Large Language Models for Telecom: Forthcoming Impact on the Industry

Ali Maatouk, Nicola Piovesan, Fadhel Ayed, Antonio De Domenico, and Merouane Debbah

Abstract

Large language models (LLMs) — AI-driven models that can achieve general-purpose language understanding and generation — have emerged as a transformative force, revolutionizing fields well beyond natural language processing (NLP) and garnering unprecedented attention. As LLM technology continues to progress, the telecom industry is facing the prospect of its impact on the landscape. To elucidate these implications, we delve into the inner workings of LLMs, providing insights into their current capabilities and limitations. We also examine the use cases that can be readily implemented in the telecom industry, streamlining tasks, such as anomaly resolution and technical specification comprehension, which currently hinder operational efficiency and demand significant manpower and expertise. Furthermore, we uncover essential research directions that deal with the distinctive challenges of utilizing the LLMs within the telecom domain. Addressing them represents a significant stride toward fully harnessing the potential of LLMs, and unlocking their capabilities to the fullest extent within the telecom domain.

Introduction

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP) and Artificial Intelligence (AI), propelling text generation, comprehension, and interaction to unprecedented levels of sophistication. The history of LLMs can be traced back to the early developments in Machine Learning (ML) and NLP, which encompassed the emergence of statistical language models and the advancements in neural networks. However, it was the rise of transformer architectures [1], which paved the way for the development of language models capable of processing and generating vast amounts of text. Among the notable advancements in this domain, OpenAI's Generative Pre-trained Transformer (GPT) series and open-source LLMs like LLaMA and its successor LLaMA2 have garnered significant attention [2]. Specifically, they have surpassed earlier models in terms of scale and capability, empowering human-like language understanding and generation.

Thanks to their language understanding capabilities, LLMs have the potential to revolutionize diverse domains [3], surpassing traditional NLP applications like machine translation and sentiment analysis. In fact, through domain-specific data, they can excel in tasks related to that particular domain. For instance, in medicine, LLMs may play a crucial role in encoding clinical knowledge and supporting medical decision-making processes. Similarly, researchers in finance have investigated how LLMs can provide insights into market trends and assist in risk analysis. Also, educational organizations have recently developed an LLM-based virtual tutor and classroom assistant.

Although LLMs have already demonstrated their potential in various fields, their application in the telecom industry has been relatively scarce. However, this situation is changing as more researchers are beginning to explore the capabilities of LLMs in this domain. For instance, a Bidirectional Encoder Representations from Transformers (BERT)-like language model was adapted to the telecom domain [4] to test its ability to answer a small, manually curated dataset of telecom questions. In another work, language models such as BERT and GPT-2 were leveraged to classify working groups within the Third Generation Partnership Project (3GPP) based on analysis of technical specifications [5]. Moreover, the potential of LLMs in facilitating Field-Programmable Gate Array development within wireless systems was highlighted in [6]. Additionally, the authors in [7] provided a vision where LLMs, along with multi-modal data (e.g., images), can significantly contribute to the development of Radio Access Network (RAN) technologies such as beamforming and localization. In this future, by combining different data types like text and visuals, LLMs can assist in optimizing and improving RAN functionalities.

In parallel to the work initiated by the research community, telecom ecosystem industries offer the first products based on LLM technologies. Huawei has released Pangu, an LLM that has been tested in mining, government, vehicles, weather, and R&D applications. Qualcomm has released an AI engine to support up to 10 billion parameters of generative AI models on mobile handsets, allowing AI assistant with NLP capabilities and image generations based on Stable Diffusion. Moreover, Google has introduced generative AI capabilities in its cloud platform to offer Mobile Network Operators (MNOs) the opportunity to integrate NLP functionalities in applications such as root cause analysis, information retrieval in legal documents, and conversational chatbot for customer experience improvement.

Ali Maatouk, Nicola Piovesan, Fadhel Ayed, and Antonio De Domenico are with Huawei Technologies, France; Merouane Debbah is with Khalifa University of Science and Technology, UAE.

Digital Object Identifier: 10.1109/MCOM.001.2300473

0163-6804/25/$25.00 © 2025 IEEE

In light of these applications, a fundamental question arises regarding the immediate and future impact of LLMs on the telecom industry. In this article, we aim to answer this question by providing a view of LLMs and their impeding influence on the industry. Our objective is to demystify their current abilities, highlight their existing limitations, and showcase several use cases in the telecom industry where they can provide substantial assistance today. Additionally, we highlight the telecom data within the industry that can be harnessed to leverage the capabilities of LLMs. Moreover, we shed light on the technical difficulties that arise in implementing these use cases and outline the research directions that need to be pursued to fully harness the potential of LLMs.

Demystifying LLMs

To explore the potential of LLMs in the telecom industry, it is essential to begin by gaining an understanding of their intrinsic behavior. To do so, we delve into the intricacies of LLMs architecture and training, exploring their capabilities as well as their limitations.

Fundamentals of LLMs

LLMs are Deep Learning (DL) models with the ability to process information and demonstrate human-like text generation capabilities. Typically, LLMs utilize transformer-based architectures, where self-attention plays a pivotal role [1]. In self-attention, each word in an input sequence attends to all the other words, calculating attention scores that signify the importance of each word relative to the others. This mechanism allows to effectively capture long-range dependencies and grasp the contextual usage of each word. Interested readers can refer to the paper in [1] for a mathematical description of the selfattention mechanism. In Fig 1, a high-level illustration of an LLM is presented, along with the accompanying self-attention mechanism. Another essential component in the transformer architecture is multi-head attention, which expands upon the concept of self-attention. Often, a sequence element needs to attend to multiple distinct aspects, and relying on a single attention mechanism alone is inadequate to accomplish this objective. The multi-head attention provides the flexibility by enabling the model to attend to different aspects of the input, capturing diverse patterns and dependencies within the input sequence. This capability allows the model to learn complex interactions between words and comprehensively understand the input.

In addition, LLMs undergo extensive pretraining on vast amounts of text to acquire an understanding of the statistical properties inherent in the language at hand. During this phase, the models are mainly trained with data crawled from the internet, which provides them with diverse linguistic information. The primary goal of this pretraining is to enable the model to predict the next word in a sentence based on the preceding words. Through this process, the model captures both syntactic and semantic relationships, thereby enhancing its grasp of contextual nuances. Due to the range of corpora used during training and the large number of model parameters involved, LLMs can develop a comprehensive understanding of grammar, reasoning abilities, and even comprehend intricate language structures.

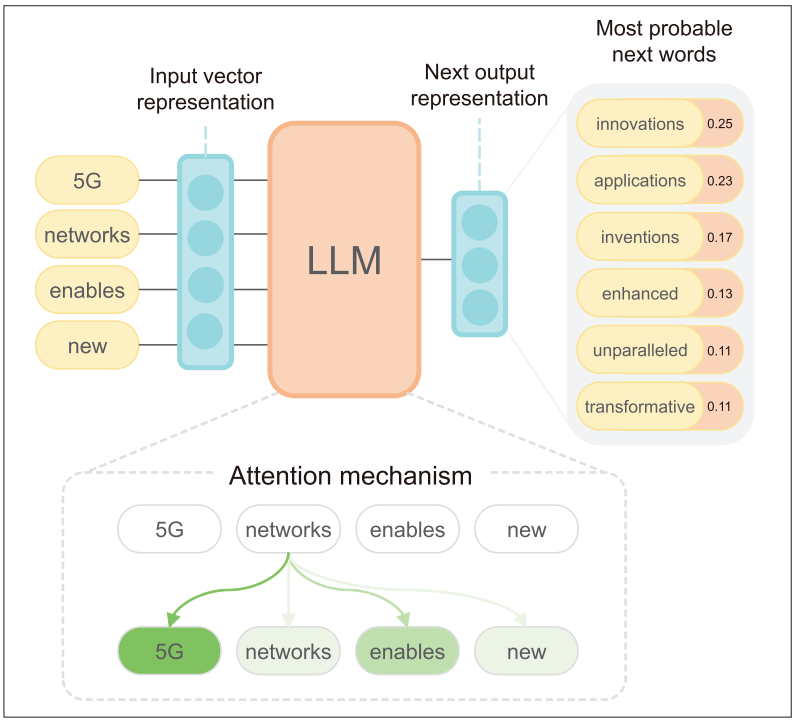

FIGURE 1. A high-level overview of LLMs

Although the pretrained LLM has a comprehensive understanding of the statistical properties within the language, it needs specific domain knowledge to be applied to industrial processes. To achieve this, the pretrained LLM's parameters, including attention blocks, are fine-tuned using domain-specific datasets and similar training techniques employed during the pretraining phase. Through this procedure, referred to as knowledge fine-tuning, the LLM can adapt the learned representations, denoted to as embeddings, from the pretraining phase to better align with the intricacies of the specific domain. In addition, researchers have designed prompt engineering solutions, such as chain-of-thought (CoT) prompting and Retrieval Augmented Generation (RAG), to enhance the capability of LLMs on a wide range of tasks. This topic and related open challenges are further discussed below.

LLMs Functionalities

The LLM potential shines through its three core competencies: an extensive understanding of the intricacies of language, cross-disciplinary knowledge, and the emerging ability to reason, albeit less developed than the former two. While we discuss three distinct functionalities: semantic comprehension, intelligent knowledge retrieval, and orchestration capabilities, it is important to note their inherent overlap in practical applications, as highlighted below.

Semantic Abilities: LLMs develop an internal representation of textual data in the form of real-valued vectors called embeddings. This representation conveniently encapsulates the input text's semantics, syntax, and contextual interpretation. These embeddings provide a simplified representation of textual data suitable for algorithmic procedures and data analysis. For example, a large city's telecom network generates millions of daily trouble tickets. Many disruptions are symptomatic of the

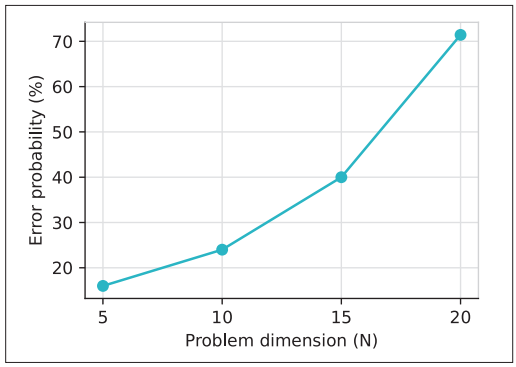

FIGURE 2. An illustration of an LLM output inconsistency. GPT-3.5 was provided a vector reporting the strengths of N beams. It was tasked with selecting the beam with the highest strength, while instructed to avoid a particular beam, which was always set as the strongest.

same core issues; however, due to compartmentalization within the network infrastructure, there is no automated system for categorizing these tickets. By converting them into embeddings from a domain-specific LLM, clustering algorithms like K-Means can effectively group the tickets, potentially tying them back to singular faults.

Intelligent Access to Knowledge: By understanding the specific intention conveyed through the prompt, an LLM can effectively apply its knowledge base to craft a response tailored to the user's needs. LLMs can process and comprehend intricate information, such as the content within standard documents, discern patterns, and infer logical conclusions from the given inputs. LLMs thus transition from passive language processors to active and intelligent agents, functioning as assistants or co-pilots that enhance professionals' productivity. For instance, in an operations and maintenance scenario, an operator faced with a trouble ticket may benefit from the model's ability to summarize the issue automatically, suggest possible solutions, and even draft a template email for field engineers to act upon, requiring only the operator's review and approval.

LLMs as Orchestrators: LLMs can utilize their reasoning to deconstruct complex tasks into manageable subtasks and deploy suitable (external) tools for each. They manage workflows by identifying the most appropriate tool for each segmented operation. Take, for instance, a task such as forecasting the next day's energy consumption for a Base Station (BS) undergoing hardware upgrades. Various tools are accessible, including data collection from available features and ML model training. The LLM can formulate a two-phase strategy: predict the traffic load and estimate the energy consumption for a specified load and hardware. It chooses the relevant ML model for each subtask and indicates the data needed to train it. After devising the strategy, the LLM orchestrates available tools to collect the relevant data and train the ML models.

LLMs Limitations

Given the structure and functionalities of LLMs, certain limitations become apparent. It is crucial to shed light on these shortcomings to utilize and interpret content generated by LLMs. The following are noteworthy flaws associated with them.

Hallucinations and Fabrications: One of the key concerns with LLMs is their tendency to generate hallucinations or fabrications. LLMs rely on statistical patterns and associations learned from vast text data during training. Consequently, they may produce responses that abide to these patterns, but are incorrect or nonexistent [8].

Limited Explainability: The complex architecture and massive number of parameters in these models render it difficult to trace the decision-making process. In fact, LLMs lack transparency in terms of the specific features or patterns they rely on to generate responses. This opacity hinders the ability to understand why a particular answer or response was chosen over others. This limited explainability raises concerns, especially in domains where transparency and accountability are crucial.

Computational Complexity: LLMs may consist of millions or even billions of parameters, making them resource-intensive to train and deploy. Even after training, running inference with LLMs can be computationally demanding. Generating responses with these models involves complex computations across multiple layers, which can strain available resources, especially for real-time applications.

Sensitivity to Updates: LLMs display sensitivity to adjustments in their parameters, leading to unforeseen variations in outputs and behaviors. A compelling illustration of this phenomenon can be found in [9], which showcased how the performance and behavior of both GPT-3.5 and GPT-4 underwent dramatic shifts over time: in March 2023, GPT-4 excelled at identifying prime numbers, but by June 2023, it faltered in handling the same questions. This inconsistency serves as a clear illustration of the susceptibility of LLMs to updates and alterations introduced to the model.

Output Inconsistency: This is a phenomenon that arises when the output generated by the model fails to fully align with the user's intent or the desired task, even when the prompt explicitly specifies the required output [10]. This is illustrated in Fig. 2, where GPT-3.5 was tested to answer a simple constrained maximization question. The LLM provided a wrong response with a given probability. Importantly, the error probability was observed to increase with the dimension of the problem. This can hamper the applicability of LLMs in areas such as telecom system optimization. Therefore, addressing this limitation becomes of utmost importance.

Potential LLM Applications in the Telecom Industry

With the understanding of how LLMs function, their capabilities, and their limitations, we can now delve into the applications that can have a large impact on the telecom industry.

Network Anomalies Resolution

Solving anomalies in the mobile network is a tedious task. With a vast infrastructure spanning across large geographical areas, maintaining and monitoring the BSs is challenging. Each BS is susceptible to a wide array of issues, including hardware malfunctions, software glitches, and environmental factors. For this reason, rectifying these anomalies necessitates extensive expertise, as arriving at appropriate solutions demands significant investments of manpower, meticulous

analysis, and troubleshooting efforts. Leveraging LLMs can enhance the capabilities of MNOs in addressing these challenges and enable more efficient troubleshooting. Particularly, MNOs have at their disposal a rich repository of tickets accumulated over time from dealing with network anomalies. These tickets capture real-world scenarios, encompassing diverse problems and equipment malfunctions. An illustrative example of such a ticket is shown in Fig. 3. By utilizing this repository with product manuals as training data, the LLM can be fine-tuned to comprehend the intricacies of network issues and grasp the unique context of anomaly resolution. Consequently, the LLM becomes an anomaly-solving tool for telecommunications professionals, furnishing them with diagnoses of network issues and their corresponding solutions. Furthermore, leveraging the time-stamped data from the tickets, the LLM can estimate the duration required to address network faults, accounting for the product type, hardware specificities, and the attributes of the involved BSs. As a result, the LLM becomes an asset for the MNO, enhancing the efficiency and effectiveness of resolving network problems.

3GPP Specifications Comprehension

The 3GPP produces the specifications that define cellular telecommunications technologies, including radio access, core network and service capabilities. 3GPP documents are known for their elaborateness, encompassing many details and specifications. Due to the sheer volume of these documents, keeping track of all the specificities, especially in the context of new releases, can be daunting and time-consuming. For engineers attempting to implement technologies and features in the product, this challenge becomes even more apparent, as they must invest considerable time in searching for relevant information within the extensive documentation. LLMs offer a resolution to this problem, providing promising solutions for engineers grappling with 3GPP documents. Through finetuning to the 3GPP documents and incorporating all relevant reports, these models can become adept at processing the vast 3GPP standard knowledge. Then, a significant application of these fine-tuned LLMs revolves around the development of interactive chatbots tailored for answering 3GPP standards queries. These chatbots, built upon the fine-tuned LLMs, empower engineers to streamline their research processes, saving valuable time and facilitating more efficient and accurate implementations of 3GPP standards.

Network Modeling

The optimization of mobile networks is a complex task that requires multiple models for capturing different Key Performance Indicators (KPIs) of the network and the interactions between various network configuration parameters. Such optimization often relies on white-box models, where interactions between multiple features are mathematically formulated to ensure explainability. Developing such models requires expert engineers with deep domain knowledge to identify relevant information and relationships driving the interactions between features. Leveraging LLMs can support the development of these models.

To better clarify this aspect, we provide an explanatory example. Let us consider a simple scenario with a network composed of 90 single-carrier BSs. We used GPT-3.5 as the LLM. The LLM was provided with a list of 12 data features, such as BS location, frequency, and load, and tasked to select the relevant features for creating a model to estimate energy consumption based on the selected features. Additionally, we asked the LLM to provide a mathematical formula capturing the relationship between inputs and outputs and a script to fit the model on a dataset containing real network data. GPT-3.5 successfully identified the 5 relevant inputs among the provided features, while discarding the irrelevant ones. Notably, this was achieved solely on the basis of its knowledge, without using any data samples. The model provided by GPT-3.5 consisted of a weighted sum of the selected features for regression.

```
Troubleshooting_Ticket.JSON =
{
'eNodeB_ID':              'B01'
'Alarm_ID':               'RAN_Site01_LTE_RF_VSWR'
'Alarm Characteristics':  'BS_Type:BTS5900,
                           Board_Type=MRRU,
                           Affected_RAT=GL,
                           VSWR_Alarm_Threshold(0.1)=14,
                           VSWR(0.1)=14,
                           Output_Power(0.1 dBm)=432'
'Alarm Time':             '4/6/2023 2:10:04 PM'
'Dispatch Time':          '4/6/2023 3:48:36 PM'
'Resolution Time':        '4/6/2023 4:36:14 PM'
'Diagnose Summary':       'RF Unit Voltage Standing Wave
                           Ratio threshold crossed due to the
                           feeder being bent or damaged'
}
```

FIGURE 3. An example of a network anomaly troubleshooting ticket. Information related to the anomaly is automatically generated by the system (in blue). Input regarding the dispatch and the resolution of the anomaly is provided by the engineer (in orange).

Figure 4 shows the real energy consumption measured by the BSs at different downlink loads. The real data reveal three different trends, corresponding to the three different configurations of maximum transmit powers in the considered network. Figure 4a shows the estimations performed by the model provided by GPT-3.5, which achieved a relative error of 7.8%. The estimations produced by this model resulted in a single average trend, as the selected inputs were summed, overlooking the relationship between the load and the maximum transmit power: in fact, these two terms should be multiplied and not summed. To address this limitation, we provided contextual data related to the dynamics driving the energy consumption of a generic BS. By leveraging this, GPT- 3.5 produced a different model where the two terms were multiplied instead of summed, significantly reducing the error to 3%. The improved model correctly captures all three trends (Fig. 4b), highlighting the importance of providing telecom related context.

Figure 5 illustrates the average hourly energy consumption in the selected network and the estimates performed by the two models provided by GPT-3.5 (i.e., with and without context). To provide a basis for comparison, we present the estimations from two alternative models: a naive model, and an expert designed ML model [11]. The naive model estimates the energy consumption in a given hour by averaging the energy consumption measured at the same hour of the day in the previous week. While simple, this model

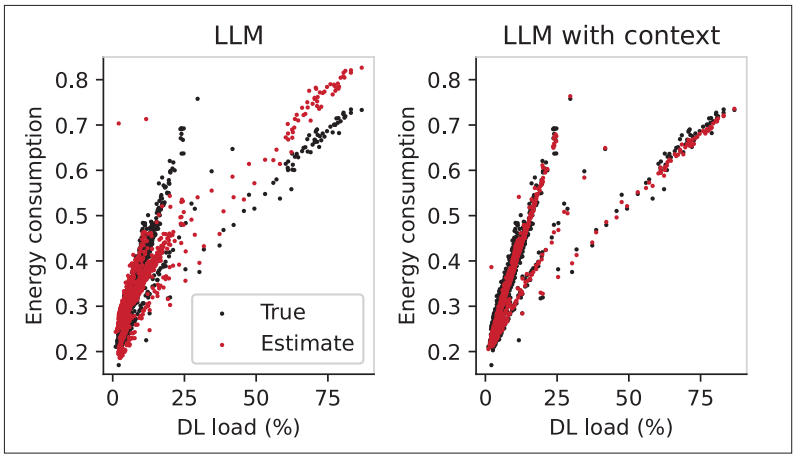

FIGURE 4. Normalized energy consumption measured at different downlink loads and estimated by the model provided by: a) GPT-3.5; b) GPT-3.5 with context.

lacks knowledge of the telecom field and consequently yields an error rate of 12%. On the other hand, the expert-designed ML model employs a ML model designed to handle more intricate scenarios, such as multi-carrier BSs utilizing multiple energy-saving features. In this simplistic setup, the expert-designed ML model achieves a relative error of 2.3%. Significantly, GPT-3.5 capitalized on its knowledge to develop a model that surpassed the limitations of the naive approach, realizing a 75% improvement in accuracy, closely approaching the performance of the expert-designed ML model.

As a final point, it is crucial to highlight that the choice of the LLM employed for a task significantly influences the quality of the achieved solution. To illustrate this, we conducted the same experiment using LLaMA-70B as the LLM. In this case, LLaMA identified additional input features compared to those selected by GPT-3.5, including location, and the year of production of the BS. The model proposed by LLaMA took the form of a weighted sum of the chosen features, similar to the approach proposed by GPT-3.5, resulting in a similar error rate of 7.6%. However, akin to GPT-3.5, the LLaMA model struggled to recognize the relationship between load and maximum transmit power. In contrast to GPT-3.5, though, LLaMA was unable to rectify this issue even when provided with additional contextual information.

Open Research Directions

Extending upon the previously discussed limitations of LLMs and their forthcoming use cases in the telecom industry, a set of open research directions presents itself. These avenues of investigation are crucial to unlock the full potential of LLMs in the telecom industry and harness their capabilities to the utmost extent.

Telecom Foundation Model

While the most advanced foundation models exhibit a reasonable grasp of the telecommunications theory, they fall short on practical implementation knowledge [12]. Besides, our findings, illustrated in Fig. 4, have demonstrated the performance gap between a context-aware LLM and a generic counterpart, shedding light on the necessity of a specialized telecom foundation model. This is further validated in [12]. Such a specialized model should leverage standards, white papers, research literature, and even exclusive proprietary materials or synthetic datasets produced through simulators like digital twins.

Three approaches are available to integrate further knowledge into a language model: full model training, fine-tuning, and RAG. Full model training achieves a profound understanding of the additional knowledge at the expense of substantial energy and complexity costs. Fine-tuning offers a pragmatic balance, enabling model specialization via training a minimal number of parameters using methods like low-rank adaptation (LoRA). Meanwhile, RAG is the most convenient solution. It is cost-efficient and does not require access to the model weights. It incorporates external knowledge using a more surface-level comprehension by querying a database for context to append to the prompt, which may limit the depth of understanding.

It is worth mentioning that some lines of work investigate non-language-based telecom foundational models. Notably, graph-based foundational models could natively capture the natural topology of telecom networks. Such approaches remain in an early exploratory phase.

Benchmarking LLMs for Telecom

In the last years researchers have proposed a number of tests to evaluate LLM capabilities in terms of NLP, for example, text understanding and reasoning. Recent LLMs are already close to human-level performance on several of these tests such that HellaSwag, a test of commonsense inference, and GLUE/SuperGLUE, which evaluate LLM linguistic understanding. MMLU, instead, evaluates LLMs' multitask accuracy and capabilities across a broad range of subjects, and show that top-performing LLMs have still significant room for improvement before achieving expert-level accuracy across specialized tasks. In all these tests, accuracy on multiple choice questions is computed to provide a simple to determine and understand evaluation. Some researchers have suggested that the future of NLP evaluation should focus on text generation: however, while some metrics exist for testing these capabilities such as BLEU and perplexity, text generation is notoriously difficult to assess and still lacks a standard evaluation methodology. Although ML researchers have mainly focusing on NLP capabilities of LLMs, the success of LLMs is the telecom industry depends on benchmark datasets designed to assess their proficiency in this specific domain. These datasets are expected to play a pivotal role in determining the optimal architectural design for LLMs and guiding the pretraining procedure in the development of telecom foundational models. The framework in [12] proposes a multiple-choice question dataset to simply evaluate the accuracy of telecom knowledge of LLMs; future works will need to extend this framework and allow the evaluation of LLMs across specialized telecom tasks such as those discussed previously.

LLMs Compression

As highlighted earlier, LLMs can be comprised of billions of parameters and require powerful devices to be trained and inferred. This limitation becomes relevant in critical scenarios where LLMs need to be deployed in edge devices with lim-

ited storage and computational capabilities. As a result, it is imperative to address the substantial size of LLMs and develop compression techniques [13], which can reduce their size while retaining their knowledge of the telecom domain. Pruning, quantization, and knowledge distillation are the three most popular model compression techniques for DL models. Today, researchers believe that quantization outperforms pruning in most of LLM architectures. Then, due to the large costs of training LLMs, post-training quantization, where weights and activation tensors are encoded with a low-level of precision, for example, 8-bit or 4bit instead of 16-bit, is the main adopted scheme. Indeed most of the open source LLMs offer quantized versions of larger models. In addition, knowledge distillation is currently explored to develop compact LLMs that can run on devices with limited resources. To conclude, compression methods reduce memory and computational resource usage but can degrade LLM performance, and thus accuracy pre- and post compression has to be evaluated to analyse pros and cons of the existing and future techniques.

Privacy Considerations

Adapting LLMs to address specific telecom-related tasks may require the use of datasets containing sensitive user information. In light of this, it becomes imperative to implement measures to protect privacy when handling such data. Included among these measures are data anonymization and aggregation, effectively removing personally identifiable information to protect individual privacy. The incorporation of techniques such as differential privacy is essential to ensure that these models remain impervious to leaking sensitive information during queries. Additionally, the development of smaller LLMs that can run on edge devices will further enhance the end user's privacy.

Behavior Alignment

Solving the problem of output inconsistency is essential to enable the adoption of LLMs in the telecom industry, especially in accuracy-critical areas. It has been shown that grounding LLMs with use-case-specific external tools, such as querying external knowledge with RAG, reduces hallucinations [14]. Besides, it is crucial to incorporate mechanisms and metrics to assess the model's prediction confidence. Such mechanisms enable the identification of uncertain cases, triggering additional verification from humans in the loop. In order to measure prediction confidence, methods include using the LLM's internal evaluation of the likelihood of the output, generating multiple responses to a single query to assess consistency, or using one LLM to review and refine the output of another. Additionally, rigorous testing of LLMs against adversarial inputs and scenarios can help to reveal vulnerabilities and guide the development of reliable models. Finally, understanding prompt engineering is necessary, given that well-designed queries and instructions play a crucial role in shaping the model's behavior and ensuring accurate outputs.

LLMs Explainability

The need for explainability in LLMs within the telecom industry is paramount due to stakeholder concerns regarding trust and reliance on ML outputs, especially considering their limitations previously discussed previously. The adoption of LLMs for critical operations require a clear understanding of how and why specific outputs are generated. This necessitates the incorporation of explainability techniques such as referencing, where LLMs can provide sources or justifications for their responses. Additionally, explicitly integrating explainability objectives into the training process is crucial for this purpose.

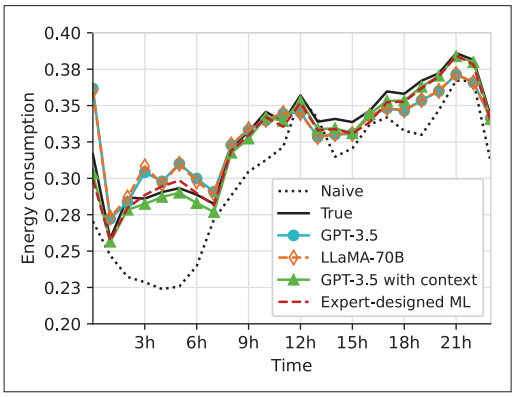

FIGURE 5. Normalized hourly energy consumption in the network - Actual measurements (in black) and estimations from various models.

Real-time Context

By design, LLMs are trained offline on large corpora of data and, therefore, are not aware of new findings that may be accessible through search engines. Consequently, prompting these LLMs can lead to potentially outdated answers, especially considering that the telecom industry continuously evolves with releases of new technical specifications. One approach to address this issue is to enable LLMs to access external tools. For instance, allowing LLMs to access the internet through dedicated channels, as OpenAI has done with ChatGPT. However, this approach confines the quality of LLM generation to the outcomes derived from search queries. A more fundamental strategy is to create data pipelines to gather new relevant telecom knowledge. This knowledge can then be utilized by either augmenting queries through RAG approaches (e.g., as done by Grok, the LLM developed by XAI, with tweets) or by conducting additional training of the LLM to refine its parametric knowledge. The latter approach introduces various research possibilities, such as identifying the optimal frequency for updating the LLM's parametric knowledge and developing efficient methodologies for model updates to integrate the new material.

Sustainability and Environmental Impact

Given their large parameter count, LLMs pose a substantial environmental concern in terms of carbon footprint. To mitigate these challenges, prioritizing smaller and more efficient models (e.g., Phi-2 that compete with larger models) is recommended. Furthermore, incorporating efficient implementations of attention mechanisms and overall model architecture can substantially alleviate computational demands during both training and inference. For instance, adopting the FlashAttention mechanism [15] or employing the mixture of experts architec-

From another perspective, tackling the sustainability challenge also involves the development of KPIs and regulations that effectively measure, evaluate, and compare the environmental footprint of different LLMs.

ture, as demonstrated by models like Mixtral, offers promising avenues for reducing computational loads. From another perspective, tackling the sustainability challenge also involves the development of KPIs and regulations that effectively measure, evaluate, and compare the environmental footprint of different LLMs.a

LLMs as Orchestrators

Leveraging even further the reasoning capabilities of the LLMs, an open research direction involves transitioning from a strict parametric knowledge framework to a different paradigm where LLMs serve as orchestrators, as introduced earlier. In this scenario, LLMs are granted access to fine-grained blocks, such as code interpreters, optimizers, signal processing blocks, and network models. Their role then shifts to translating user prompts into actionable steps by leveraging both their knowledge and harnessing the accessible blocks. In this context, the research avenues revolve around defining these fine-grained blocks and ensuring seamless integration between LLMs and these blocks to unlock their potential.

Conclusions

In this article, we have delved into the inner workings of LLMs, shedding light on their current capabilities and limitations. Additionally, we explored various use cases of LLMs that can be promptly leveraged within the industry using the available data at vendors' disposal. Furthermore, we discussed the specific open research directions tailored to the peculiarities of the telecom domain, which must be addressed to fully harness the potential of LLMs. As the technology behind LLMs continues to evolve, the telecom industry is poised to seize the opportunity and leverage these advancements to enhance operational efficiency within the sector.

References

[1] A. Vaswani *et al.*, "Attention is All you Need," *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[2] OpenAI, "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.

[3] M. U. Hadi *et al.*, "Large Language Models: A Comprehensive Survey of Its Applications, Challenges, Limitations, and Future Prospects," Nov. 2023; available: http://dx.doi.org/10.36227/techrxiv.23589741.v4.

[4] H. Holm, "Bidirectional Encoder Representations from Transformers (BERT) for Question Answering in the Telecom Domain: Adapting a BERT-like language model to the telecom domain using the ELECTRA Pre-Training Approach," *KTH, School of Electrical Engineering and Computer Science, Tech. Rep.*, 2021.

[5] L. Bariah *et al.*, "Understanding Telecom Language Through Large Language Models," arXiv preprint arXiv:2306.07933, 2023.

[6] Y. Du *et al.*, "The Power of Large Language Models for Wireless Communication System Development: A Case Study on FPGA Platforms," arXiv preprint arXiv:2307.07319, 2023.

[7] L. Bariah *et al.*, "Large Language Models for Telecom: The Next Big Thing?" arXiv preprint arXiv:2306.10249, 2023.

[8] Y. Bang *et al.*, "A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity," arXiv preprint arXiv:2302.04023, 2023.

[9] L. Chen *et al.*, "How is ChatGPT's Behavior Changing Over Time?" arXiv preprint arXiv:2307.09009, 2023.

[10] Y. Wolf *et al.*, "Fundamental Limitations of Alignment in Large Language Models," arXiv preprint arXiv:2304.11082, 2023.

[11] N. Piovesan *et al.*, "Machine Learning and Analytical Power Consumption Models for 5G Base Stations," *IEEE Commun. Mag.*, vol. 60, no. 10, 2022, pp. 56–62.

[12] A. Maatouk *et al.*, "TeleQnA: A Benchmark Dataset to Assess Large Language Models Telecommunications Knowledge," arXiv preprint arXiv:2310.15051, 2023.

[13] T. Dettmers *et al.*, "QLoRA: Efficient Finetuning of Quantized LLMs," arXiv preprint arXiv:2305.14314, 2023.

[14] K. Shuster *et al.*, "Retrieval Augmentation Reduces Hallucination in Conversation," arXiv preprint arXiv:2104.07567, 2021.

[15] T. Dao *et al.*, "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness," arXiv preprint arXiv:2205.14135, 2022.

Biographies

Ali Maatouk is a Researcher with Huawei Technologies, France.

Nicola Piovesan is a Senior Researcher with Huawei Technologies, France.

Fadhel Ayed is a Senior Researcher with Huawei Technologies, France.

Antonio De Domenico is a Senior Researcher with Huawei Technologies, France.

Mérouane Debbah is a Professor at Khalifa University in Abu Dhabi.