TelecomGPT: Revolutionizing Telecommunications with Large Language Models

Prof. Merouane Debbah
Khalifa University of Science and Technology

LLM Telecom Papers

Large Language Models for Telecom: The Next Big Thing?

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

***Abstract*—The evolution of generative artificial intelligence (GenAI) constitutes a turning point in reshaping the future of technology in different aspects. Wireless networks in particular, with the blooming of self-evolving networks, represent a rich field for exploiting GenAI and reaping several benefits that can fundamentally change the way how wireless networks are designed and operated nowadays. To be specific, large language models (LLMs), a subfield of GenAI, are envisioned to open up a new era of autonomous wireless networks, in which a multimodal large model trained over various Telecom data, can be fine-tuned to perform several downstream tasks, eliminating the need for dedicated AI models for each task and paving the way for the realization of artificial general intelligence (AGI)-empowered wireless networks. In this article, we aim to unfold the opportunities that can be reaped from integrating LLMs into the Telecom domain. In particular, we aim to put a forward-looking vision on a new realm of possibilities and applications of LLMs in future wireless networks, defining directions for designing, training, testing, and deploying Telecom LLMs, and reveal insights on the associated theoretical and practical challenges.**

and Contrastive Language-Image Pre-Training (CLIP), have strongly impacted how AI is employed for inference and decision-making purposes, and laid down a new base for novel applications that can exploit the potential of GenAI models. This is rooted to the generative and predictability capabilities of these LLMs, in which large models (mainly based on the transformer architecture) are trained over a vast amount of unlabeled multimodal data (primarily textual and/or visual data), and therefore, are enabled to understand and generate human-like languages. Through the self-attention mechanism of transformers and the large amount of training data, the developed large models are able to capture the statistical patterns and relationships in the provided data, and hence, to predict and generate the required data. Similar approach applies to visual models, where variational autoencoders (VAEs) and generative adversarial networks (GANs) can be leveraged to map contextual data with images and vice versa.

Understanding Telecom Language Through Large Language Models

Lina Bariah, Hang Zou, Qiyang Zhao, Belkacem Mouhouche, Faouzi Bader, and Merouane Debbah

Technology Innovation Institute, 9639 Masdar City, Abu Dhabi, UAE
Email: firstname.lastname@tii.ae

***Abstract*—The recent progress of artificial intelligence (AI) opens up new frontiers in the possibility of automating many tasks involved in Telecom networks design, implementation, and deployment. This has been further pushed forward with the evolution of generative artificial intelligence (AI), including the emergence of large language models (LLMs), which is believed to be the cornerstone toward realizing self-governed, interactive AI agents. Motivated by this, in this paper, we aim to adapt the paradigm of LLMs to the Telecom domain. In particular, we fine-tune several LLMs including BERT, distilled BERT, RoBERTa and GPT-2, to the Telecom domain languages, and demonstrate a use case for identifying the 3rd Generation Partnership Project (3GPP) standard working groups. We consider training the selected models on 3GPP technical documents (Tdoc) pertinent to years 2009-2019 and predict the Tdoc categories in years 2020-2023. The results demonstrate that fine-tuning BERT and RoBERTa model achieves 84.6% accuracy, while GPT-2 model achieves 83% in identifying 3GPP working groups**

multi-functional schemes that are capable of handling diverse network conditions. Accordingly, conventional AI algorithms are highly probable to fall behind in fulfilling the required performance, and therefore, a radical departure to more innovative AI-driven approaches is anticipated to shape the future of next generation wireless networks.

Foundation models (FMs) was coined by Stanford Center for Research on Foundation Models (CRFM) in 2021 and have attracted a considerable attention as generalized models that are capable of handling a wide range of downstream tasks [14]. In particular, FMs are extremely large neural networks that are trained over massive unlabeled datasets, in a self-supervised fashion, allowing several opportunities to be reaped with reduced time and cost (that would be unbearable in case of human labeling). Rapidly after being developed, FMs have

WIRELESS WORLD RESEARCH FORUM

OUTLOOK

Visions and research directions for the Wireless World

Foundation Models for Telecom: paving the way for self-built networks.

Editor: Merouane Debbah.

Contributors: Hang Zou, Yu Tian, Lina Bariah, Qiyang Zhao, Belkacem Mouhouche, Faouzi Bader, Ebtesam Almazrouei, Merouane Debbah,

Technology Innovation Institute.

Wireless Multi-Agent Generative AI: From Connected Intelligence to Collective Intelligence

Hang Zou, Qiyang Zhao, Lina Bariah, Mehdi Bennis, and Mérouane Debbah.

***Abstract*—The convergence of generative large language models (LLMs), edge networks, and multi-agent systems represents a groundbreaking synergy that holds immense promise for future wireless generations, harnessing the power of collective intelligence and paving the way for self-governed networks where intelligent decision-making happens right at the edge. This article puts the stepping-stone for incorporating multi-agent generative artificial intelligence (AI) in wireless networks, and sets the scene for realizing on-device LLMs, where multi-agent LLMs are collaboratively planning and solving tasks to achieve a number of network goals. We further investigate the profound limitations of cloud-based LLMs, and explore multi-agent LLMs from a game theoretic perspective, where agents collaboratively solve tasks in competitive environments. Moreover, we establish the underpinnings for the architecture design of wireless multi-agent generative AI systems at the network level and the agent level, and we identify the wireless technologies that are envisioned to play a key role in enabling on-device LLM. To demonstrate the promising potentials of wireless multi-agent**

telecom, and healthcare [1]. The maturity of LLMs constitutes a stepping stone towards realizing the vision of artificial general intelligence (AGI), which is a broader concept than AI, that encompasses highly autonomous systems of machines that possess general intelligence and cognition capabilities that are comparable to humans. Within this context, LLMs will play an essential role in AGI-based systems through their abilities to perform complex tasks on multi-modal data across many domains, with only a few examples [5].

The remarkable capabilities of LLMs are owed to the Transformer architecture, rooted in the self-attention mechanism [6]. Leveraging multi-head attention, Transformers capture long-range dependencies in parallel by attending a word to all previous words, or to all other words in the text. With pre-training on huge unlabeled corpus, generative models can learn universal knowledge representation, and can be further

LLM Telecom Papers

1

Artificial General Intelligence (AGI)-Native Wireless Systems: A Journey Beyond 6G

Walid Saad, *Fellow, IEEE,* Omar Hashash, *Graduate Student Member, IEEE,*
Christo Kurisummoottil Thomas, *Member, IEEE,* Christina Chaccour, *Member, IEEE,*
Mérouane Debbah, *Fellow, IEEE,* Narayan Mandayam, *Fellow, IEEE,* and Zhu Han, *Fellow, IEEE*

***Abstract*—Building next-generation wireless systems that could support metaverse services like digital twins (DTs) and holographic teleportation is challenging to achieve exclusively through incremental advances to conventional wireless technologies like meta-surfaces or holographic antennas. While the 6G concept of artificial intelligence (AI)-native networks promises to overcome some of the limitations of existing wireless technologies, current developments of AI-native wireless systems rely mostly on conventional AI tools like auto-encoders and off-the-shelf artificial neural networks. However, those tools struggle to manage and cope with the complex, non-trivial scenarios appearing in the network environment and the growing quality-of-experience requirements of the aforementioned, emerging wireless use cases. In contrast, in this paper, we propose to fundamentally revisit the concept of AI-native wireless systems, equipping them with the *common sense* necessary to transform them into *artificial general intelligence (AGI)-native systems*. Our envisioned AGI-native wireless systems acquire common sense by exploiting different cognitive abilities such as perception, analogy, and reasoning, that can enable them to effectively generalize and deal with unforeseen scenarios. The proposed AGI-native wireless system is mainly founded on three fundamental components: A perception module, a world model, and an action-planning component. Collectively, these three fundamental components enable the four pillars of common sense that include dealing with unforeseen scenarios through horizontal generalizability, capturing intuitive physics, performing analogical reasoning, and integrated information theory and hierarchical abstractions play a crucial role in the proposed intent-driven and objective-driven planning methods that maneuver the AGI-native network to plan its actions. Next, we discuss how an AGI-native network can be further exploited to enable three use cases related to human users and autonomous agents applications: a) analogical reasoning for next-generation DTs, b) synchronized and resilient experiences for cognitive avatars, and c) brain-level metaverse experiences exemplified by holographic teleportation. Finally, we conclude with a set of recommendations to ignite the quest for AGI-native systems. Ultimately, we envision this paper as a roadmap for the next-generation of wireless systems beyond 6G.**

***Index Terms*—artificial general intelligence (AGI), metaverse, AGI-native, cognitive avatars, AGI-augmented digital twins (DTs), reasoning, planning, common sense, beyond 6G**

I. Introduction

In the next decade, novel wireless use cases, such as the metaverse and holographic societies, are anticipated. Those use cases will largely strain the communication limits of modern-day wireless systems due to their unique performance requirements, which are quite different from conventional use cases like smartphone-centric services or intelligent transportation, that were the key drivers for 5G and early 6G systems [1].

Causal Reasoning: Charting a Revolutionary Course for Next-Generation AI-Native Wireless Networks

Christo Kurisummoottil Thomas, *Member, IEEE,* Christina Chaccour, *Member, IEEE,*
Walid Saad, *Fellow, IEEE,* Mérouane Debbah, *Fellow, IEEE,* and Choong Seon Hong, *Senior Member, IEEE*

***Abstract*—Despite the basic premise that next-generation wireless networks (e.g., 6G) will be artificial intelligence (AI)-native, to date, most existing efforts remain either qualitative or incremental extensions to existing "AI for wireless" paradigms. Indeed, creating AI-native wireless networks faces significant technical challenges due to the limitations of data-driven, training-intensive AI. These limitations include the black-box nature of the AI models, their curve-fitting nature, which can limit their ability to reason and adapt, their reliance on large amounts of training data, and the energy inefficiency of large neural networks. In response to these limitations, this article presents a comprehensive, forward-looking vision that addresses these shortcomings by introducing a novel framework for building AI-native wireless networks; grounded in the emerging field of causal reasoning. Causal reasoning, founded on causal discovery, causal representation learning, and causal inference, can help build explainable, reasoning-aware, and sustainable wireless networks. Embracing such a human-like AI foundation can revolutionize the design of AI-native wireless networks, laying the foundations for creating self-sustaining networks that ensure uninterrupted connectivity. Towards fulfilling this vision, we first highlight several wireless networking challenges that can be addressed by causal discovery and representation, including ultra-reliable beamforming for terahertz (THz) systems, near-accurate physical twin modeling for digital twins, training data augmentation, and**

machine learning (ML) and AI algorithms to design, optimize, and operate various aspects of the wireless system, including transceiver design, resource allocation, interference management, and many more. However, developing AI native 6G wireless systems necessitates the design of novel AI frameworks that are tailored to several unique challenges of wireless systems: 1) *dynamic adaptability* to ensure rapid adjustments to changing network conditions, user demands, and other environmental factors; 2) *time criticality*, whereby 6G systems must deliver ultra-low latency and unwavering reliability, particularly for applications demanding split-second responsiveness; 3) *intent management*, enabling networks to autonomously translate high-level business intents into network configurations in a closed-loop fashion, ensuring intent assurance across the network while maintaining overall network reliability; 4) *resilience*, enabling 6G networks to withstand disruptions and maintain connectivity even in challenging scenarios; 5) *non-linear signal dynamics*, that must be properly modeled to accurately capture the time-varying nature of multi-modal wireless signals, e.g., audio, video, haptics and olfactory signals; and 6) *human-level cognition and reasoning*

Reasoning over the Air: A Reasoning-based Implicit Semantic-Aware Communication Framework

Yong Xiao, *Senior Member, IEEE,* Yiwei Liao, Yingyu Li, Guangming Shi, *Fellow, IEEE,* H. Vincent Poor, *Life Fellow, IEEE,* Walid Saad, *Fellow, IEEE,* Mérouane Debbah, *Fellow, IEEE,* and Mehdi Bennis, *Fellow, IEEE*

***Abstract*—Semantic-aware communication is a novel paradigm that draws inspiration from human communication focusing on the delivery of the meaning of messages. It has attracted significant interest recently due to its potential to improve the efficiency and reliability of communication and enhance users' quality-of-experience (QoE). Most existing works focus on transmitting and delivering the explicit semantic meaning that can be directly identified from the source signal. This paper investigates the implicit semantic-aware communication in which the hidden information, e.g., hidden relations, concepts and implicit reasoning mechanisms of users, that cannot be directly observed from the source signal must be recognized and interpreted by the intended users. To this end, a novel implicit semantic-aware communication (iSAC) architecture is proposed for representing, communicating, and interpreting the implicit semantic meaning between source and destination users. A graph-inspired structure is first developed to represent the complete semantics, including both explicit and implicit, of a message. A projection-based semantic encoder is then proposed to convert the high-dimensional graphical representation of explicit semantics into a low-dimensional semantic constellation space for efficient physical channel transmission. To enable the destination user to learn and imitate the implicit semantic reasoning process of source user, a generative adversarial imitation learning-based solution, called G-RML, is proposed. Different from existing communication solutions, the source user in G-RML does not focus only on sending as much of the useful messages as possible; but, instead, it tries to guide the destination user to learn a reasoning mechanism to map any observed explicit semantics to the corresponding implicit semantics that are most relevant to the semantic meaning. By applying G-RML, we prove that the destination user can accurately imitate the reasoning process of the source user and automatically generate a set of implicit reasoning paths following the same probability distribution as the expert paths. Compared to the existing solutions, our proposed G-RML**

*This work has been accepted at IEEE Transactions on Wireless Communications. Copyright may be transferred without notice, after which this version may no longer be accessible.

An abridged version of this paper was presented at the Proc. of the IEEE ICC Workshop on Data Driven Intelligence for Networks and Systems, Seoul, South Korea, May 2022 [1]. The work of Yong Xiao was supported in part by the National Natural Science Foundation of China under Grant 62071193, in part by the Major Key Project of Peng Cheng Laboratory under Grant PCL2023AS1-2, and in part by the Key Research and Development Program of Hubei Province of China under Grant 2021EHB015. The work of Yingyu Li was supported in part by the Major Key Project of Peng Cheng Laboratory under grant PCL2023AS1-2, in part by the National Natural Science Foundation of China under Grant 62301516, in part by the Key Research and Development Program of Hubei Province under Grant 2021EHB015, and in part by the National Key Research and Development Program of China under Grant 2022YFB2903201. The work of Guangming Shi was supported in part by the National Natural Science Foundation of China under Grant 62293483 and Grant 61976169, in part by the National Key Research and Development Program of China under grant 2019YFA0706604, and in part by the Major Key Project of Peng Cheng Laboratory under Grant

AI Embodiment Through 6G: Shaping the Future of AGI

Lina Bariah and Mérouane Debbah

***Abstract*—In the ever-evolving field of technologies, the emergence of Artificial General Intelligence (AGI), often referred as strong artificial intelligence (AI), stands as a breakthrough in the realm of machines intelligence, promising to witness a new era of capabilities and possibilities. In particular, AGI ventures into human-level cognition, and expands to thinking, reasoning, and awareness. This imminent evolution is envisioned to be manifested through the embodiment of AI machines, allowing machines to transcend their purely computational nature and interact with the world through the different senses. Accordingly, AI agents will be grounded in the physical environment, going through subjective experiences and acquiring the needed knowledge that will lead to understanding and cognition. In our article, we explore the path towards realizing the true vision of AGI through AI embodiment, where we dig into the different types of thinking required to achieve knowledge, and hence, cognition and understanding. Furthermore, we look through the evolution of generative AI models, and shed lights on the limitations of auto-regression in large language models (LLMs), with the aim to answer the question: is sensory grounding (through 6G) necessary, and *enough*, to achieve understanding in LLMs? Finally, we identify the main pillars of AGI and unveil how 6G networks will orchestrate the development of AGI systems.**

I. Introduction

The culmination of years of artificial intelligence (AI) research has led to the emergence of artificial general intelligence (AGI) as the definitive frontier that aspires to mimic human intelligence from all aspects. The notion of AGI will rather possess a physical instantiation that will allow it to interact with the surrounding environment. Embodying AI has profound implications on AGI systems, in the sense that it will not only allow the AGI agents to understand the physical world, but to engage with it and learn from such an engagement.

The aim of this article is to delve deeper into the concept of AI embodiment as a gateway to the ultimate vision of AGI, where we open avenues for exploring how sensory grounding can forge a robust link between the machines and the environment, and therefore, paves the way to machine understanding. We set the scene for the true vision of AGI, where we unfold the reason why 6G is essential to realize this vision, and how 6G will be the key to the convergence of computing, collective intelligence, reinforcement learning (RL), sensing, and virtualization, in pursuit of AGI. To the best of the authors' knowledge, this is the first article to approach the AGI paradigm from embodiment, LLMs, and 6G perspectives.

A. Slow & Fast Thinking: Two Ways to Knowledge

Humans brains comprise two cognition systems, system I, referred as fast thinking, which is intuitive, automatic, and tied to heuristics and past experiences. Meanwhile, system II (slow thinking) is slow, analytical, and tied to reasoning and

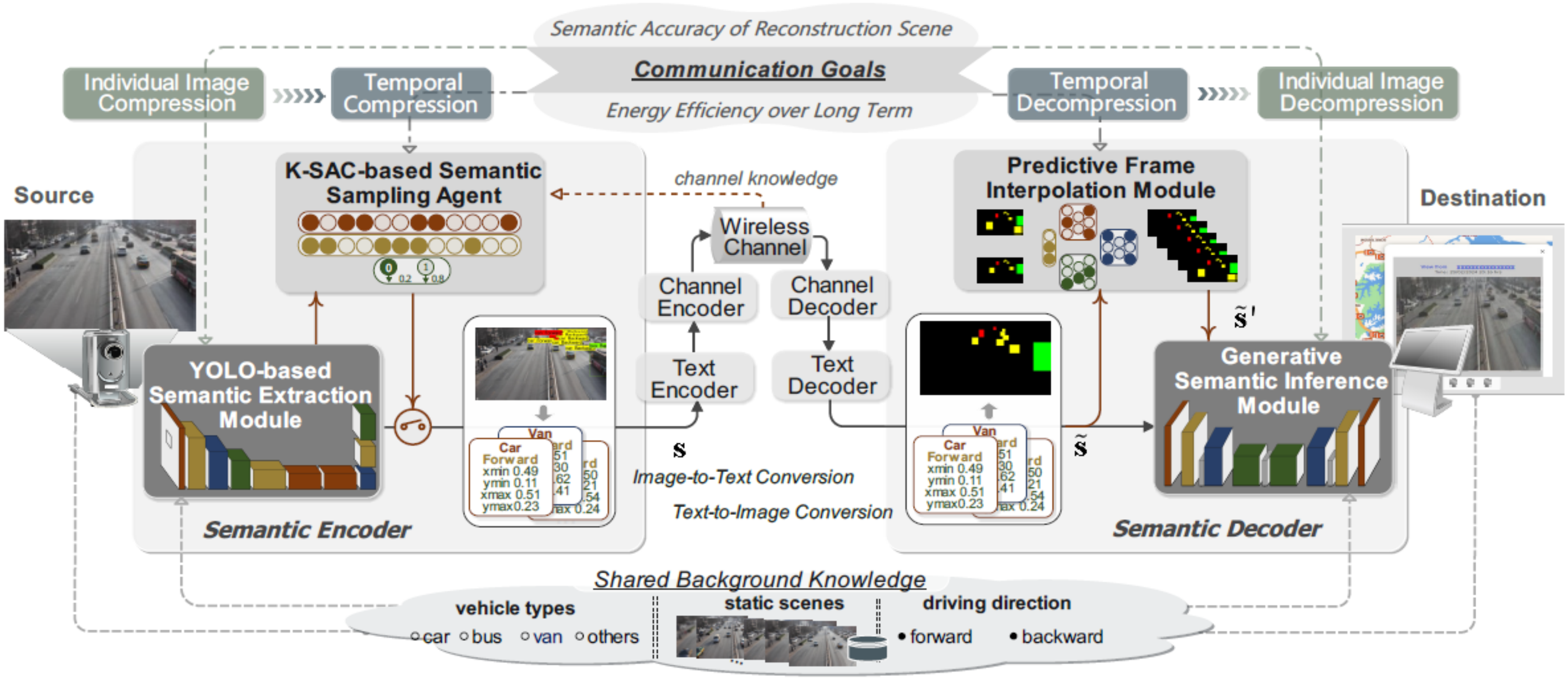
6G: Connected Intelligence

“G” Waves

- 2G: Mobile for Voice
- 3G: Mobile for Data
- 4G: Mobile for Internet
- 5G: Mobile for Things
- 6G : Mobile for Intelligence

	Human	Machine
Display resolution	< 290 ppi	Limits??
Display refresh rate	< 60 fps	
3D effect	Left / right image	
Latency	< 100 ms	
Audible frequency	250 ~ 2000 Hz	
Visible frequency	280 ~ 780 nm	
Viewing angle	Horizontal 200°, vertical 130°	
Senses	5 senses (see, hear, smell, touch, taste)	

Semantic Communication: LLM for Compression



Telecom AI agents: end of apps…

Media | 02-15-2024 | Niels Hafenrichter | 2 Comments

AI phone: Deutsche Telekom wants to free smartphones from apps

Share Print Read out

- Visionary showcase at MWC 2024 shows the world of an app-free AI smartphone
- Digital assistant helps in (almost) all situations in life
- Cooperation with Qualcomm and Brain.ai

Portable Large Language Models – not the iPhone 15 – are the future of the smartphone

Personal AI can redefine the handheld experience and perhaps preserve privacy too

Mark Pesce Wed 13 Sep 2023 // 07:38 UTC

COLUMN Smartphone innovation has plateaued. The iPhone 15, launched overnight, has some nice additions. But my iPhone 13 will meet my needs for a while and I won't rush to replace it. My previous iPhone lasted four years.

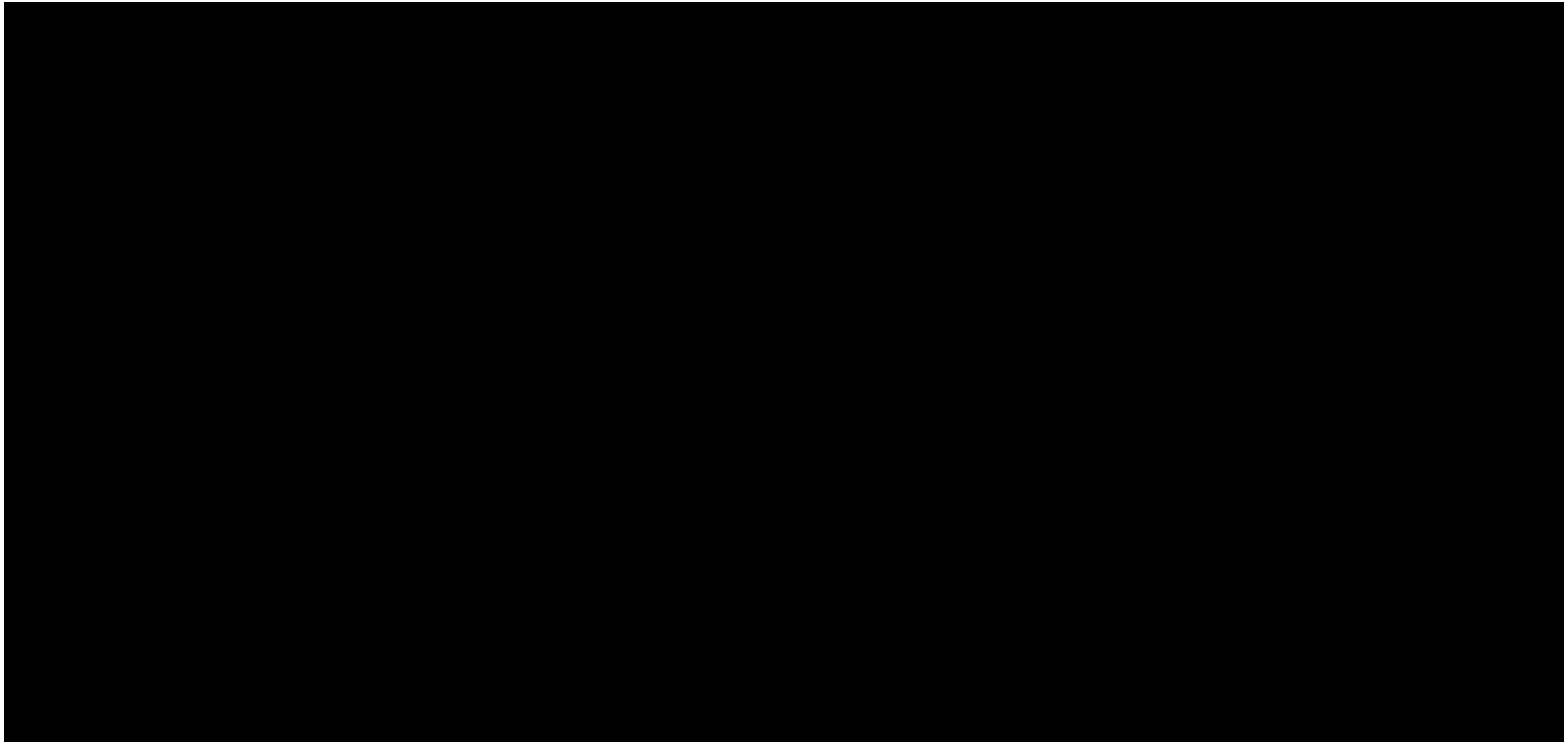
Before that phone I could justify grabbing Cupertino's annual upgrade. These days, what do we get? The iPhone 15 delivered USB-C, a better camera, and faster wireless charging. It's all nice, but not truly necessary for most users.

Yet smartphones *are* about to change for the better – thanks to the current wild streak of innovation around AI.

Pretty much everyone with a smartphone can already access the "Big Three" AI chatbots – OpenAI's ChatGPT, Microsoft's Bing Chat and Google's Bard – through an app or browser.

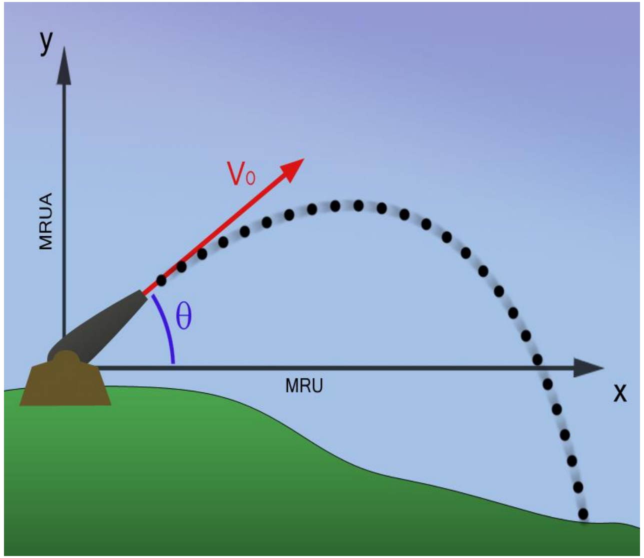
That works well enough. Yet alongside these "general purpose" AI chatbots, a

Telecom AI agents: end of apps…and the rise of LLM solvers



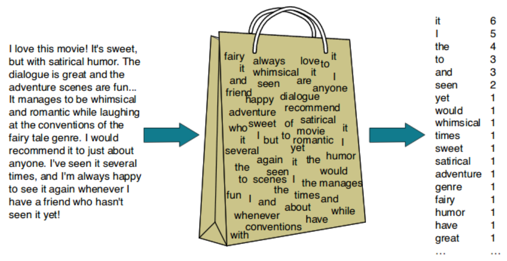
2023: The Generative AI Revolution

In a nutshell…

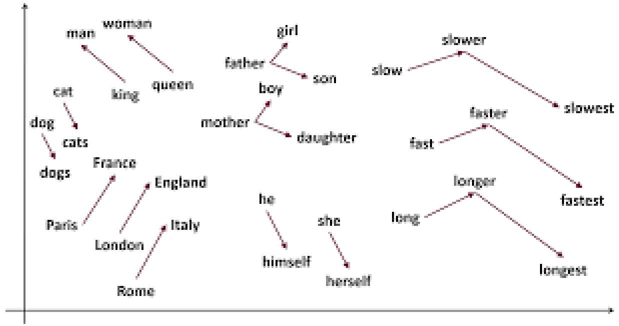


NLP In The Past
Statistical NLP

Bag of words



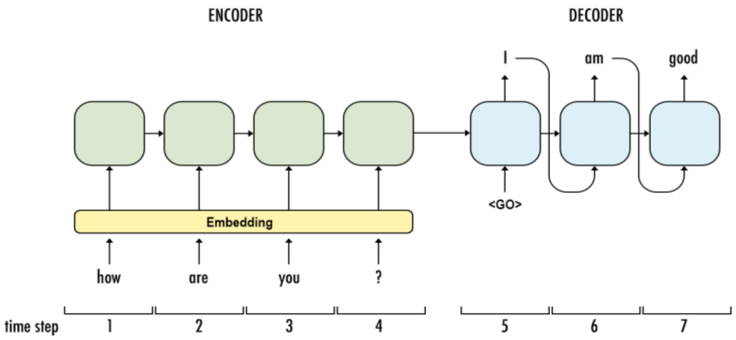
Word2vec

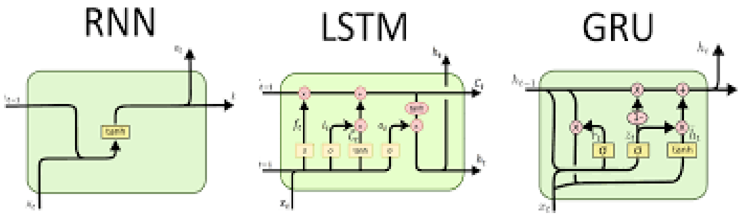


Most of natural language processing systems were based on simple statistical rules or non-complex Machine learning algorithms. The capabilities of these systems were limited to few tasks.

NLP In The Past
Before 2017

Seq-to-Seq modeling



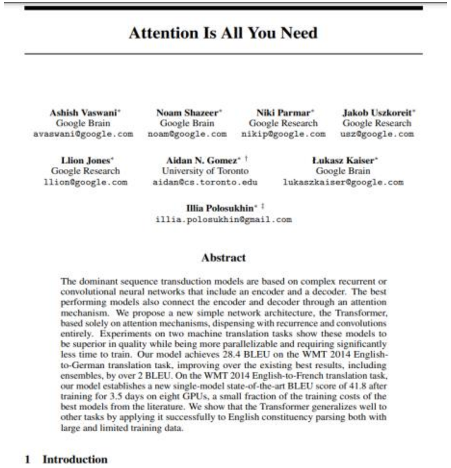


Emergence of new tasks with these new architectures :
- Translation
- Summarization
- Text completion
- …

- Attention to the rescue
  - Cannot learn Long dependencies
  - Fails in Long sentences
- Recurrent
  - Sequential
  - Parallelization : not parallelizable

NLP Today
Attention Is All You Need, 2017

Google Brain, Google Research, and University of Toronto



Attention Is All You Need

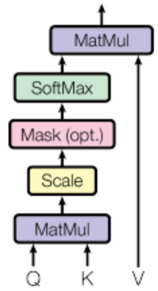
Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

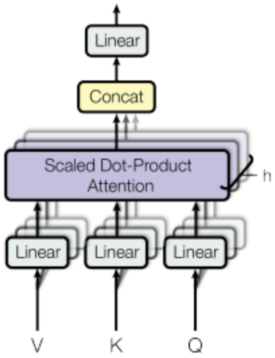
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

1 Introduction

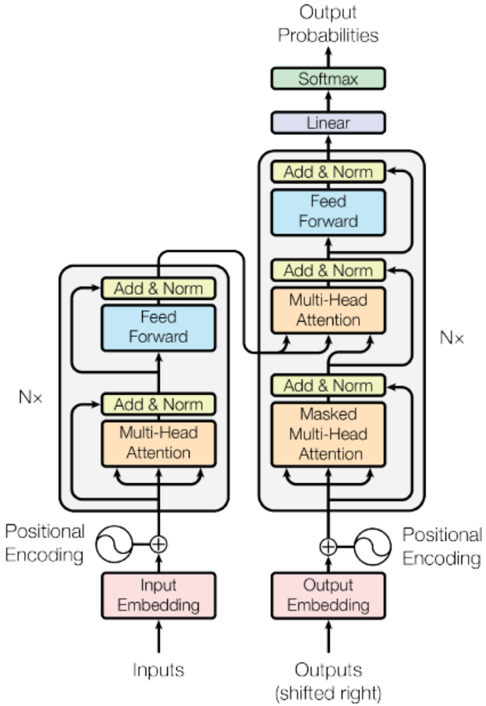
Attention mechanism





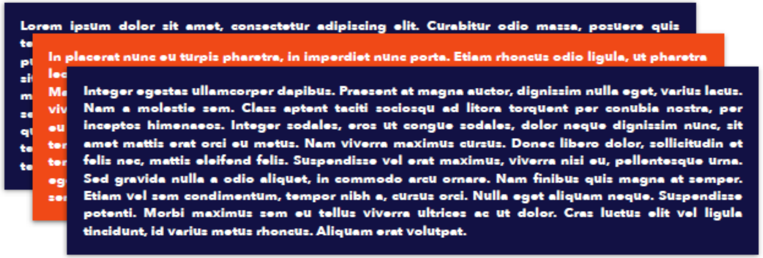
$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

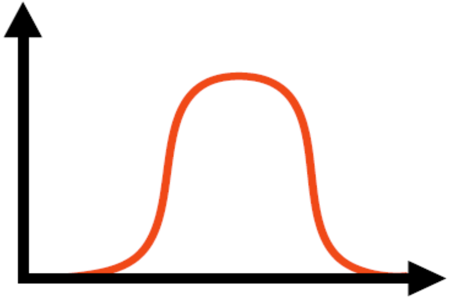
Transformers



- **Can learn Long dependencies**
- **Parallelizable**

What Does Modeling Language Really Mean?





Text corpus

Conditional distribution

The capital of the United Arab Emirates is Abu X

$\boxed{P(X \mid \underline{\text{context}})}$

Autoregressive/causal modelling

How Do We Learn/Model This Conditional Probability P?

🧠 Learning is easy: it's a self-supervised problem!

we can crawl vast amount of texts and simply predict the next word.

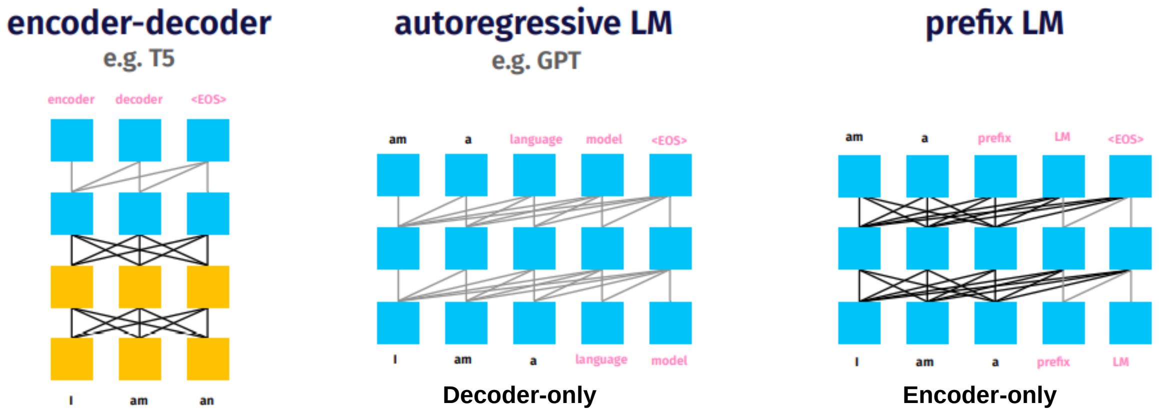
1st reason why extreme-scale is possible!

🤔 We need a data structure to learn the distribution... neural network?

MLP/FC	RNNs	Transformers
❌ not sequence-aware	❌ not scalable SOTA until ~2018, but slow & hard to train, no long-term dependencies	✅

Natural Language Processing: The Age of Transformers
Attention Mechanisms Revolutionized The Way We Do NLP

SOTA NLP models today are composed of a set of stacked transformers : Encoder-based, Decoder-based or Encoder-Decoder models.

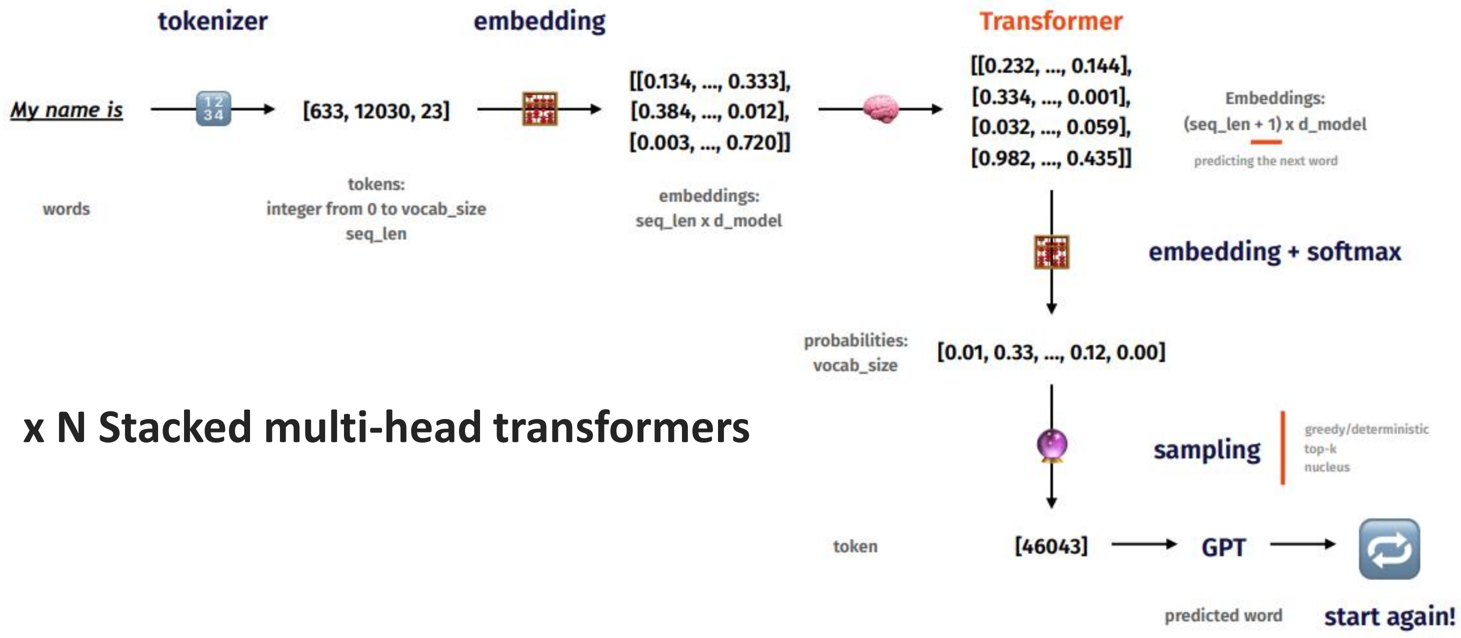


- **Self-supervised models : MLM, next word, sentences order …**
- **Parallelizable with teacher forcing**

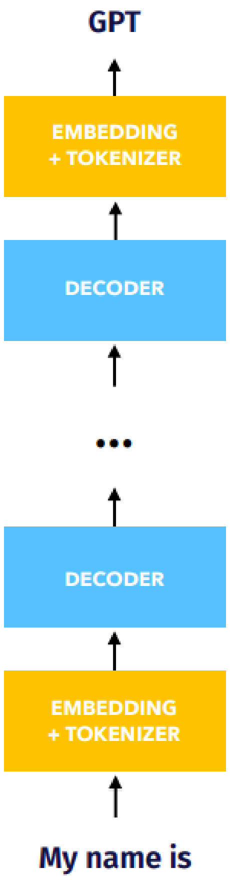
Examples of transformers-based language models :

- **BERT**
- **ROBERTA**
- **GPT-2**
- **T5**
- **GPT-3**
- **PagNol**
- **Megatron-Turing NLG**
- **Noor**

Focus on Decoder-only Architectures (E.g. GPT Models) Predicting The Next Token



x N Stacked multi-head transformers



Cost of the Models

Model	Training end	Chip type	TFLOP/s (max)	Chip count	Wall clock (days)	Total time (years)	Retail (US$)	MMLU
GPT-3 175B	Apr/2020	V100	130	10,000	15 days	405y	$9M	43.9
Llama 1 65B	Jan/2023	A100	312	2,048	21 days	118y	$4M	63.4
Llama 2 70B	Jun/2023	A100	312	*2,048*	35 days	196y	$7M	68.0
Titan 200B	Apr/2023	A100	312	13,760	48 days	1,319y	$45M	70.4
GPT-4 1.7T	Aug/2022	A100	312	25,000	95 days	6,507y	$224M	86.4
Gemini	Nov/2023	TPUv4	275	57,000	100 days	15,616y	$440M	90.0
Llama 3 405B	Apr/2024	H100	989	24,576	50 days	3,366y	$125M	*85+*
GPT-5	*Apr/2024*	H100	989	*50,000*	*120 days*	*16,438y*	$612M	
Grok 2	*Jun/2024*	H100	989	20,000	*50 days*	*6,571y*	$245M	
Olympus	*Aug/2024*	H100	989					
Gemini 2	*Nov/2024*	TPUv6	*1,847*					
Grok 3	*Dec/2024*	H100	989	100,000	*50 days*	*32,855y*	$1.2B	

Alan D. Thompson. May/2024. LifeArchitect.ai

Table. Model training compute (see working, with sources[8]).

Choose your Weapon: Survival Strategies for Depressed AI Academics

Julian Togelius and Georgios N. Yannakakis*

April 14, 2023

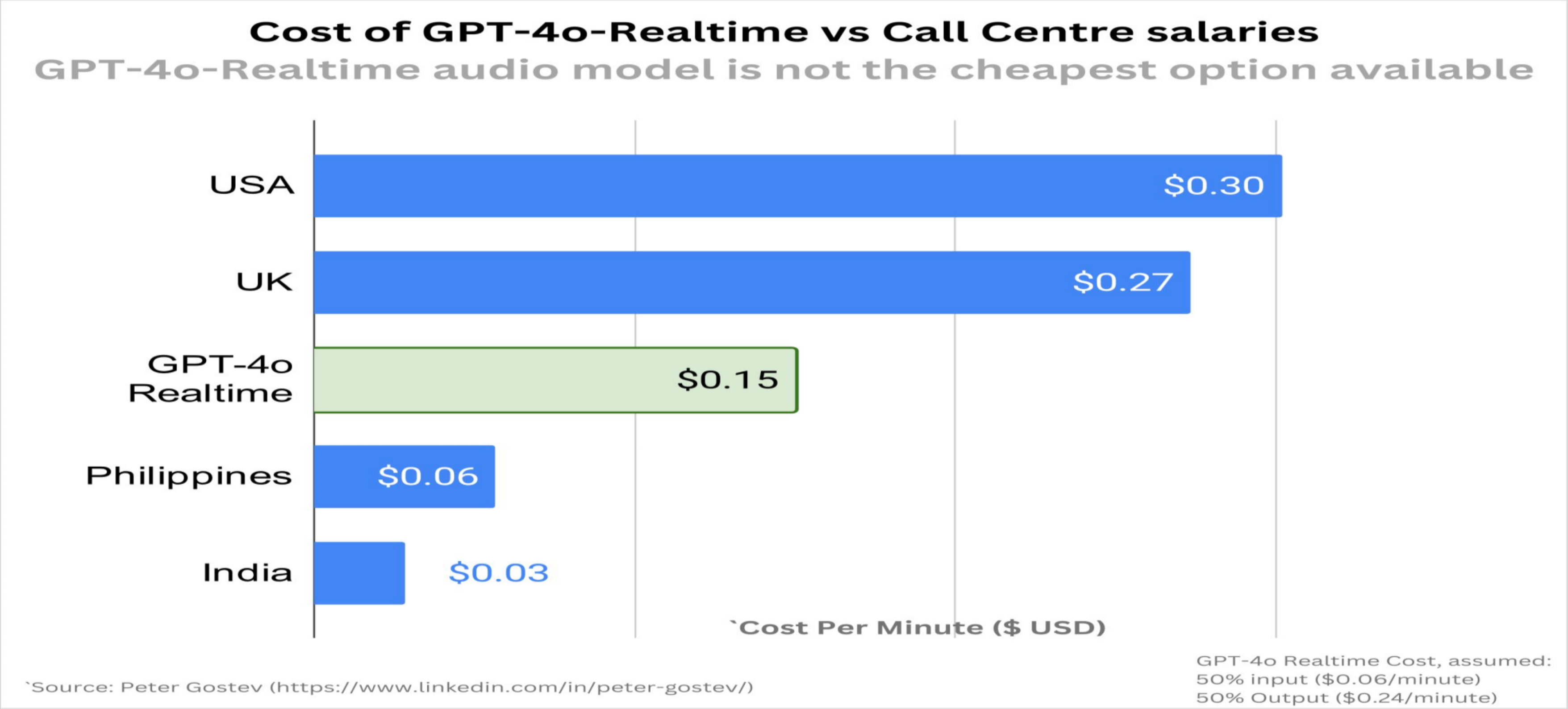
Abstract

Are you an AI researcher at an academic institution? Are you anxious you are not coping with the current pace of AI advancements? Do you feel you have no (or very limited) access to the computational and human resources required for an AI research breakthrough? You are not alone; we feel the same way. A growing number of AI academics can no longer find the means and resources to compete at a global scale. This is a somewhat recent phenomenon, but an accelerating one, with private actors investing enormous compute resources into cutting edge AI research. Here, we discuss what you can do to stay competitive while remaining an academic. We also briefly discuss what universities and the private sector could do improve the situation, if they are so inclined. This is not an exhaustive list of strategies, and you may not agree with all of them, but it serves to start a discussion.

1 Introduction

As someone who does AI research in a university, you develop a complicated relationship to the corporate AI research powerhouses, such as DeepMind, Open AI, Google Brain and Meta AI. Whenever you see one of these papers that train some kind of gigantic neural net model to do something you weren't

Inference Cost?



2023: The 6G Kick-off Race

2023: The Kick-off Race

ITU-R WP 5D agrees on "IMT-2030 Framework" (June 2023)

At its June 2023 meeting, ITU-R WP 5D *has agreed* the draft new Recommendation *"Framework and overall objectives of the future development of IMT for 2030 and beyond"**, which can be considered as the basis for the standardisation fora to develop the next generation of IMT standards.

This draft Recommendation addresses:

- Trends of IMT-2030
- Usage scenarios of IMT-2030
- Capabilities of IMT-2030
- Considerations of ongoing development

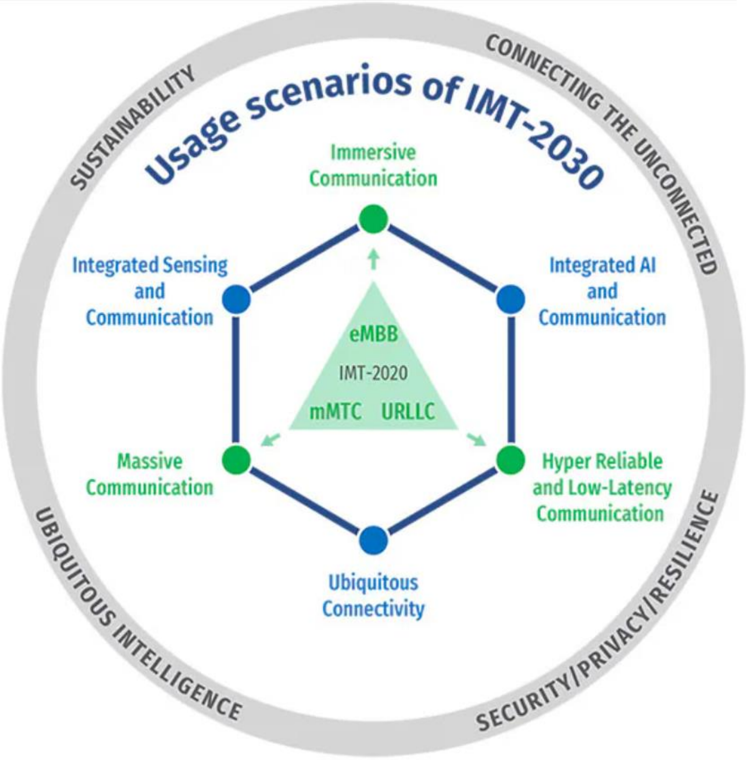
21 Nov 2023

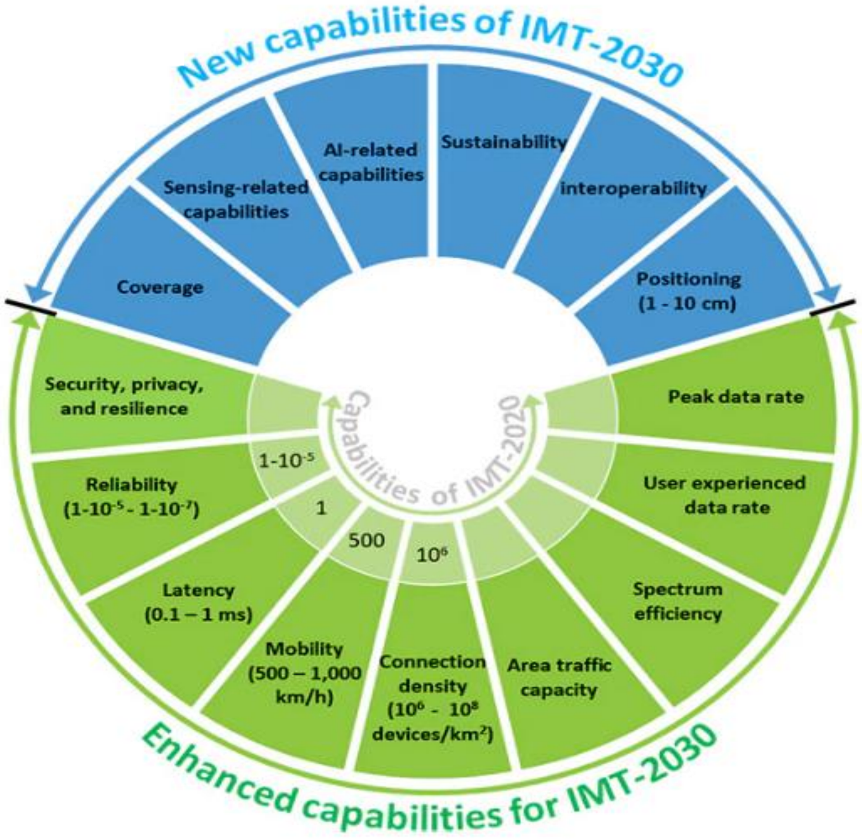
ITU adopts resolution to guide development of the 6G standard

The International Telecommunication Union (ITU) has adopted a resolution to guide the development of a 6G standard. The resolution is in the focus at the ongoing World Radiocommunication Conference (WRC-23) in Dubai.

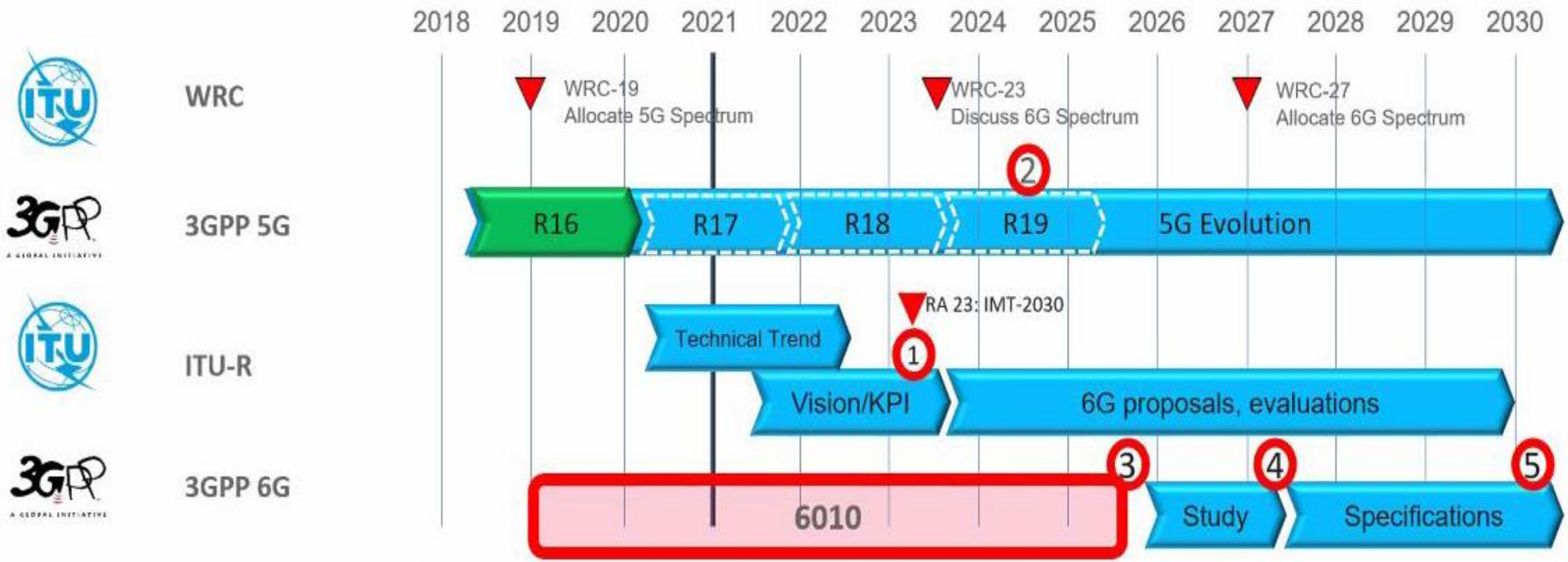
The International Telecommunication Union (ITU) has adopted ITU-R Resolution 65, which aims to guide the development of a 6G standard. This resolution enables studies on the compatibility of current regulations with potential 6th generation International Mobile Telecommunications (IMT) radio interface technologies for the year 2030 and beyond. The adoption of this resolution is particularly significant during the World Radiocommunication Conference (WRC-23), taking place in Dubai, where discussions are being held on radio regulations and frequencies essential for advancements in smart cities, the digital economy, knowledge society, and space.

Next-Gen Connectivity Metrics





6G Timeline



1. Per ITU process, vision finished before WRC-23
2. Early 2025, 6G Workshop @3GPP
3. Early 2026 to end of 2027, 3GPP study item
4. Early 2028 to early 2030, 3GPP work item
5. 1st specification of 6G finished in 3GPP @2030

Future Technologies

Broadband Connectivity

- Holographic MIMO
- Tbps coding
- OAM
- Cell Free Massive architectures
- Large Intelligent Surfaces
- mmWave
- LEO Satellite (3D Cellular)
- Beyond OFDM waveforms
- Ultra-Massive Access (Rate Splliting)
- Time Reversal

Sensing and Control

- Joint sensing and Communication Waveforms
- THz Communications
- Wireless Imaging
- Extreme URLLC
- Age of Information
- Beyond Sampling

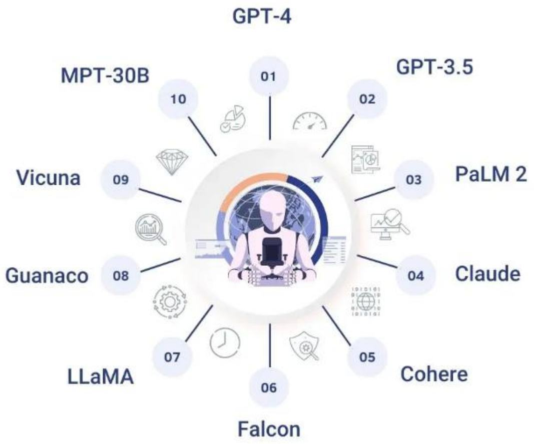
Intelligence

- AI Transceiver
- Multi-Agent Generative AI
- Distributed GPT
- Edge Learning
- Semantic source codes
- Semantic channel Codes

TelecomGPT Pillars

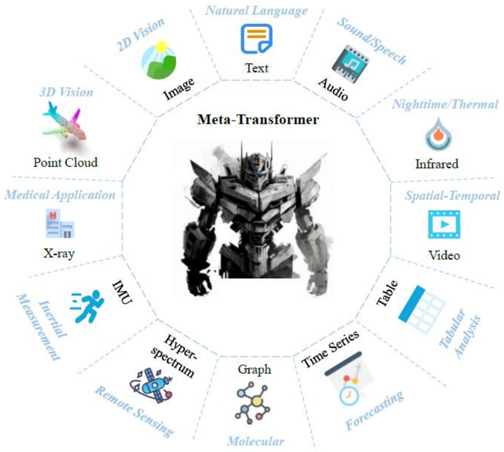
TELECOM FOUNDATION MODELS

Large Model at Small Devices



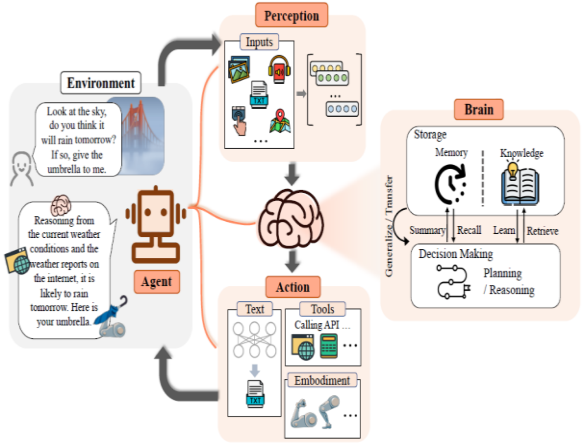
TELECOM BIG DATA

Multi-Modality



TELECOM AUTONOMY

Autonomous Agents



TelecomGPT

UAE's TelecomGPT: The AI Breakthrough Set to Transform Telecoms

17 JULY 2024

TELECOM Review
THE TELECOMS INDUSTRY MEDIA PLATFORM
telecomreview.com

TelecomGPT, designed explicitly for the telecom sector, outperforms general-purpose models like GPT-4 and Llama-3 in telecom-specific tasks, particularly in the Telecom Math Modelling benchmark.

This specialisation allows for improved performance in industry-related applications, such as Radio Resource Management (RRM) in Open Radio Access Network (O-RAN) and more sophisticated and adaptive communication protocols. These capabilities enhance the efficiency and resilience of telecom networks.

EMERGENT TECH July 17, 2024 09:26 AM GST

Why UAE's TelecomGPT could be a global game-changer in not just the telecom world but even the AI space

The Global Telco AI Alliance's introduction of TelecomGPT marks a pivotal moment for the telecom industry. This advanced AI model outshines general-purpose models like GPT-4 and is key to the UAE's 6G aspirations.

By Sindhu V Kashyap

Middle East AI News

Abu Dhabi researchers create first-of-its-kind telecom LLM

Khalifa University's 6G Centre and TII collaborate on building TelecomGPT

CARRINGTON MALIN
JUL 16, 2024

Share

Edge/

TelecomGPT: A Framework to Build Telecom-Specfic Large Language Models

Hang Zou[1], Qiyang Zhao[1], Yu Tian[1], Lina Bariah[2], Faouzi Bader[1], Thierry Lestable[1], and Merouane Debbah[1,2]

[1]Technology Innovation Institute, 9639 Masdar City, Abu Dhabi, UAE
[2]Khalifa University, Abu Dhabi 127788, UAE

***Abstract*—Large Language Models (LLMs) have the potential to revolutionize the Sixth Generation (6G) communication networks. However, current mainstream LLMs generally lack the specialized knowledge in telecom domain. In this paper, for the first time, we propose a pipeline to adapt any general purpose LLMs to a telecom-specific LLMs. We collect and build telecom-specific pre-train dataset, instruction dataset, preference dataset to perform continual pre-training, instruct tuning and alignment tuning respectively. Besides, due to the lack of widely accepted evaluation benchmarks in telecom domain, we extend existing evaluation benchmarks and proposed three new benchmarks, namely, Telecom Math Modeling, Telecom Open QnA and Telecom Code Tasks. These new benchmarks provide a holistic evaluation of the capabilities of LLMs including math modeling, Open-Ended question answering, code generation, infilling, summarization and analysis in telecom domain. Our fine-tuned LLM TelecomGPT outperforms state of the art (SOTA) LLMs including GPT-4, Llama-3 and Mistral in Telecom Math Modeling benchmark significantly and achieve comparable performance in various evaluation benchmarks such as TeleQnA, 3GPP technical documents classification, telecom code summary and generation and infilling.**

***Index Terms*—Generative AI, Large Language Models, 3GPP, Telecom Foundation Models**

I. Introduction

Recent advances of Large Language Models (LLMs) have attracted significant attention across different domains including telecommunications community. LLMs, such as GPT-4 [1], the Llama series [2]–[4], the Mistral series [5], [6], and the Falcon series [7], have demonstrated remarkable capabil-

LLM-QAT [8], one bit quantization in BitNet [9] makes it possible to deploy mainstream LLMs on edge devices, e.g., mobile phones. Second, the long inference time of LLMs is unbearable to meet the requirement of Ultra Reliable and Low Latency Communication (URLLC) in beyond 5G networks. Taking the example of Vehicle to Everything (V2X) communication networks, it would be impossible for autonomous vehicles to wait for the generation completion of LLMs when taking crucial decision or transmitting important information to surrounding vehicles. Inference acceleration techniques and architectures on both system level and algorithm level [10] e.g., KV caching [11], FlashAttention [12] and Mixture of Experts (MoEs) [13] could largely increase the throughput of LLMs (tokens per second) to alleviate this issue. Finally, even physical challenges such as insufficient memory and low high latency are mitigated by the combination of various techniques, it remains a fundamental difficulty for LLMs to accomplish telecom-specific tasks in wireless networks due to a general lack of knowledge in telecom domain. Therefore, it would be natural to anticipate the existence of telecom-specialized or telecom-specfic LLMs, which is exactly the core problem this paper tries to tackle with. Before diving into the technical details of our proposed methods, we briefly review the recent advances including domain-specific LLMs, applications of LLMs in telecom and the the challenge of building telecom-specific LLMs.

OpenTelecom Data Pipeline

Source Collection

Extraction & Preprocessing

Nougat

UnstructuredIO

https://facebookresearch.github.io/nougat/

Get text from specific tags

DSIR

Select relevant documents w.r.t. a seed Telco dataset

https://github.com/p-lambda/dsir

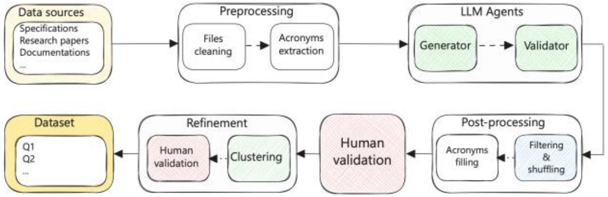
Discard lines containing too many uppercases numbers / punctuations / duplicated words / spaces

Discard lines with emails / phone numbers / incomplete sentences

Keep english corpus only with fasttext

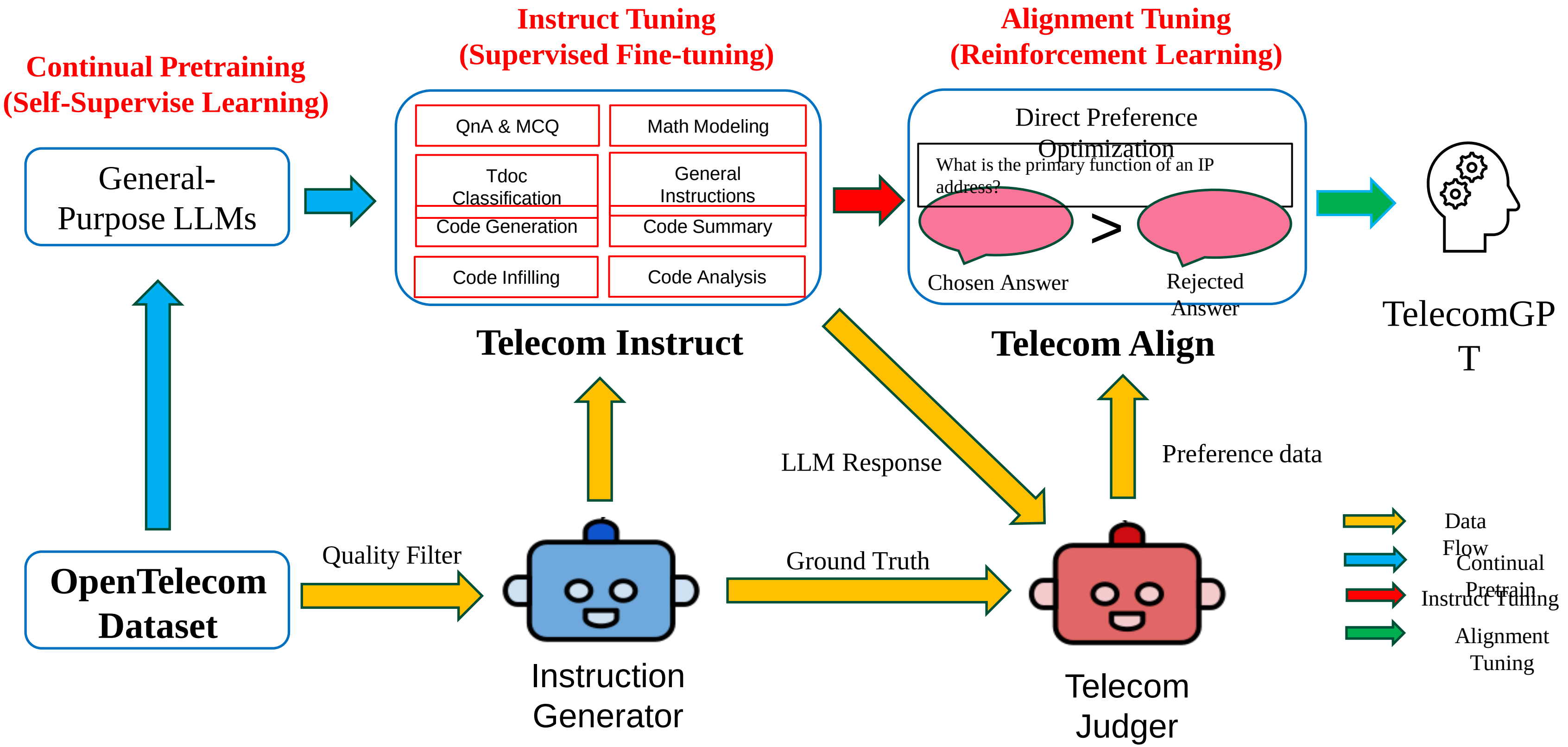
Fuzzy deduplication with LSH-MinHash & exact deduplication

Task Specific Data Creation



Instruction	Identify the issue with the router's configuration and provide a solution.
Input	Pinging 192.168.1.1 with 32 bytes of data: Reply from 74.x.x.254: TTL expired in transit. Reply from 74.x.x.254: TTL expired in transit. Reply from 74.x.x.106: TTL expired in transit. Reply from 74.x.x.254: TTL expired in transit.Ping statistics for 192.168.1.1: Packets: Sent = 4, Received = 4, Lost = 0 (0% loss)
Output	The issue is that a WAN port is connected to an existing network, and the router received an IP from DHCP along with a default gateway. When trying to ping 192.168.1.1, the packet is routed to the default gateway, which then routes the packet out to the Internet. The solution is to connect the correct type of port (LAN or WAN) to the existing network, and ensure the router is configured with the correct IP address and default gateway settings.
Original Text	Q: I just got this router, fresh out of the box. I turn it on and set it up and I get these weird IPs: Pinging 192.168.1.1 with 32 bytes of data: Reply from 74.x.x.254: TTL expired in transit. Reply from 74.x.x.254: TTL expired in transit. Reply from 74.x.x.106: TTL expired in transit. Reply from 74.x.x.254: TTL expired in transit. Ping statistics for 192.168.1.1: Packets: Sent = 4, Received = 4, Lost = 0 (0% loss), Never seen this before. Is this a Cisco thing? A: My best guess here without more information is that you hooked up a WAN port to your existing network. The new router got an IP from DHCP, along with a default gateway. When you try to ping 192.168.1.1, the packet is routed to your default gateway. For whatever reason, your network routes the packet out to the Internet, where it loops around in no-man's land until the 30-hop TTL expires. A traceroute to the same IP should tell us more about what's going on in case my guess is incorrect. =)
Corpus	Network Engineering from Stack Exchange
Model Used	mixtral

TelecomGPT: Training Pipeline



OpenTelecom Dataset

Overview of pretraining Telecom dataset (M Tokens)

	Training	Validation	Percentage
3GPP standard	193	1.9	11.49
IEEE standard	7.5	0.07	0.45
Paper (arxiv)	893	9	53.17
Book	1.9	0.02	0.11
Patent (c4)	253.2	2.6	15.08
StackExchagne	51.9	0.5	3.09
Wikipedia	18.9	0.2	1.13
Code (github)	260.1	2.6	15.49
Total	1679.5	16.89	-

Training Parameters

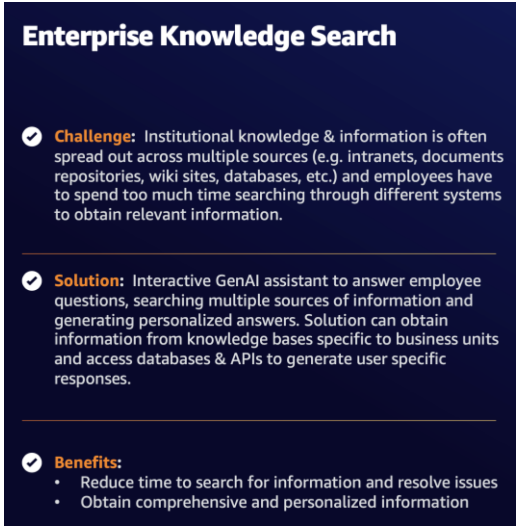
	Contine Pretrain	Instruct Tuning	Alignment Tuning
Training	Self-Supervised Learning	Supervised Fine-Tuning	Reinfocement Learning
Epoches	1	3	1
Steps	93000	1020	180
PEFT	Full Parameters	QLoRA (2%)	QLoRA (2%)
Learning rate	2e-5 Constant	2e-5 Constant	5e-6 Cosine Decay
Training data	3263 M tokens	5M tokens	8K tokens
GPU hours (A100)	1536	46	16

Use Case & Benchmark: Telecom Knowledge Query

Queries of Telecom standads, research papers, patents, lexicons

Question	What does IM stand for, and why is it considered desirable for constellations in communication systems?
Context (optional)	Option 1: Intermittent Messaging; because it allows for messages to be sent intermittently. Option 2: Index Modulation; because its combinatorial structure enables higher SNR efficiency, low-complexity implementation, and compatibility with existing coding and modulation techniques. Option 3: Instant Messaging; because it allows for real-time communication Option 4: Internal Modulation; because it modulates signals within the device itself.
Answer	Index Modulation (IM) is desirable for communication system constellations due to its higher signal-to-noise ratio efficiency compared to conventional techniques, low-complexity implementation, and compatibility with existing coding and modulation techniques.
Category	RAN1

Potential application
(Source: AWS)



Use Case & Benchmark : Telecom Task Completion and Planning

Explain techniques, identify problems, propose solutions in Telecom

Prompt	Describe the process of providing authorization to a mobile device for a pending transaction request.
Response	To provide authorization to a mobile device for a pending transaction request, the system searches the database for the mobile device's phone number, generates and sends an authorization request message to the mobile device, and awaits a reply message indicating approval or denial of the transaction from the user.

Plan tasks in Telecom specificaiton to complete a high-level instruction

Prompt	Initiate event based charging with decentralized and centralized unit determination, centralized rating
Response	1. Request for resource usage: A request for session establishment is received in the NF (CTF). The service is configured to be authorized by the CHF to start. 2. Units Determination: the NF (CTF) determines the number of units depending on the service requested by the UE in "Decentralized Units determination" scenario. 3. Charging Data Request [Event, Units]: The NF (CTF) sends the request to the CHF for th service to be granted authorization, and to allow the number of units, if determined in item 2, to be rated and accounted.

Potential application (Source: AWS)

Field Technician Assistant

- **Challenge:** Field technicians encounter multiple challenges during their work in the field, from technical problems to unforeseen situations that requires consulting documentation and procedures. Failure to obtain timely information in the field generate delays and repeat visits that impact costs and customer experience
- **Solution:** Interactive GenAI assistant that can provide timely information to field technicians on their mobile phones, searching through technical manuals, standard procedures, etc. The GenAI assistant also has access to specific information related to the current field assignment and past field work performed to provide relevant suggestions
- **Benefits:**
  - Reduce operational costs, service time and repeat visits
  - Improve field technician experience and increase succeed rates of field visits

Use Case & Benchmark : Telecom Code Understanding

Analyze code functionality and operational logic of Telecom features

Prompt

Analyze this script's contribution to telecom workflows, specifying its primary functionalities.

```
def bsc(input_bits, p_t):
    """
    Binary Symmetric Channel.

    Parameters
    ----------
    input_bits : 1D ndarray containing {0, 1}
        Input array of bits to the channel.

    p_t : float in [0, 1]
        Transition/Error probability of the channel.

    Returns
    -------
    output_bits : 1D ndarray containing {0, 1}
        Output bits from the channel.
    """
    output_bits = input_bits.copy()
    flip_locs = (random(len(output_bits)) <= p_t)
    output_bits[flip_locs] = 1 ^ output_bits[flip_locs]
    return output_bits
```

Response

The script defines a function `bsc` for Binary Symmetric Channel within the telecom domain. It takes `input_bits` array of bits and `p_t` transition/error probability as parameters and returns `output_bits` after applying the channel operations. The function simulates a binary symmetric channel by flipping bits based on the given probability, which is a crucial concept in telecom communication systems for modeling channel errors and noise.

Potential application (Source: AWS)

Developer Productivity

- **Challenge**: Solution developers have to spend significant amounts of time repetitive tasks like developing unit tests, creating documentation and maintaining legacy code. These tasks keep developers from focusing on building new solutions that bring business value
- **Solution**: Leverage the power of GenAI to generate code from plain instructions, explain, document or debug existing code and improve the quality and test coverage of solutions.
- **Benefits**:
  - Improve developer productivity, reducing the time to create and maintain solutions
  - Improve quality and consistency of code

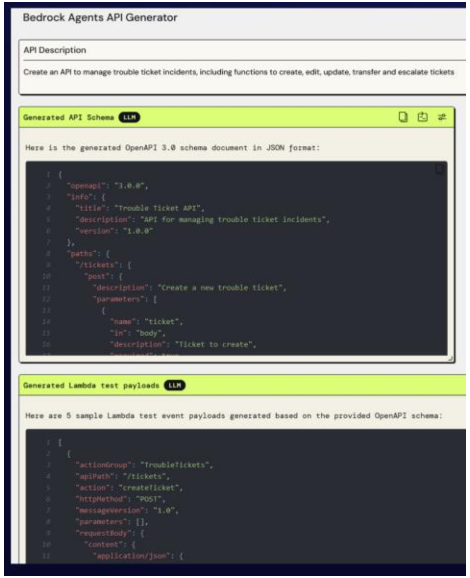
Use Case & Benchmark : Telecom Code Generation

Generate script or function for a given task or functionality in Telecom

Prompt	Develop a C function that retrieves and returns the transmit (TX) power value for a wireless device. The function should update a data structure representing wireless request (wrqu) with the TX power information including the power value, the fixed state, flags, and disabled status based on the device's status. This function is intended to be used in the context of configuring and managing wireless interfaces within a telecommunications network.
Response	*(code below)*

```c
union iwreq_data *wrqu, char *extra)
{
    struct ipw_priv *priv = ieee80211_priv(dev);
    mutex_lock(&priv->mutex);
    wrqu->power.value = priv->tx_power;
    wrqu->power.fixed = 1;
    wrqu->power.flags = IW_TXPOW_DBM;
    wrqu->power.disabled = (priv->status & STATUS_RF_KILL_MASK) ? 1 : 0;
    mutex_unlock(&priv->mutex);
    IPW_DEBUG_WX("GET TX Power -> %s %d \n",
                 wrqu->power.disabled ? "OFF" : "ON", wrqu->power.value);
    return 0;
}
```

Potential application
(Source: AWS)



Use Case & Benchmark : Telecom Math Modeling

Generate math solution according to Telecom scenario descriptions

Prompt

Please fill in the missing mathematical equation represented by $< \mathrm{MASK} >$ placeholder in the provided sections of a telecom-relevant research paper that discuss the system model or problem formulation.

System Model: We consider a discrete-time zero-mean jointly proper Gaussian flat-fading channel with the following input-output relation

$$< \mathrm{MASK} >$$

with the diagonal matrices $\mathbf{H} = \mathrm{diag}(\mathbf{h})$ and $\mathbf{X} = \mathrm{diag}(\mathbf{x})$. Here the $\mathrm{diag}(\cdot)$ operator generates a diagonal matrix whose diagonal elements are given by the argument vector. The vector $\mathbf{y} = [y_1, \ldots, y_N]^T$ contains the channel output symbols in temporal order. Analogous, $\mathbf{x} = [x_1, \ldots, x_N]^T$, $\mathbf{n} = [n_1, \ldots, n_N]^T$, and $\mathbf{h} = [h_1, \ldots, h_N]^T$ contain the channel input symbols, the additive noise samples and the channel fading weights. All vectors are of length N. The samples of the additive noise process are assumed to be i.i.d. zero-mean jointly proper Gaussian with variance σ_n^2 and, thus, $\mathbf{R}_n = \mathrm{E}\left[\mathbf{n}\mathbf{n}^H\right] = \sigma_n^2 \mathbf{I}_N$, with $\mathbf{I}_N$ being the identity matrix of size $N \times N$. The channel fading process is zero-mean jointly proper Gaussian with the temporal correlation characterized by

$$< \mathrm{MASK} >$$

Its variance is given by $r_h(0) = \sigma_h^2$. For mathematical reasons we assume that the autocorrelation function $r_h(l)$ is absolutely summable, i.e.,

$$< \mathrm{MASK} >$$

Response

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} = \mathbf{X}\mathbf{h} + \mathbf{n}$$

$$r_h(l) = \mathrm{E}[h_{k+l} \cdot h_k^*].$$

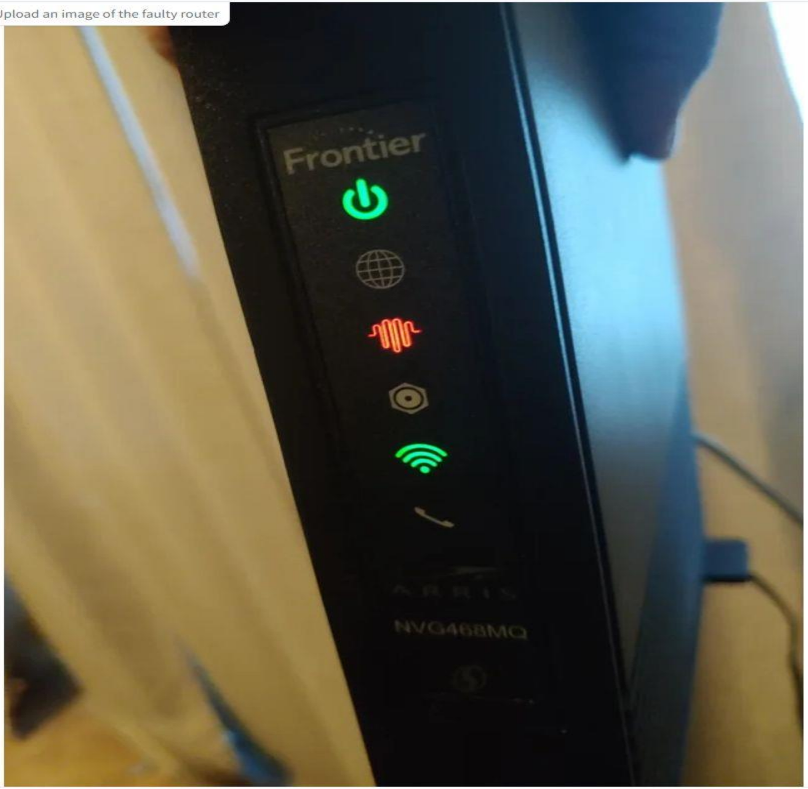
$$\sum_{l=-\infty}^{\infty} |r_h(l)| < \infty.$$

Spaces yasserrmd / RouterFix like 0 Running on ZERO Logs App Files Community Settings

Router Diagnosis

Upload a photo of your router to receive a professional diagnosis and troubleshooting steps displayed in a structured, easy-to-read format.

Upload an image of the faulty router



Clear

Submit

Issue Description The image shows a Frontier router with a series of lights indicating a potential issue. The lights are red and green, which typically indicate a problem with the router's connection or configuration.

Root Cause Analysis

- **LED Analysis**
  - Color: Red and Green
  - Pattern: Alternating red and green lights
  - Indicates: Potential connection or configuration issue
- **Error Code:** None visible
- **Possible Cause:** The router may be experiencing a connection issue or configuration problem.

Step-by-Step Troubleshooting

1. **Restart the Router**
   - Action: Press the power button to turn off the router, wait for 30 seconds, and then press it again to turn it back on.
   - Details: This step is often the simplest and most effective way to resolve connectivity issues.
   - Expected Outcome: The router should boot up normally and the lights should stabilize.
2. **Check Ethernet Cable**
   - Action: Inspect the Ethernet cable connecting the router to the modem for any signs of damage or loose connections.
   - Details: Ensure the cable is securely plugged into both the router and modem.
   - Expected Outcome: A secure connection should be established, and the lights on the router should stabilize.
3. **Check Wi-Fi Settings**
   - Action: Log in to the router's web interface and check the Wi-Fi settings to ensure they are correct and not set to "off" or "disabled". +

Use via API · Built with Gradio

Generative AI for RF Sensing in IoT systems

Li Wang, Chao Zhang, Qiyang Zhao, Hang Zou,
Samson Lasaulce, Giuseppe Valenzise, Zhuo He, and Merouane Debbah

***Abstract*—The development of wireless sensing technologies, using signals such as Wi-Fi, infrared, and RF to gather environmental data, has significantly advanced within Internet of Things (IoT) systems. Among these, Radio Frequency (RF) sensing stands out for its cost-effective and non-intrusive monitoring of human activities and environmental changes. However, traditional RF sensing methods face significant challenges, including noise, interference, incomplete data, and high deployment costs, which limit their effectiveness and scalability. This paper investigates the potential of Generative AI (GenAI) to overcome these limitations within the IoT ecosystem. We provide a comprehensive review of state-of-the-art GenAI techniques, focusing on their application to RF sensing problems. By generating high-quality synthetic data, enhancing signal quality, and integrating multi-modal data, GenAI offers robust solutions for RF environment reconstruction, localization, and imaging. Additionally, GenAI's ability to generalize enables IoT devices to adapt to new environments and unseen tasks, improving their efficiency and performance. The main contributions of this article include a detailed analysis of the challenges in RF sensing, the presentation of innovative GenAI-based solutions, and the proposal of a unified framework for diverse RF sensing tasks. Through case studies, we demonstrate the effectiveness of integrating GenAI models, leading to advanced, scalable, and intelligent IoT systems.**

***Index Terms*—Generative AI, RF sensing, cross-modal estimation, multi-modal fusion, large language models.**

I. Introduction

With the development of the Internet of Things (IoT), many kinds of wireless sensing signals (e.g., Wi-Fi, Infrared images, visible images, Radio Frequency (RF) signal) are filling our living and working spaces nowadays. Recently, researchers have also utilized RF signals to capture events in the IoT environment (i.e., RF sensing). While RF signals are transmitted, reflected, blocked, and scattered by objects like walls, furniture, vehicles, and human bodies, it is possible to extract useful information, such as position, movement direction, speed, and vital signs of a human subject, from received RF signals. Unlike traditional hardware sensors, RF sensing provides users with low-cost and unobtrusive services. Furthermore, due to the broadcast nature of RF signals, RF sensing can be used not only to monitor multiple subjects, but also to capture changes in the environment over a large area costs of deploying and maintaining extensive sensor networks make large-scale implementations expensive. Additionally, unstable environments cause signal weakening and multipath propagation, reducing reliability. These challenges necessitate advanced solutions like Generative AI (GenAI) to enhance the robustness, efficiency, and scalability of IoT systems.

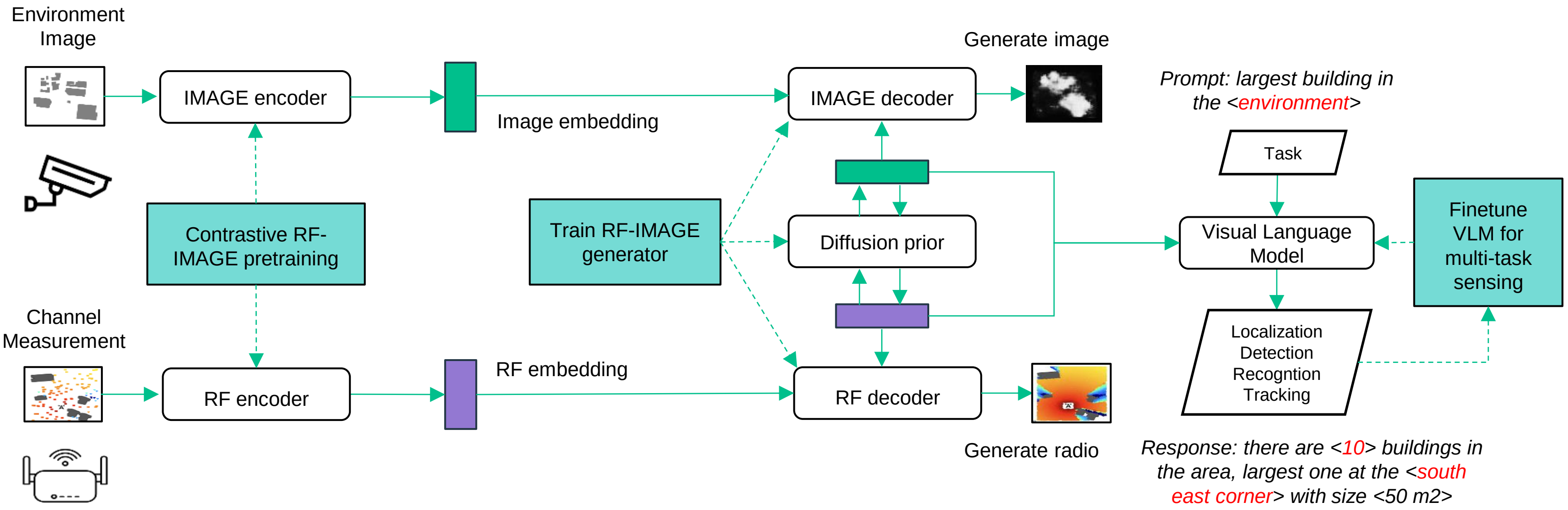
GenAI refer to neural network models designed to generate new data similar to a given dataset, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Autoregressive Models, flow-Based Models, Diffusion Models (DMs), and Transformer-based Large Language Models (LLMs). These techniques offer significant advantages in data-intensive applications by creating high-quality synthetic data, improving data quality through denoising, and filling in missing values. Generative AI is particularly effective in both cross-modal and multi-modal applications: integrating diverse data types into unified representations for better decision-making and translating information between modalities to enhance robustness. This capability supports innovative IoT applications, smart cities, healthcare, and autonomous systems, showcasing generative AI's transformative potential.

GenAI's ability to enhance data quality and integrate various data types makes it ideal for IoT applications, which require universality. With the advent of smarter devices, advanced sensors, and enhanced connectivity technologies like 5G and 6G, IoT systems can greatly benefit from GenAI. It extends conventional deep learning to manage diverse and unforeseen tasks with limited data and resources. GenAI's generalization capability is crucial for IoT devices to adapt to new environments and tasks. Additionally, GenAI's natural language processing enhances multi-modal sensing by integrating text, audio, and visual data, creating more comprehensive and intelligent IoT systems.

As shown in Fig. 1, we discuss the main challenges in RF sensing applications and explore how GenAI can address these issues using unimodal and multi-modal datasets, including reviewing the most relevant works and proposing feasible solutions for the potential use of GenAI. The main contributions of this article can be summarized as follows:

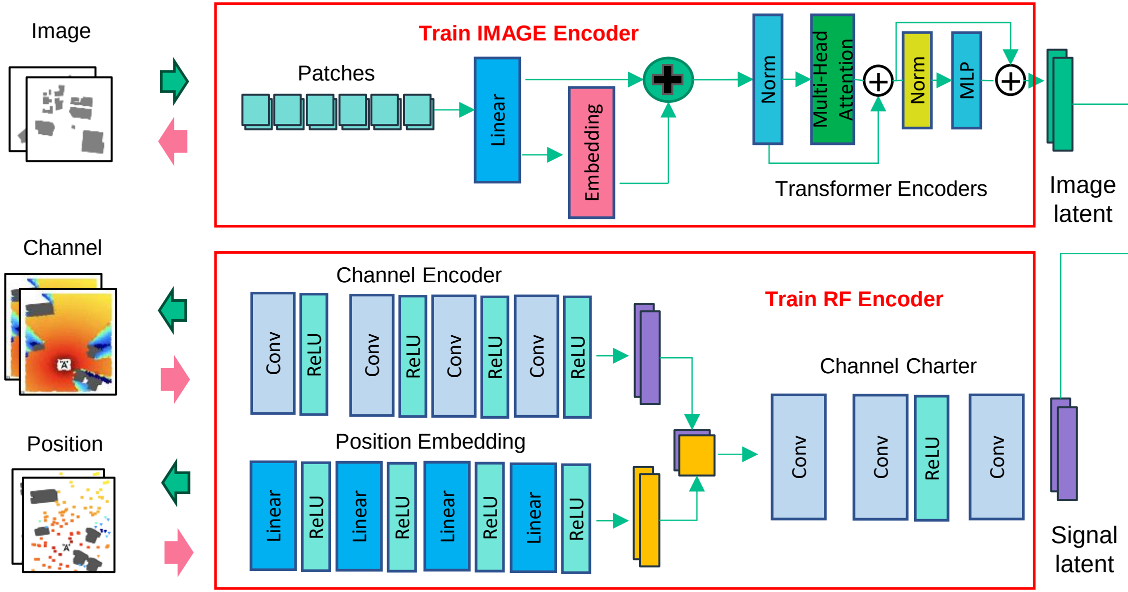
Multi-Task Wireless Sensing

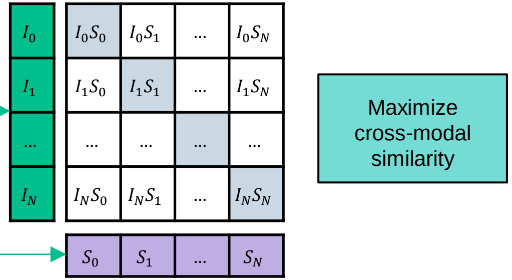
- RF-Visual-Language model generalize different wireless sensing tasks with prompts
  - Contrastive cross modality pretraining to connect RF and image on a common latent space
  - Cross-attention embeds RF, image, text with LLM to generate objects, location from prompt



Multi-Task Wireless Sensing

- Contrastive cross-modality pretraining
  - Maximize latent similarity of RF and image pairs sampled in the same environment
  - Capable of zero-shot classification on larger datasets without specific fine-tuning

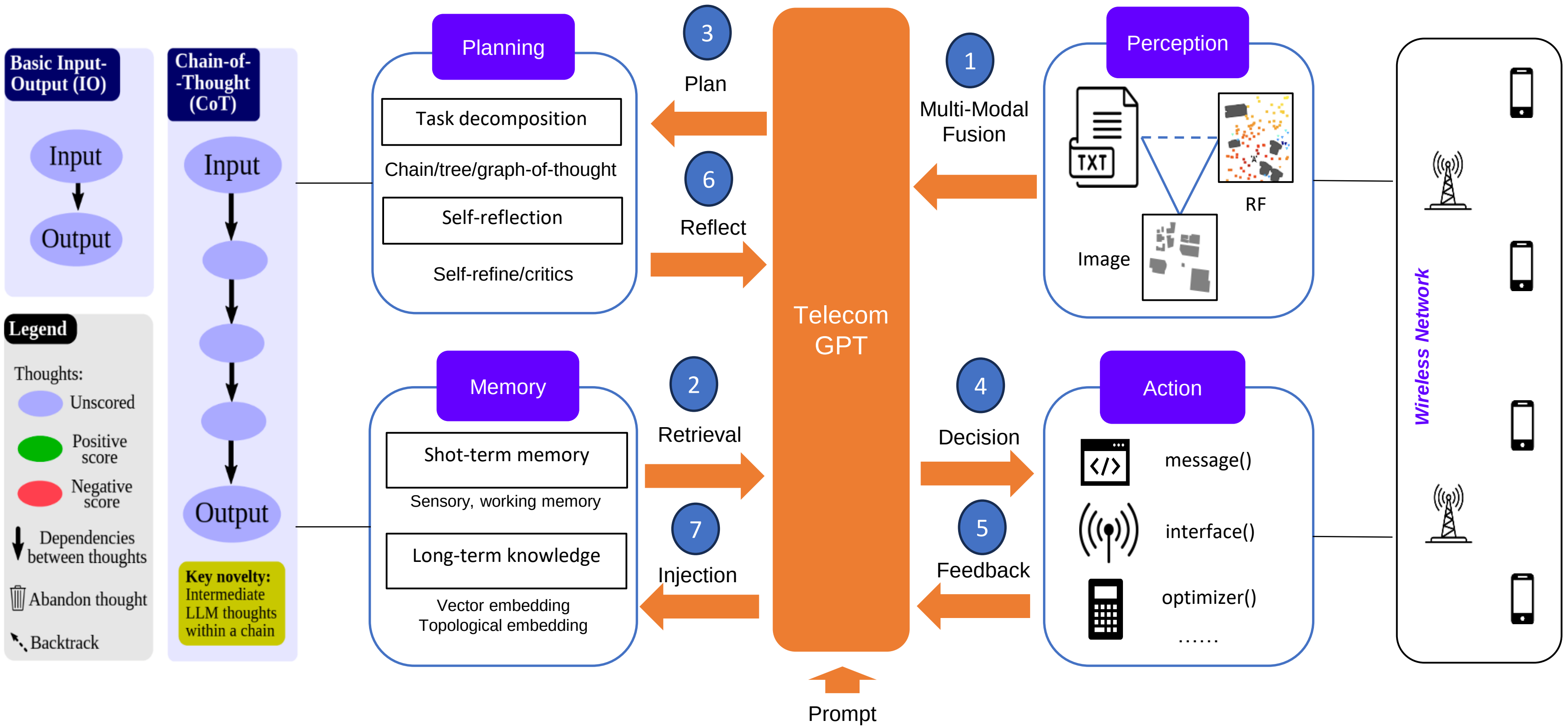




Zero-Shot Latent Space Classification

Pairs	100	1000	10000
Signal-to-Image	90.6	88.1	86.9
Image-to-Signal	88.3	88.5	86.1

TelecomGPT 3.0: Telecom AI Agent Solver



Diffusion Models as Network Optimizers: Explorations and Analysis

Ruihuai Liang, *Student Member, IEEE,* Bo Yang, *Member, IEEE,* Pengyu Chen, Xianjin Li, Yifan Xue, Zhiwen Yu, *Senior Member, IEEE,* Xuelin Cao, *Member, IEEE,* Yan Zhang, *Fellow, IEEE*, Mérouane Debbah, *Fellow, IEEE*, H. Vincent Poor, *Fellow, IEEE*, and Chau Yuen, *Fellow, IEEE*

***Abstract*—Network optimization is a fundamental challenge in the Internet of Things (IoT) network, often characterized by complex features that make it difficult to solve these problems. Recently, generative diffusion models (GDMs) have emerged as a promising new approach to network optimization, with the potential to directly address these optimization problems. However, the application of GDMs in this field is still in its early stages, and there is a noticeable lack of theoretical research and empirical findings. In this study, we first explore the intrinsic characteristics of generative models. Next, we provide a concise theoretical proof and intuitive demonstration of the advantages of generative models over discriminative models in network optimization. Based on this exploration, we implement GDMs as optimizers aimed at learning high-quality solution distributions for given inputs, sampling from these distributions during inference to approximate or achieve optimal solutions. Specifically, we utilize denoising diffusion probabilistic models (DDPMs) and employ a classifier-free guidance mechanism to manage conditional guidance based on input parameters. We conduct extensive experiments across three challenging network optimization problems. By investigating various model configurations and the principles of GDMs as optimizers, we demonstrate the ability to overcome prediction errors and validate the convergence of generated solutions to optimal solutions.**

***Index Terms*—Internet of things, network optimization, diffusion models, generative artificial intelligence.**

and energy consumption [7], [8], reducing execution costs [9]–[11], or enhancing quality of service (QoS) [12], [13]. The complexity of real-world environments and application requirements introduces challenging mathematical properties to these problems, such as convexity and non-convexity, linearity and non-linearity, and solution spaces that may be discrete, continuous, or mixed-integer. Additionally, these problems can involve both differentiable and non-differentiable components. These properties impact not only the objective functions but also the constraints, resulting in a complex feasible solution space that makes solving such problems particularly challenging.

Existing network optimization methods primarily include numerical approaches based on optimization theory [5], [8], [10], [12]–[14] and fitting algorithms based on machine learning [4], [9], [11], [15]–[22], with some work exploring the use of deep learning to enhance numerical methods [23]–[25]. For problems without complex characteristics, it is often straightforward to apply classical algorithms. However, when dealing with more typical challenges—such as non-convex, mixed-integer, multi-objective, or Pareto optimization—custom algorithm design becomes necessary [5], [8], [10], [12]–[14]. This process demands a strong understanding of optimization theory and can be labor-intensive in developing effective solutions. In the case of fitting algorithms, standard supervised

Large Language Model Based Multi-Objective Optimization for Integrated Sensing and Communications in UAV Networks

Haoyun Li, Ming Xiao, Kezhi Wang, Dong In Kim, and Merouane Debbah

***Abstract*—This letter investigates an unmanned aerial vehicle (UAV) network with integrated sensing and communication (ISAC) systems, where multiple UAVs simultaneously sense the locations of ground users and provide communication services with radars. To find the trade-off between communication and sensing (C&S) in the system, we formulate a multi-objective optimization problem (MOP) to maximize the total network utility and the localization Cramér-Rao bounds (CRB) of ground users, which jointly optimizes the deployment and power control of UAVs. Inspired by the huge potential of large language models (LLM) for prediction and inference, we propose an LLM-enabled decomposition-based multi-objective evolutionary algorithm (LEDMA) for solving the highly non-convex MOP. We first adopt a decomposition-based scheme to decompose the MOP into a series of optimization sub-problems. We second integrate LLMs as black-box search operators with MOP-specifically designed prompt engineering into the framework of MOEA to solve optimization sub-problems simultaneously. Numerical results demonstrate that the proposed LEDMA can find the clear trade-off between C&S and outperforms baseline MOEAs in terms of obtained Pareto fronts and convergence.**

***Index Terms*—Integrated sensing and communications, unmanned aerial vehicle, multi-objective optimization, large language model.**

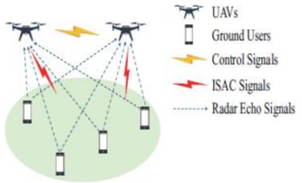
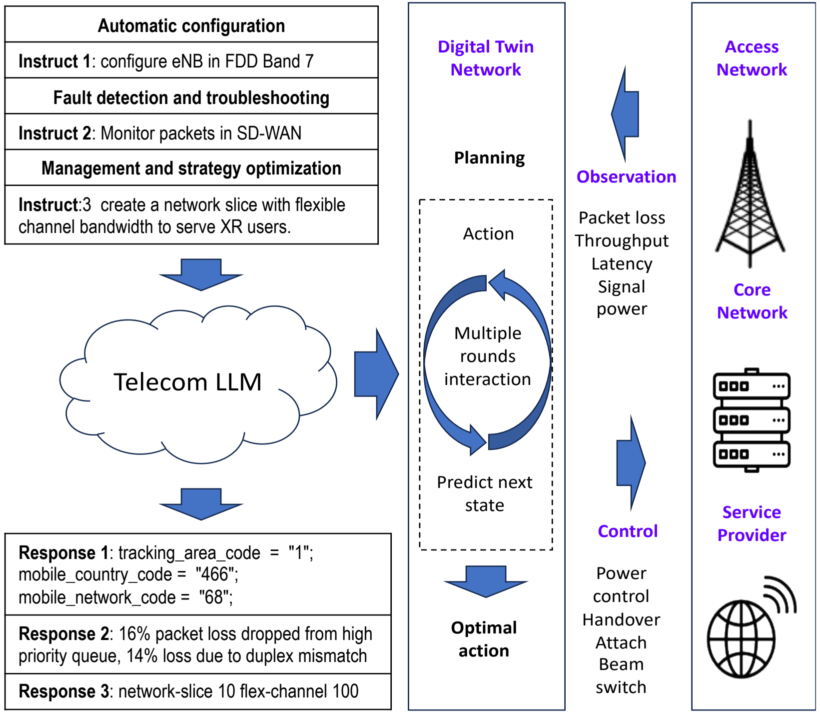


Fig. 1. Illustration of the multi-UAV-assisted ISAC system.

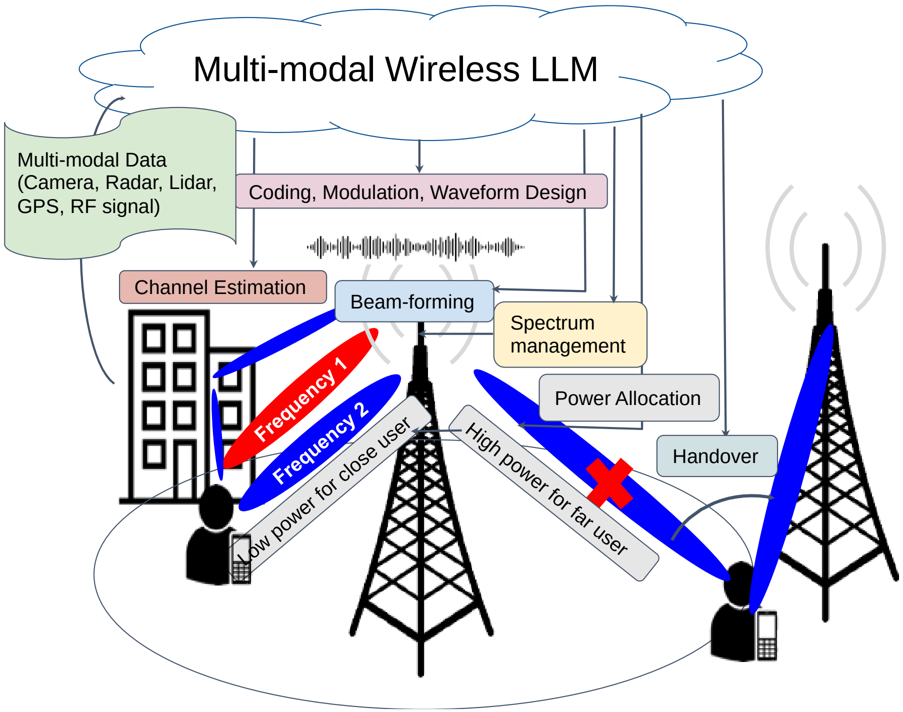
the non-negligible trade-off in the UAV networks, a multi-objective optimization problem (MOP) could be applied. To solve the resulting MOP, the bio-inspired multi-objective evolutionary algorithm (MOEA) is considered a promising approach to simultaneously dealing with a set of solutions and finding several Pareto-optimal solutions even for the non-convex Pareto front in a single run of the algorithm [4]. However, the complexity of MOEA is still high in general. Recently, large language models (LLM) have demonstrated remarkable capabilities in reasoning and prediction, which

TelecomGPT 3.0

Autonomous Network Management



Radio Network Optimization



TelecomGPT 3.0: Telecom AI Agent Solver

LLM

LLM

LLM

......

1. Retrieve from BS memory*: network status and goals*
Consider the downlink channel with one base station and 4 users. The noise level is the same for all users and equals to 1.258 W. The bandwidth is 15 kHz. The channel gain vector is [1.21, 2.01, 0.58, 0.13]. The initial transmission power vector is [2.0, 4.0, 5.0, 6.0] W. Base station aims at reducing the total transmission power by 0.85 W at least..

2. Reasoning & multi-agent prompting: BS to UE
After calculation of transmission rate for each user in the previous round, it shows that User 1 and User 4 need to reduce their transmitting power to help the base station reduce the total transmission power by 0.85 W at least.

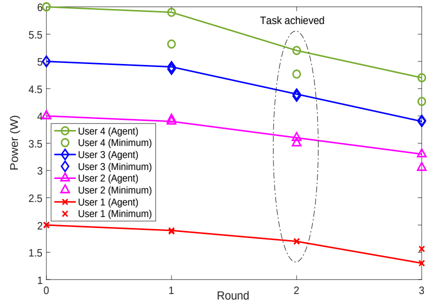
3 retrieve from UE memory: previous decisions and reasoning
{Round 1: 4.5W, 15.8 kbps, Round 2: 4.5 W, 13.5 kbps}

4. Reasoning & decision making: call functions
I am User 2, I will ensure that my transmission rate is no less than my minimum rate of 15.8 Kbit/s. I will adjust my transmitting power to 4.2 W. I will help the base station to reduce the total transmission power and achieve the target of at least 0.85 W reduction.

5. Update UE memory

6. Prompt decision to BS
My transmitting power is 4.8 W.

7. Update BS memory



- LLM powered AI agents on BSs and UEs
- Non-cooperative games on power saving
- Agent communciates and memorizes strategies

Thank you