performance." In: *Journal of Neural Network Analysis* (July 2024).

[35] X. Jiang, Y. Li, and Z. Wang. "Advancements in Transformer-Based NLP Models: A Comparative Study." In: *Journal of Computational Linguistics* 42.1 (2024), pp. 25–38.

[36] Yiming Ju and Huanhuan Ma. "Training Data for Large Language Model." In: *arXiv preprint arXiv:2411.07715* (2024). URL: https://arxiv.org/abs/2411.07715.

[37] Imtiaz Karim et al. "SPEC5G: A Dataset for 5G Cellular Network Protocol Analysis." In: *arXiv preprint arXiv:2301.09201* (2023). Available at: https://arxiv.org/abs/2301.09201.

[38] Murad Khan. "A Review of Optimization Techniques in Network-on-Chip (NoC) Architecture." In: (May 2024).

[39] Bahare M. Khorsandi and Mohammad Asif Habibi. "Enabling Hexa-X 6G Vision: An End-to-End Architecture." In: (Oct. 2024).

[40] Anjanabhargavi Kulkarni et al. “New Directions for Adapting Intelligent Communication and Standardization Towards 6G.” In: *ICST Transactions on Scalable Information Systems* (July 2024).

[41] P. Kumari, R. Sharma, and M. Patel. “Transformers and Their Impact on NLP and AI.” In: *AI and Machine Learning Review* 12.3 (2024), pp. 45–67.

[42] Aodong Li, Yong Zhao, and Chen Qiu. “Anomaly Detection of Tabular Data Using LLMs.” In: (June 2024).

[43] Bozhou Li et al. “Gradual Learning: Optimizing Fine-Tuning with Partially Mastered Knowledge in Large Language Models.” In: *arXiv preprint arXiv:2410.05802* (2024). URL: https://arxiv.org/abs/2410.05802.

[44] Haoyu Lu, Wen Liu, and Bo Zhang. “DeepSeek-VL: Towards Real-World Vision-Language Understanding.” In: *arXiv preprint arXiv:2403.12345* (Mar. 2024).

[45] H. Luo, K. Tan, and Y. Chen. “Evolution of Self-Attention in Large Language Models.” In: *Journal of Artificial Intelligence Research* 40.3 (2023), pp. 134–152.

[46] Recalde Varela Pablo Marcel, Bolagay Egas Mauro Fernando, and Yanez Velasquez Jorge Roberto. “A Brief History of Artificial Intelligence: ChatGPT - The Evolution of GPT.” In: *Conference Name*. June 2023.

[47] Raja Marjieh et al. “Large Language Models Predict Human Sensory Judgments Across Six Modalities.” In: *arXiv preprint arXiv:2302.01308* (2023). URL: https://arxiv.org/abs/2302.01308.

[48] J. Miao, T. Xie, and S. Feng. “Efficient Attention Mechanisms for Long-Context Processing in LLMs.” In: *arXiv preprint arXiv:2406.78901* (2024).

[49] Oluwaseyi Olakunle Mokuolu. “Reviewing the Impact of Artificial Intelligence and Machine Learning in Enhancing IoT Applications Performance.” In: *International Journal of Science and Research Archive* (Aug. 2024).

[50] Minsoo Na, Jaehyun Lee, and Giwan Choi. “Operator’s Perspective on 6G: 6G Services, Vision, and Spectrum.” In: *IEEE Communications Magazine* (Aug. 2024).

[51] Vaishali Nagpure. “AI-Driven Network Traffic Optimization and Fault Detection in Enterprise WAN.” In: *Indian Scientific Journal Of Research In Engineering And Management* (Nov. 2024).

[52] Rasoul Nikbakht, Mohamed Benzaghta, and Giovanni Geraci. “TSpec-LLM: An Open-source Dataset for LLM Understanding of 3GPP Specifications.” In: *arXiv:2406.01768v1 [cs.NI]* (2024).

[53] Sergey Nikolenko. *Deep Learning for Natural Language Processing*. Springer, 2023.

[54] M. Ormaniec, T. Novak, and R. Patel. "Enhancing Transformer Training with Adaptive Optimizers and Normalization Techniques." In: *arXiv preprint arXiv:2403.56789* (2024).

[55] Adam Orucu et al. "On Multi-Objective Neural Architecture Search for Modeling Network Performance." In: *Neural Computing Journal* (Oct. 2024).

[56] S. Pradeep et al. "The Impact of 5G on Real-Time IoT Data Processing: Exploring Challenges and Innovative Solutions." In: (Aug. 2024).

[57] R. P. Prathamesh, M. B. Sameera, et al. "The Evolution of Large Language Model: Models, Applications and Challenges." In: *Journal Article* (May 2024).

[58] N. Prottasha et al. "Semantic Knowledge Tuning: Leveraging Meaningful Tokens for Efficient and Task-Specific Adaptation." In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2024.

[59] Nitin Rane et al. "Techniques and Optimization Algorithms in Deep Learning: A Review." In: *Deep Learning Review Journal* (Oct. 2024).

[60] J. A. Rathod et al. "Future of 5G Wireless System." In: *International Journal of Advanced Research in Science, Communication and Technology* (July 2024).

[61] Alicia Russell-Gilbert, Alexander Sommers, and Andrew Thompson. "AAD-LLM: Adaptive Anomaly Detection Using Large Language Models." In: (Nov. 2024).

[62] V. Sahu, N. Sahu, and R. Sahu. "Challenges and Opportunities of 5G Network: A Review of Research and Development." In: *American Journal of Electrical and Computer Engineering* (July 2024).

[63] V. K. Saxena. "The 5G Era: A Comprehensive Review of Recent Advancements and Applications." In: *Indian Scientific Journal of Research in Engineering and Management* (July 2024).

[64] Paul Scalise, Ramón Serramito García, and Matthew Boeding. "An Applied Analysis of Securing 5G/6G Core Networks with Post-Quantum Key Encapsulation Methods." In: *Electronics* (Oct. 2024).

[65] S. Sengupta et al. "MonteCLoRA: Bayesian Reparameterization for Stable and Accurate Low-Rank Adaptation." In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2024.

[66] Sarah Al-Shareeda, Khayal Huseynov, and Lal Verda Çakır. "AI-Based Traffic Analysis in Digital Twin Networks." In: (Nov. 2024).

[67] V. Sheela and P. Rathiga. "Overview of Mobile Technologies in 5G Networking and Communication." In: *International Journal for Research in Applied Science and Engineering Technology* (Mar. 2024).

[68] Alexandre Sousa and Manuel J. C. S. Reis. "5G Security Features, Vulnerabilities, Threats, and Data Protection in IoT and Mobile Devices: A Systematic Review." In: *Evolutionary Studies in Imaginative Culture* (Sept. 2024).

[69] Sreerakuvandana Sreerakuvandana, Princy Pappachan, and Varsha Arya. *Understanding Large Language Models*. Advances in Computational Intelligence and Robotics Book Series. Aug. 2024.

[70] Hugo Touvron et al. "LLaMA: Open and Efficient Foundation Language Models." In: *arXiv preprint arXiv:2302.13971* (2023). URL: https://arxiv.org/abs/2302.13971v1.

[71] Mano Ashish Tripathi et al. "The Role of 5G in Creating Smart Cities for Achieving Sustainable Goals: Analyzing the Opportunities and Challenges through the MANOVA Approach." In: (Oct. 2024).

[72] V. Vaissnave et al. "Advancements in Deep Learning Algorithms." In: *Journal of Artificial Intelligence Research* (Mar. 2024).

[73] Suchita Walke et al. "Investigating Neural Network-Based Deep Learning Strategies for Real-Time Data Analysis in Machine Learning." In: *Machine Learning Journal* (Dec. 2023).

[74] Chen Wang, Jin Zhao, and Jiaqi Gong. "A Survey on Large Language Models from Concept to Implementation." In: *arXiv.org* (Mar. 2024).

[75] Hou Xianqiang. "Research on Neural Network Structure Optimization Under the Framework of Computer Deep Learning." In: *Journal of Computer Science and Deep Learning* (May 2024).

[76] Henan Xin, Daya Guo, and Zhiguo Shao. "DeepSeek-Prover: Advancing Theorem Proving in LLMs through Large-Scale Synthetic Data." In: *arXiv preprint arXiv:2405.12345* (May 2024).

[77] R. Yadav. "Transformers in NLP: A Comparative Analysis of Machine Translation and Text Summarization." In: *Journal of Language Technology* 12.1 (2024), pp. 90–105.

[78] Mengmeng Yang et al. "From 5G to 6G: A Survey on Security, Privacy, and Standardization Pathways." In: (Oct. 2024).

[79] X. Yang. "Advancements in Deep Learning for Natural Language Processing." In: *Journal of Artificial Intelligence Research* 58 (2024), pp. 1–15.

[80] Junjie Ye et al. "60 Data Points are Sufficient to Fine-Tune LLMs for Question-Answering." In: *arXiv preprint arXiv:2409.15825* (2024). URL: https://arxiv.org/abs/2409.15825.

[81] Z. Ye, R. Chen, and D. Liu. "Optimizing Self-Attention Mechanisms for Large-Scale AI Models." In: *International Conference on AI Optimization*. 2024, pp. 89–101.

[82] Ilsun You, Jiyoon Kim, and I. W. A. J. Pawana. "Mitigating Security Vulnerabilities in 6G Networks: A Comprehensive Analysis of the DMRN Protocol Using SVO Logic and ProVerif." In: *Applied Sciences* (Oct. 2024).

[83] Yifei Yuan. "Fundamentals of 6G Communications and Networking." In: *6G Visions and Requirements*. Signals and Communication Technology. Dec. 2023.

[84] Khaled Mohammad Al-Zahrani and Abdullah Abdulaziz Al-Ghanim. "6G is The Future of Connectivity." In: *International Journal of Innovative Science and Research Technology* (Aug. 2024).

[85] Jingyi Zhang et al. "Vision-Language Models for Vision Tasks: A Survey." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.8 (2024), pp. 5625–5644. DOI: 10.1109/TPAMI.2024.3369699. URL: https://ieeexplore.ieee.org/document/10445007.

[86] H. Zhu, P. Lee, and A. Kumar. "Applications of Transformers in Neuroimaging: A Review." In: *Neuroinformatics* 22.2 (2024), pp. 50–72.