جامعة خليفة
Khalifa University

KU 6G RESEARCH CENTER

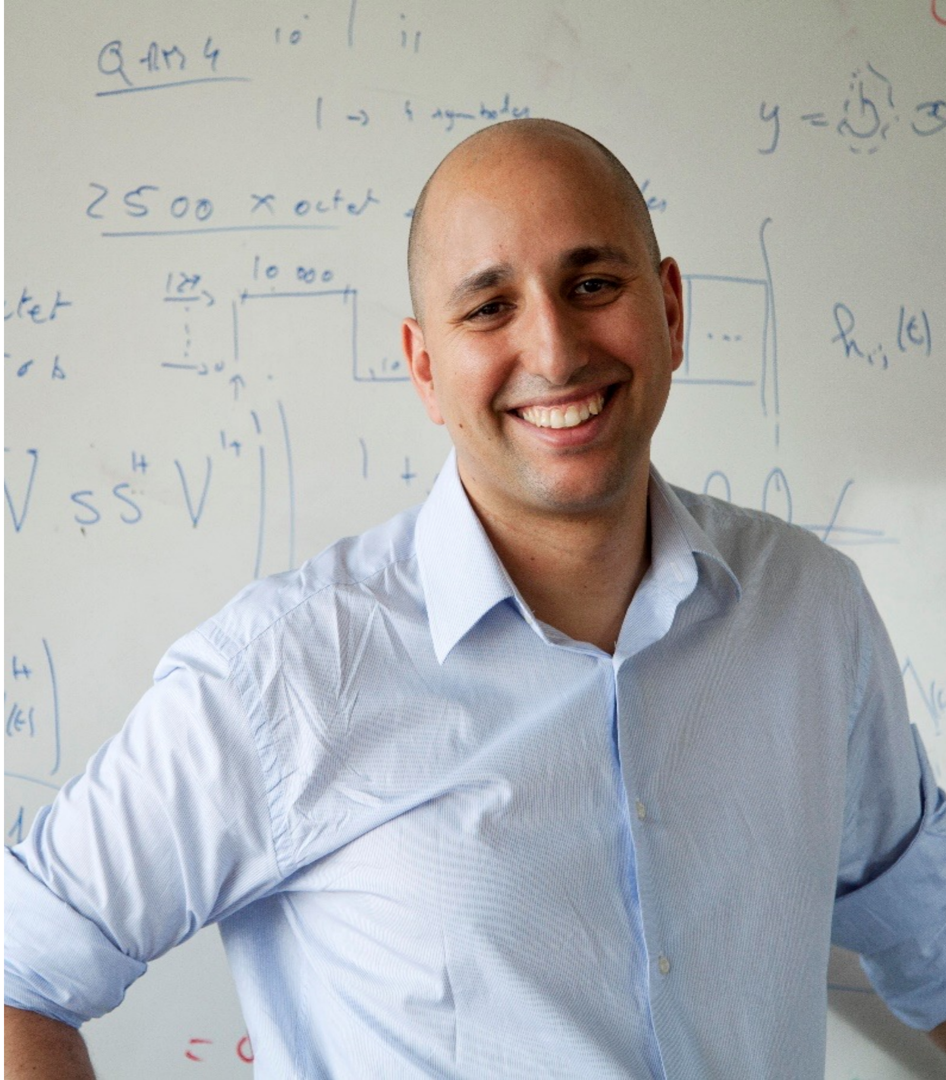
TelecomGPT

En Route to Building Telecom-Specific LLMs

Prof. Merouane Debbah
Director, 6G Research Center
Khalifa University

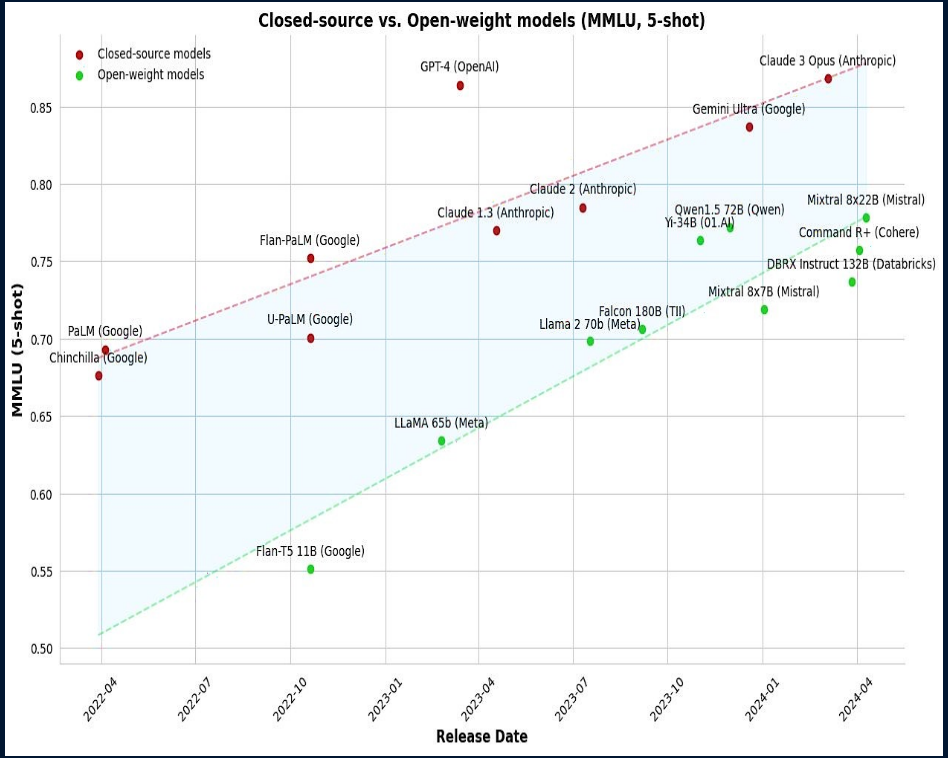
ABOUT ME

- Professor & Founding Director, Khalifa University 6G Center
- Senior AI Advisor at TII
- IEEE, EURASIP and WWRF Fellow
- Citations: 70000+, h-index:115+
- More than 40 Best papers Awards
- More than 50 patents
- IEEE Signal Processing Society Distinguished Industry Speaker (2021-2022)
- Field of Research: AI and beyond 5G Systems



2023: THE GENERATIVE AI RACE

LLM MODELS APPEARING AT THE UNPRECEDENT PACE



Falcon 180B: The Largest Open Language Model Surpasses Llama 2 and GPT 3.5

6 September 2023 State-of-the-Art

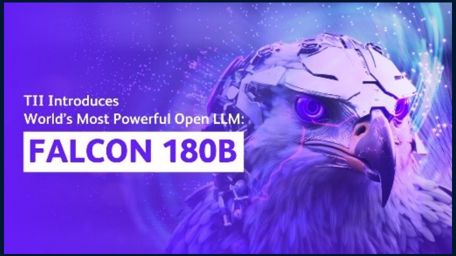
The Institute of Technological Innovations from the UAE has unveiled Falcon 180B, the largest open language model, displacing Llama 2 from the top spot in the rankings of pre-trained open-access language models by HuggingFace. The model was trained on 3.5 trillion tokens using the

The Falcon Series of Open Language Models: 7B, 40B, 180B ..

The World's Most Powerful Open Access Large Language Model

- Falcon LLM is currently at the ***top of the Hugging Face Leaderboard for pre-trained Open Large Language Models.***
- Falcon-180B, has been trained on over 3.5 trillion tokens of text–***the largest openly documented pretraining run***.
- **Falcon-180B** significantly outperforms models such as *PaLM* or *Chinchilla*, and improves upon concurrently developed models such as *LLaMA 2* or *Inflection-1*.
- The production training for Falcon 180B was run on 4096 GPUs using the Amazon SageMaker cloud machine learning platform for a total of about 7,000,000 GPU hours.

	Falcon-7B	Falcon-40B	Falcon-180B
Pretraining [tokens]	1,500B	1,000B	3,500B
Compute [PF-days]	730	2,800	43,500
Training [A100s]	384	384	4,096
Availability	Apache 2.0	Apache 2.0	Responsible use license
Agg. performance	60.8	67.1	70.3
Closest model	<GPT-3	Chinchilla	PaLM-2 Large



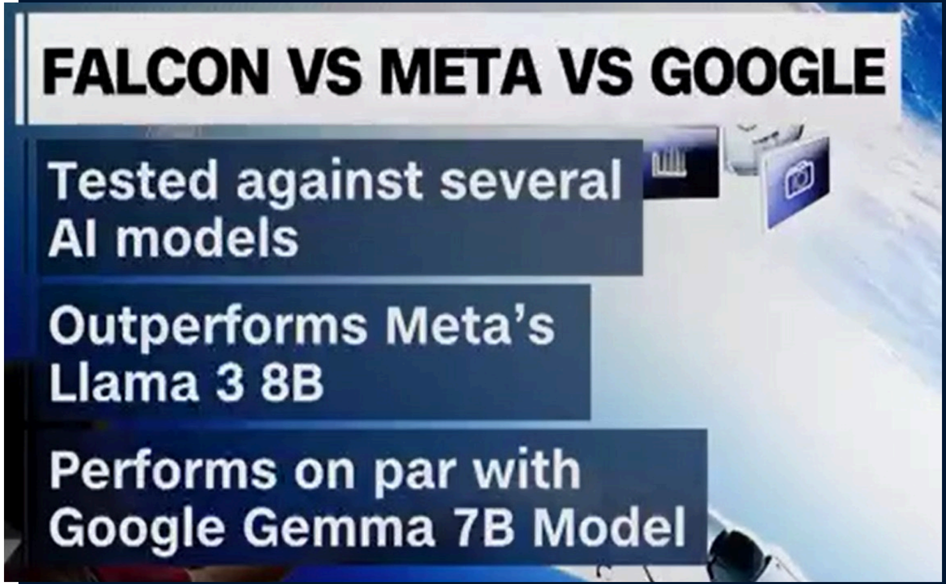
FALCON 2: 13th May 2024

TOP RANKED OPEN-SOURCE MODEL

Falcon 2: UAE's Technology Innovation Institute Releases New AI Model Series, Outperforming Meta's New Llama 3

- *Next-Gen Falcon 2 Series launches AI Model that is Open-Source, Multilingual, and Multimodal – and is only AI Model with Vision-to-Language Capabilities*
- *New Falcon 2 11B Outperforms Meta's Llama 3 8B, and Performs on par with leading Google Gemma 7B Model, as Independently Verified by Hugging Face Leaderboard*
- *Immediate Plans Include Exploring 'Mixture of Experts' for Enhanced Machine Learning Capabilities*

Abu Dhabi-UAE: 13 May, 2024 - The Technology Innovation Institute (TII), a leading global



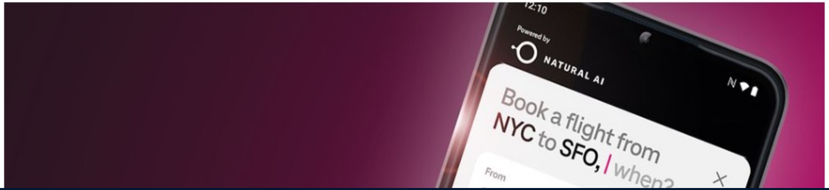
PORTABLE LLM AND AI AGENTS: END OF APPS...

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

2023: THE 6G RACE

2023: The 6G 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

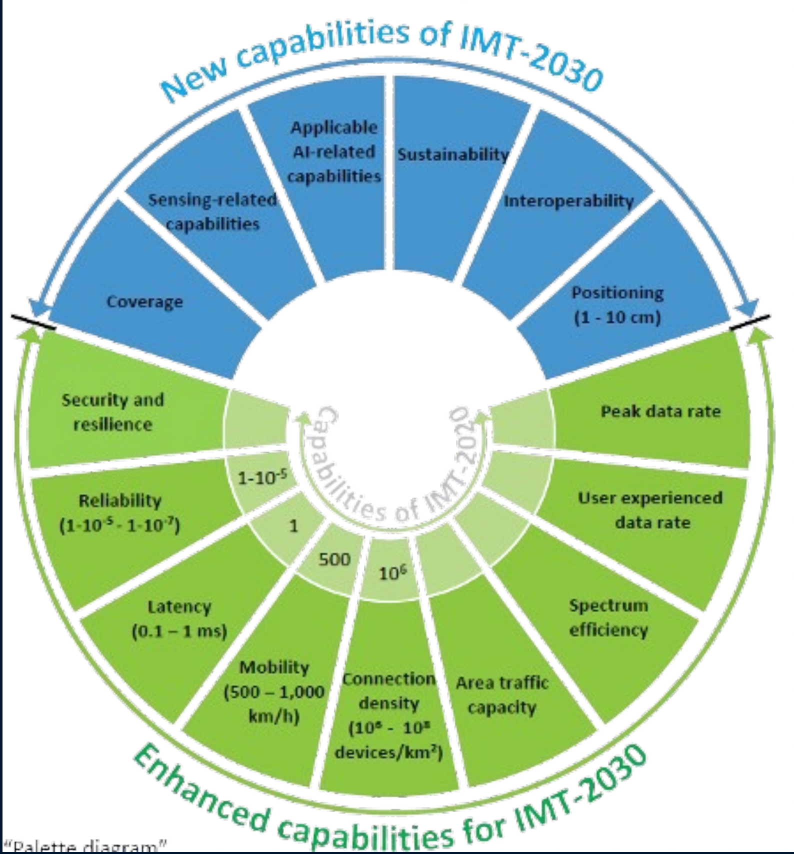
21 Nov 2023

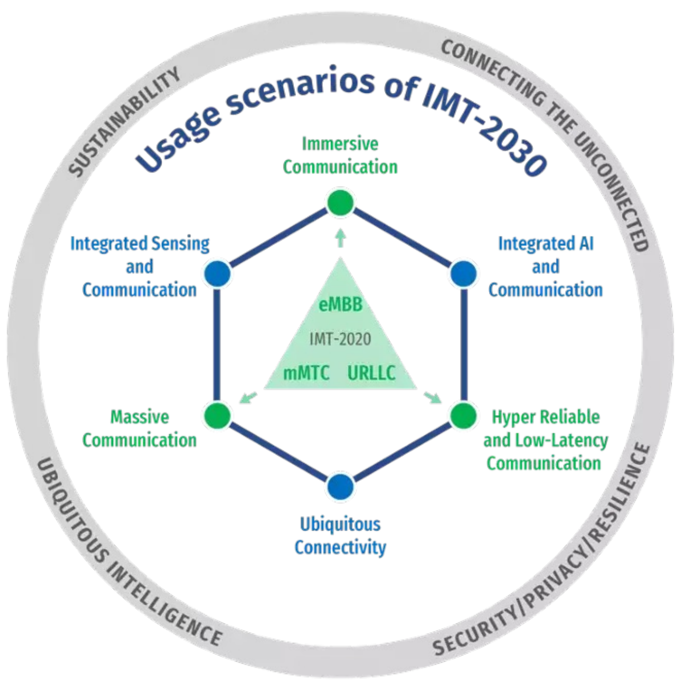
ITU adopts resolution to guide development of the 6G standard

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

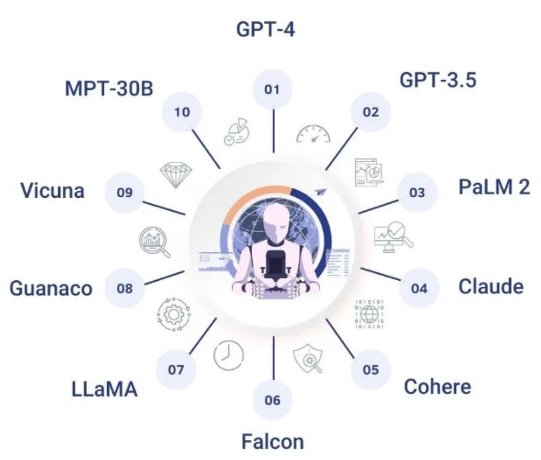




LLM TELECOM PILLARS

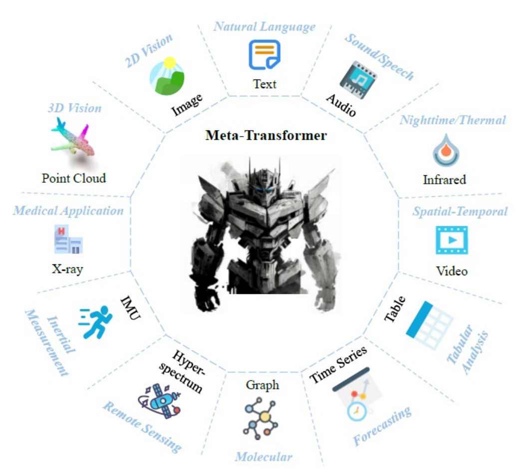
TELECOM FOUNDATION MODELS

Large Model at Small Devices



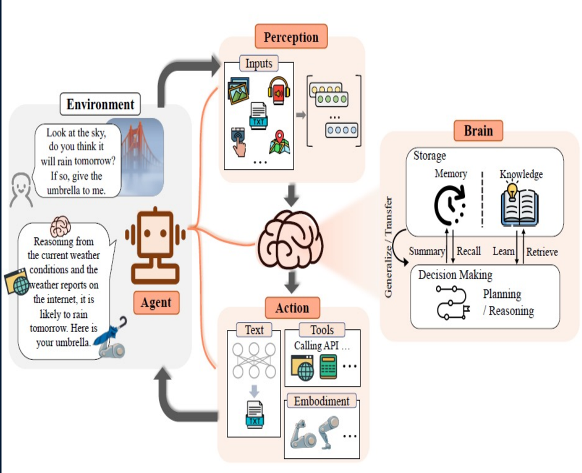
TELECOM BIG DATA

Multi-Modality



TELECOM AUTONOMY

Autonomous Agents



DOMAIN-SPECIFIC LLMs

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.

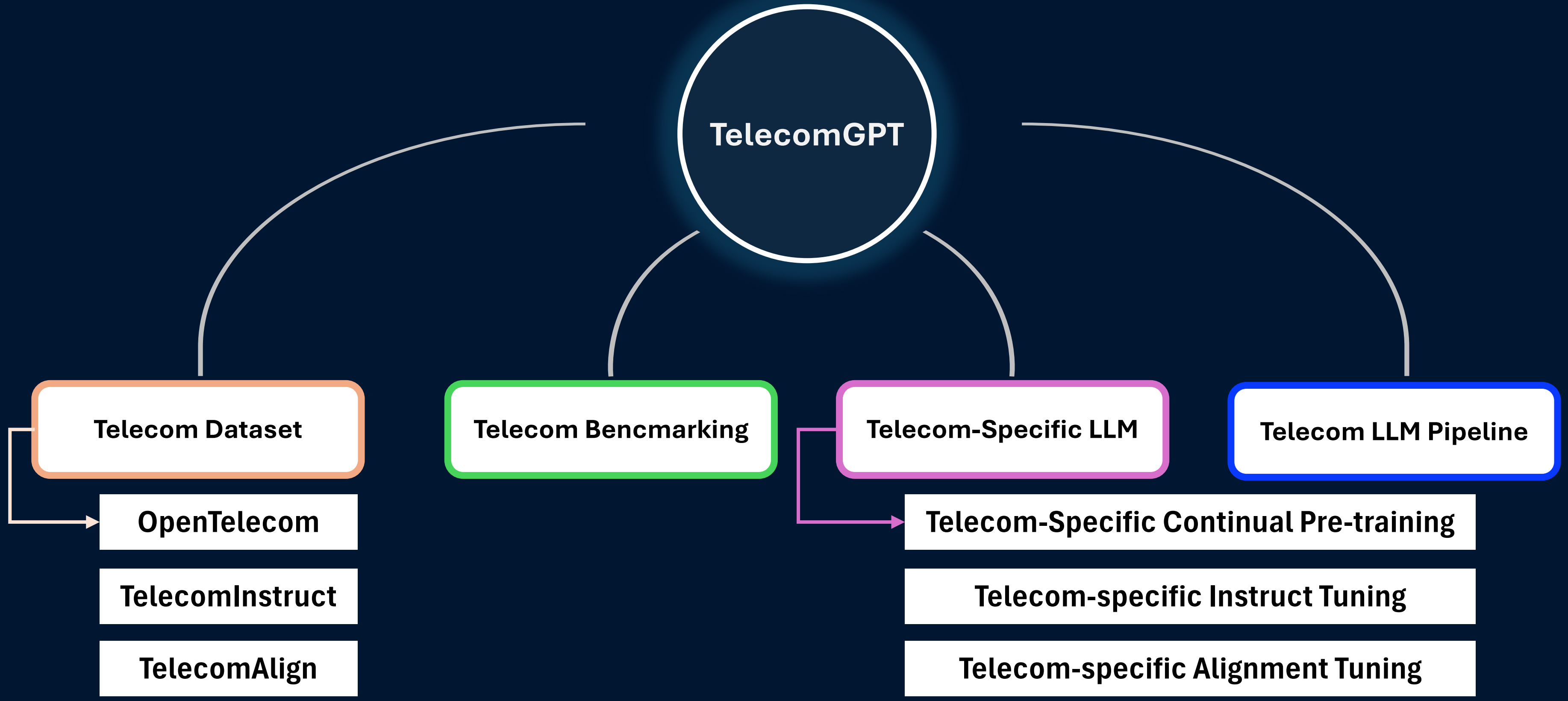
LLM-QAT [8], one bit quantization in BitNet [9] makes it possible to deploy mainstream LLMs on edge devices, e.g., mobile phones. Second, the long inference time of LLMs is unbearable to meet the requirement of Ultra Reliable and Low Latency Communication (URLLC) in beyond 5G networks. Taking the example 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

TelecomGPT

To a new era of Telecom Agents

TelecomGPT FRAMEWORK

3 Datasets, 3 Models, 5 Benchmarks!



TelecomGPT DATASET

Khalifa University

For Continual pre-training

OpenTelecom

3GPP · GitHub · arXiv · IEEE · Wikipedia · StackExchange + Books, Patents (C4)

- Release 8 - 19
  - RAN
  - SA
  - CT
- IEEE 802.11
- IEEE 802.3
- IEEE 802.15
- C95.1 series

Keywords associate with Telecom industry & Theory (e.g., MIMO, 5G, etc)

Keywords frequently appearing in Telecom publications (e.g., broadband, LTE, roaming)

Keywords uniquely identifying Telecom technologies (e.g., spectrum, fiber-optic, data trans.)

- Standards (IEEE, ITU, etc)
- Telecom Advances (quantum, 5GNR)
- Avoid broad terms (convex, efficiency)

TelecomGPT DATASET

Khalifa University

TelecomInstruct

For instruction tuning

Multiple-Choice QnA	Open-ended Question Answering	Technical Doc Classification
Select all that apply	Answer Telecom-specific questions from standards, research papers, patents, in an open-ended manner	Classify text from different Tdocs into related working group of the relevant SDOs.

Math Modeling	Code Generation	Code Infilling
Generate accurate math equations such as channel models for a given description of system model	Generate script for a given functionality in telecom (sending signal indicator, extract MAC address from a frame,..)	Infill incomplete script based on the con- text and the targeted functionality

Code Summary	General Instruction	Protocol Instruction
Summarize the core functionality of a given script, including identifying if the script is telecom- relevant or not.	Explain concepts, specifications, identify problems, propose solutions, in Telecom standards, patents, papers.	Generate the protocol workflows in Telecom standard following a human prompt.

TelecomGPT DATASET

TelecomInstruct

Prompt Template for Telecom Question Answering

Instruction: Please create several multiple-choice questions based on the provided texts. These created questions must be generated in this form:

- Question: XXXX
- Option 1: XXXX
- Option 2: XXXX
- ...
- Answer: Option X
- Explanation: explain why the correct answer is Option X.

These questions should not refer to any equation. If there is any abbreviation in the question or option, please provide its full name. The "Answer" must be in the format of "Option X". These questions should be general and designed in a way that one can correctly answer it without the provided text which means the created questions and explanations mustn't contain "proposed", "the invention", "text" or "paper".

TEXT: {text}

Prompt Template for Telecom Code Generation

Instruction: As a distinguished expert in telecommunications software, you are skilled in developing and optimizing programs within the telecom domain, using languages like C++, C, Python, and Matlab. When provided with a script or function, your task is to craft a concise yet comprehensive request that can guide the generation of the script, if it is telecom-relevant. Refer to the examples provided below and tailor your request accordingly:

- **Example 1:** Write a Python function to convert an IPv6 address from string format to an integer.
- **Example 2:** Develop a C function that updates the decrypt status flag based on the decryption result for a received 802.11 frame.

Your request should satisfy the following requirements:

- Clearly identify the programming language of the script;
- Describe the script's functionality in telecom domain.
- If necessary, describe the scenario or environment where the script operates.

Should the script is telecom-relevant, please respond with "response": "content of your request". If the script lacks strong relevance to telecommunications or seems to be an incomplete function, your response should be "response": "irrelevant". Please ensure your request directly pertains to the script's utility in the telecom sector, omitting any unnecessary details or commentary.

SCRIPT: {script}

Prompt Template for Telecom Instruction Following

Instruction: You are specialized in Telecommunication domain. You are familiar with topics like 5G, RAN, wireless communication, etc., as well as technical Telecom standards, specifications from 3GPP. You are given a text in a 3GPP document. Your task is to transform it into an instruction as follows: {"instruction": "...", "output": "..."}. Instruction types could be planning subtasks according to the 3GPP procedure to complete the main tasks or achieve a goal. The instruction/output must be clear. They must contain relevant context/passages in the target text needed to interpret them. Ensure that they do not mention implicit information, such as figures, tables, annexes, other sections in the document, etc.

Example: {"instruction": "Initiate event based charging with decentralized and centralized unit determination, centralized rating", "output": "1. Request for resource usage. 2. Units Determination. 3. Charging Data Request. 4. Account, Rating Control. 5. Create CDR. 6. Charging Data Response. 7. Granted Units Supervision. 8. Content/Service Delivery."}

TEXT: {text}

TelecomGPT DATASET

Khalifa University

TelecomAlign

For alignment tuning

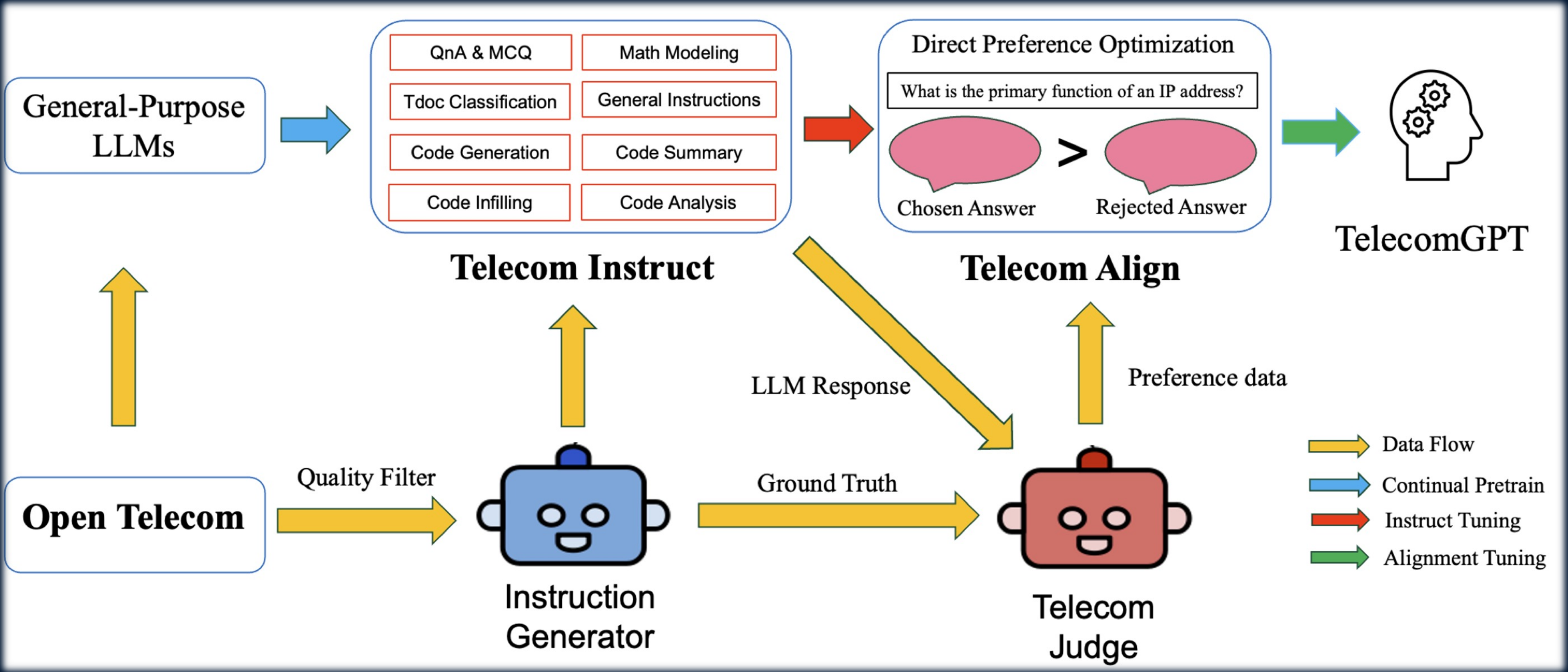
- Supervised fine-tuning models have generally learned to perform telecom-relevant tasks, **BUT**:
  - Repeated generation
  - Too short responses
  - Telecom-irrelevant content generation

- Rather than collecting real human preference data which is costly and inefficient, we define our preference:
  - provide concise and accurate answer
  - minimum amount of information unless requested otherwise

✓ Reduces latency in LLM-based communication systems
✓ Aligns with concepts like semantic communication

Preference Dataset Creation: Obtained by selecting instructions with low performance metrics

TelecomGPT PIPELINE



TelecomGPT

Ground Truth vs. Instruct vs. Align

Comparison of generated response for instruction task in Telecom research paper.

Our model provides concise and high-quality summaries of Telecom techniques compared to the general-purpose LLM

Prompt	Summarize the challenges and solutions in the field of CSI feedback in massive MIMO systems utilizing DL-based methods.
Category	Research Paper
Ground Truth	The challenges in CSI feedback in massive MIMO systems include the vast CSI matrix, compressed sensing inefficiency, and lack of standard CSI dataset collection. DL-based solutions involve using end-to-end Auto-Encoder architecture for compression and reconstruction, as well as considering factors like sampling range and network complexity in practical deployment.
Llama3-8B-Instruct	A very specific and technical question! CSI (Channel State Information) feedback is a crucial aspect in massive MIMO (Multiple-Input Multiple-Output) systems, as it enables the base station to adapt to the changing wireless channel conditions. However, in massive MIMO systems, the number of antennas and users increases, making CSI feedback a challenging task. Deep Learning (DL) based methods have been proposed to address these challenges. Here's a summary: **Challenges:**1. **Feedback overhead**: With a large number of antennas and users, the amount of CSI feedback data increases exponentially, leading to significant overhead and potential interference. 2. **Feedback delay**: The delay in receiving CSI feedback can result in outdated channel information, affecting the performance of the system. 3. **Channel estimation errors**: The accuracy of channel estimation is crucial, but errors can occur due to limited feedback resources and noisy measurements. **Solutions using DL-based methods:**1. **Compressive sensing**: DL-based compressive sensing methods can reduce the feedback overhead by compressing the CSI data while preserving its essential information. 2. **Feedback reduction**: DL-based methods can learn to reduce the feedback data by identifying the most important CSI information and discarding the rest. 3. **Feedback prediction**: DL-based feedback prediction methods can predict the CSI feedback based on historical data, reducing the feedback delay and overhead. 4. **Channel estimation**: DL-based channel estimation methods can improve the accuracy of channel estimation by learning from the available CSI feedback data and other system information. 5. **Hybrid beamforming**: DL-based hybrid beamforming methods can optimize the beamforming weights by learning from the CSI feedback data and other system information, improving the system performance. Some popular DL-based methods for CSI feedback in massive MIMO systems include: 1. Convolutional Neural Networks (CNNs) 2. Recurrent Neural Networks (RNNs) 3. Long Short-Term Memory (LSTM) networks 4. Autoencoders 5. Generative Adversarial Networks (GANs) These DL-based methods have shown promising results in reducing the feedback overhead, delay, and errors, while improving the system performance in massive MIMO systems.
Llama3-8B-TI	1. Challenges in CSI feedback for massive MIMO systems include the need for accurate channel estimation, efficient compression methods, and the ability to handle distributed data. 2. Solutions involve DL-based methods for channel estimation, such as CNN and DNN approaches, as well as DL-based feedback strategies for efficient compression and handling of distributed data.
Llama3-8B-TI-TA	Challenges in CSI feedback for massive MIMO systems using DL-based methods include the need for efficient compression strategies and robustness against channel variations. Solutions involve developing lightweight DL architectures for channel estimation and feedback, as well as exploring quantization techniques to reduce the overhead of transmitting CSI to the baseband unit.

TelecomGPT Benchmarking
Five Evaluation Benchmark Datasets!

- Telecom Question Answering
  - MCQ
  - Open-ended Questions
- Telecom Standards Documents Classification
- Telecom Math Modeling
- Telecom Code Understanding & Generation
- Telecom Instruct Following
  - General instruct
  - Protocol instruct

TelecomGPT Performance
Multiple-Choice Questions

LLMs	Lexicon	Research Overview	Research Publications	Standards Overview	Standards Specifications	Overall
GPT-4o	92	81.73	79.54	83.87	62.77	78
GPT-4	92	77	78	79	60	75
GPT-3.5	96	66.35	66.98	64.52	56.38	66
Llama3-8B	72	51.92	65.11	56.45	36.17	56.20
Llama3-8B-Instruct	80	67.31	69.77	59.68	50	64.80
Llama3-8B-TI	**96**	**69.23**	**74.88**	**74.19**	**56.38**	**71.20**
Llama3-8B-TI-TA	92	73.08	71.63	72.58	58.51	70.60
Mistral-7B	72	49.04	51.16	50	34.04	48.40
Mistral-7B-Instruct	84	64	65	56	51	62
Mistral-7B-TI	**84**	**67.3**	**70.69**	**56.45**	**51.06**	**65.2**
Mistral-7B-TI-TA	84	70.19	73.95	61.29	48.94	64
LlaMA-2-7B	62.5	52.24	49.18	48.28	40	48.94
LlaMA-2-7B-TI	**84**	**57.69**	**63.26**	**56.45**	**50**	**59.80**
LlaMA-2-7B-TP-TI	**81.82**	**63.92**	**67**	**70**	**47.48**	**63.79**

LLMs	Paper	Book	Patent	Wiki	Overall
GPT-4o	98	94	84	98	93
GPT-3.5	95	89	83	91	89
LlaMA3-8B-Instruct	97.92	89.81	80.65	87.21	88.49
LlaMA3-8B-TI	**92.52**	**97.17**	**92.16**	**94.12**	**94**
Mistral-7B-Instruct	78.5	76.42	65.69	71.64	73.25
Mistral-7B-TI	**80.37**	**82.08**	**82.35**	**81.18**	**81.5**

TelecomGPT Performance

Tdoc Classification

LLMs	RAN	SA	CT	Overall
GPT-4o	44.12	47.59	17.28	38.94
GPT-3.5	42.93	48.59	16.26	38.54
LlaMA3-8B-Instruct	39.13	38.66	16.19	33.35
LlaMA3-8B-TI	**82.76**	**68.8**	**73.61**	**75.30**
Mistral-7B-Instruct	29.59	33.20	16.87	27.84
Mistral-7B-TI	**76.62**	**76.90**	**49.89**	**70.83**

Math Modeling

LLMs	Average Score	$\geq 90\%$	$\geq 50\%$
GPT-4	49.38	3.77	50.35
GPT-3.5	43.53	1.81	40.44
Llama3-8B-Instruct	40.78	2.51	34.45
Llama3-8B-TI	46.16	9.69	46.80
Llama3-8B-TI-TA	**49.45**	**9.52**	**50.73**
Mistral-7B-Instruct	35.54	1.53	29.43
Mistral-7B-TI	47.66	8.04	48.77
Mistral-7B-TI-TA	**48.11**	**7.22**	**49.26**

Thank you