IEEE Open Journal of the Communications Society

Received 14 July 2024; revised 26 August 2024; accepted 5 September 2024. Date of publication 9 September 2024; date of current version 17 September 2024.

Digital Object Identifier 10.1109/OJCOMS.2024.3456549

LLM-Based Edge Intelligence: A Comprehensive Survey on Architectures, Applications, Security and Trustworthiness

OTHMANE FRIHA [1], MOHAMED AMINE FERRAG [2] (Senior Member, IEEE), BURAK KANTARCI [1] (Senior Member, IEEE), BURAK CAKMAK [3], ARDA OZGUN [3], AND NASSIRA GHOUALMI-ZINE [4]

[1] School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada

[2] Artificial Intelligence and Digital Science Research Center, Technology Innovation Institute, Abu Dhabi, UAE

[3] Headquarters, Edge Signal, Ottawa, ON K2K 0G7, Canada

[4] Department of Computer Science, Badji Mokhtar-Annaba University, 23000 Annaba, Algeria

CORRESPONDING AUTHOR: B. KANTARCI (e-mail: burak.kantarci@uottawa.ca)

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) Discovery Program; in part by the NSERC CREATE TRAVERSAL Program; and in part by the Innovation for Defence Excellence and Security (IDEaS) Program from the Department of National Defence (DND).

ABSTRACT The integration of Large Language Models (LLMs) and Edge Intelligence (EI) introduces a groundbreaking paradigm for intelligent edge devices. With their capacity for human-like language processing and generation, LLMs empower edge computing with a powerful set of tools, paving the way for a new era of decentralized intelligence. Yet, a notable research gap exists in obtaining a thorough comprehension of LLM-based EI architectures, which should incorporate crucial elements such as security, optimization, and responsible development. This survey aims to bridge this gap by providing a comprehensive resource for both researchers and practitioners. We explore LLM-based EI architectures in-depth, carefully analyzing state-of-the-art paradigms and design decisions. To facilitate efficient and scalable edge deployments, we perform a comparative analysis of recent optimization and autonomy techniques specifically designed for resource-constrained edge environments. Additionally, we shed light on the extensive potential of LLM-based EI by demonstrating its varied practical applications across a wide range of domains. Acknowledging the utmost importance of security, our survey thoroughly investigates potential vulnerabilities inherent in LLM-based EI deployments. We explore corresponding defense mechanisms to protect the integrity and confidentiality of data processed at the edge. In conclusion, highlighting the essential aspect of trustworthiness, we outline best practices and guiding principles for the responsible development and deployment of these systems. By conducting a comprehensive review of these key components, our survey aims to support the ethical development and strategic implementation of LLM-driven EI, paving the way for its transformative impact on diverse applications.

INDEX TERMS Edge intelligence (EI), generative AI; large language models (LLMs), security, privacy, trustworthiness, responsible AI.

I. INTRODUCTION

LANGUAGE, the foudnation of human interaction [1], is a complex system of symbols and principles. It serves as the vehicle for expressing and understanding thoughts, concepts, and emotions. This complex structure includes spoken, written, and gestural forms, enabling individuals to communicate meaning within a social setting. As we examine the historical path from the inception of the Turing Test [2], it is evident that the ongoing effort to grant machines with language intelligence remains a continual focus within the realm of Artificial Intelligence (AI) exploration.

TABLE 1. List of abbreviations.

AI	Artificial Intelligence
EI	Edge Intelligence
PLM	Pre-trained Language Model
LM	Language Models
LLM	Large Language Model
MLLM	Multi-modal LLM
ViT	Vision Transformer
ML	Machine learning
DL	Deep Learning
FL	Federated Learning
DRL	Deep Reinforcement Learning
KD	Knowledge Distillation
RL	Reinforcement Learning
IoT	Internet of Things
IoV	Internet of Vehicles
MEC	Mobile Edge Computing
GPT	Generative Pretrained Transformer
SLM	Statistical Language Model
NLM	Neural Language Model
PEFT	Parameter-Efficient FineTuning
P2P	Peer-to-Peer
AS	Autonomous Systems
PTQ	Post-Training Quantization
QAT	Quantization-Aware Training
MIA	Membership Inference Attacks
AIA	Attribute Inference Attacks
PII	Personally Identifiable Information
ASR	Attack Success Rate
JSR	Jailbreak Success Rate
DSR	Defense Success Rate
DP	Differential Privacy
SGD	Stochastic Gradient Descent
IRP	Incident Response Plan
IDS	Intrusion Detection Systems
XAI	Explainable AI
TSP	Transactional Stream Processing

Unlike humans, who naturally acquire expressive communication skills, machines lack the inherent capacity to understand and express human language. Such proficiency requires advanced AI, particularly in Natural Language Processing (NLP) algorithms. The ongoing research challenge centers on developing AI systems capable of reading, writing, and interacting with humans at a comparable level. In recent years, we have witnessed a revolution in NLP that has been propelled by language models (LM). Essentially, an LM aims to capture the probability of generating word sequences, facilitating the prediction of probabilities for forthcoming or absent tokens [3].

The ongoing pursuit for machines to truly comprehend and understand natural language, deciphering its nuanced meaning, has faced persistent challenges. The concept of "*understanding*" in the context of language models, as emphasized in the literature, is frequently misunderstood and associated with overclaims. As articulated in [4], one argument challenges these assertions by highlighting a fundamental misinterpretation of the relationship between linguistic form and meaning. The language modeling task, predominantly relying on linguistic form as training data, introduces inherent barriers that hinder machines' genuine acquisition of meaning.

NLP has served as the cornerstone for enabling computers to comprehend and generate human language, encompassing a variety of tasks such as machine translation [3]. Central to these tasks are LMs [5], which predict the likelihood of word sequences, ranging from simple n-grams to sophisticated neural networks [6], [7]. Distinctly, PLMs represent a subset of LMs that, through extensive pre-training on vast datasets, offer superior performance and efficiency when fine-tuned for specific tasks [8], [9], [10]. Examples of such models include GPT-1, GPT-2 [11], [12] and BERT [13]. Thus, while NLP covers the broad spectrum of language-related tasks, LMs focus on sequence prediction, and PLMs further enhance LMs by leveraging large-scale pre-training for task-specific optimization.

It has become apparent that scaling up these models improves their overall capacity and introduces *emergent abilities*, defined by characteristics that appear in larger models but not in the smaller ones [14]. Contemporary LMs have undergone massive scaling predominantly across three key dimensions: the magnitude of computational resources employed, the number of model parameters integrated, and the expansiveness of the training dataset utilized [14], hence the name *Large Language Models (LLM)*. This triad of factors collectively reflects the extensive efforts to enhance the capabilities and performance of LMs in today's computational landscape. This scaling endeavor has given rise to more potent LLMs (*e.g.,* GPT-4 [15]). These advanced models have transitioned beyond mere language modeling, extending their capabilities to proficiently tackling diverse, complex tasks [3], marking a significant evolution in the field.

In a landscape where global associations appear to engage in a competitive race to create increasingly more significant LMs, exploring LLMs properties and their transformations with size emerges as a compelling area of scientific interest [16]. Amidst this pursuit of innovation, a pertinent question arises within the research community – Has thorough consideration been devoted to the potential risks linked to the expansion of these models, and are existing strategies in place to alleviate these risks?

In their study, Bender et al. [17] highlighted a consequential effect wherein not only does the environmental impact scale with model size, but it also introduces challenges in comprehending the content within the extensive training data. This dual impact raises critical considerations regarding the sustainability and interpretability of LLMs, shedding light on the broader implications associated with their burgeoning scale. In addition, the prevailing scenario involves the predominant use of applications based on LLMs in the cloud,

which necessitates users to transmit data to uncontrollable (and sometimes unknown) locations; this approach encounters significant challenges. These challenges encompass prolonged response times, elevated bandwidth costs, and apprehensions regarding data privacy breaches [18].

While significant advancements have been made in AI in recent years, challenges remain regarding the transparency, fairness, environmental impact, and trustworthiness of AI systems [17], [19], [20]. These issues necessitate the development of sustainable and interpretable AI models on a global scale. One key effort in addressing these concerns comes from the regulatory framework established by the European Union (EU) in the AI Act [21]. The AI Act intends to make clear classifications of AI systems in terms of the risks they pose. The Act proposed four distinct categories: unacceptable risk, high risk, limited risk, and minimal risk. What is rated as AI systems posing an "unacceptable risk" will be prohibited, while "high-risk," which would be deployed in critical infrastructure or impact fundamental rights, would be rigorously regulated. By imposing stringent requirements on high-risk AI systems, the AI Act seeks to ensure that AI technologies used in Europe are reliable and respect fundamental rights, thus fostering an environment of trust and accountability in AI applications. This regulatory approach underlines why the issues illustrated earlier need to be resolved and, therefore, reasons the urge for sustainable and interpretable models of AI worldwide in innovation.

While cloud-based deployments have various benefits, such as processing power, they also have inherent limitations. Take the example of a self-driving car that makes decisions in real-time–which it does by querying a cloud-based LLM. In areas with low or spotty network connectivity, there would be latency in communicating with the cloud, which may cause the car to delay responding to unexpected road hazards while it is waiting on necessary information from the LLM.

Acknowledging these limitations, particularly within the realm of LLM's multimodality, the deployment on the Edge layer stands out as a compelling alternative [22]. This alternative is gaining traction not only in academic circles, as evidenced by works such as [23], [24], [25], [26], [27], [28], but also in the corporate sphere, where industry leaders like Intel [29], NVIDIA [30], Microsoft [26] and Qualcomm [31] are actively embracing this concept. Closer deployment of LLMs on the network edge (on-premise), enhances data privacy [18], reduces costs [32], reinforces autonomy [33], and provides real-time responses [34], addressing pivotal challenges associated with the current cloud-based deployment [35], offering potential solutions by adopting the edge computing paradigm.

Edge-based LLMs will empower devices with some level of autonomy in decision-making, which minimizes their reliance on centralized control. This can be very critical in scenarios characterized by limited or unreliable connectivity. For example, if a smart home is equipped with an LLM for managing its energy consumption and the Internet is out for some time–for example, there is a case of temporary Internet outage–it will still be able to optimize energy use and hence guarantee efficient operation. Also, sensitive data, like the feeds from security cameras or medical scans analyzed by LLMs, can reside on-premises at the edge, circumventing some of the pernicious privacy problems associated with storage in the cloud. For instance, if a hospital is running an LLM for the analysis of medical images at a nearby edge, this would ensure the preservation of patient data since those images will never leave the network of the respective hospital.

However, the integration of LLM into EI architectures presents a range of deployment and security challenges [20], [36]. For instance, LLMs, with their significant computational and memory requirements, can be difficult to accommodate on edge devices, which often have limited resources. This mismatch not only impacts the performance of LLMs but also exposes the EI systems to various security vulnerabilities [37]. LLMs can be integrated into EI architectures through techniques such as model compression, efficient memory management, and distributed computing [38], [39], [40]. For instance, employing methods like quantization and pruning can reduce the model size, making it more suitable for edge deployment [23], [41]. Additionally, leveraging hierarchical memory systems and advanced attention mechanisms can help manage the limited memory resources more effectively [40].

The possibilities for LLM-base EI are endless. Consider a wind farm equipped with edge devices that host LLM models. They are capable of ingesting data directly from wind turbines, ranging from sensor readings on the speed of wind to the Revolutions Per Minute (RPM) of the rotor, and the environmental conditions around it. The LLM shall use that data to identify potential maintenance issues. For example, based on the historical data and finding certain patterns in the sensor readings, an LLM could be able to determine whether these readings are indicative of a bearing failure in a wind turbine. Afterward, it generates a natural language report on the issue, the recommended maintenance actions, and the possible consequences in case those actions are not taken. Real-time analysis onsite allows for proactive maintenance to avoid undesirable production losses due to breakdowns, and to maximize energy harvesting. Furthermore, in the domain of Industrial IoT (IIoT), LLMs deployed at the edge can transform manufacturing processes. Imagine factory robots equipped with LLMs. This kind of robot would do pre-defined tasks but would also be able to analyze shop sensor data in search of inefficiencies or potential safety hazards. For instance, an LLM could assess data that is coming from a robotic arm and identify a deviation in its pattern of movement. It could then build a report suggesting recalibration of the robotic arm or even flag a potential safety risk if the deviation is significant. This real-time, on-device analysis empowers robots to do more than execute tasks: it makes them capable of participating in a rudimentary form of self-monitoring that underpins a safer and more efficient production environment.

TABLE 2. An overview of selected surveys on EI, Generative AI, and LLMs.

	Work	Year	Main focus	LLM-based EI	Arch.	App.	O/A	Security		Trust.
								Attacks	Defences	
Edge Intelligence	[42]	2019	Analyzes recent advancements in EI, focusing on architectures, frameworks, and technologies for running DL on edge devices.	○	●	●	◐	○	○	○
	[43]	2019	Explores Edge Information Systems for smart vehicles, focusing on design challenges, methods, hardware, and use cases	○	◐	◐	●	○	○	○
	[44]	2019	Explores applications, implementation methods, and challenges for combining Edge Computing and DL for on-device intelligence.	○	●	●	◐	◐	◐	○
	[45]	2020	Analyzes core functions of EI (caching, training, inference, offloading) and explores existing solutions and open challenges.	○	●	●	◐	●	◐	○
	[46]	2020	Examines challenges and solutions for FL, focusing on applications in mobile edge network optimization.	○	●	●	◐	◐	○	○
	[47]	2021	Proposes a 6G architecture with AI and network virtualization for flexible, intelligent services.	○	●	●	◐	○	○	○
	[48]	2021	Proposes a unified design for edge AI, combining communication, decentralized learning, and system architecture for real-world applications.	○	●	●	◐	●	○	●
	[49]	2021	Explores security and privacy challenges in B5G edge AI, proposing a blockchain framework for solutions.	○	◐	◐	○	●	●	○
	[50]	2022	Analyzes how EI and Blockchain can work together, exploring limitations, benefits, and implementation details for future networks.	○	●	●	○	●	●	○
Generative AI	[51]	2024	Explores deploying AI content generation on mobile edge networks for real-time, personalized experiences, considering privacy and future challenges.	●	●	●	◐	◐	◐	◐
	[52]	2024	Analyzes methods to reduce resource consumption of powerful AI models (LLMs, ViTs etc.), covering design, training, and deployment aspects.	●	●	◐	◐	○	○	○
	[20]	2024	Explores both benefits and drawbacks of LLMs from the perspective of security and privacy.	○	○	○	○	●	●	◐
	[53]	2024	Examines security and privacy risks of LLMs, exploring attack methods, potential leaks, and areas needing further research.	○	○	○	○	●	●	◐
	[54]	2024	Analyzes of recent LLMs focusing on their design, training data, evaluation metrics, and future research directions	○	●	○	●	○	○	○
Generative AI based EI	[55]	2023	Proposes using 6G edge computing to run powerful LLMs on devices, solving cloud-based issues like slowness, data privacy, and cost.	●	●	○	◐	○	○	○
	[56]	2023	Explore generative AI techniques at the wireless edge networks	●	●	●	○	○	○	○
	This survey	2024	Comprehensive Survey on LLM based EI architectures, recent optimization and autonomy techniques, practical applications, security issues, defense mechanisms and trustworthiness for responsible EI	●	●	●	●	●	●	●

Not Considered (○); Partial discussion (◐); Considered (●);
Architectures (Arch.); Applications (App.); Optimization/Autonomy (O/A); Trustworthiness (Trust.);

A. RESEARCH GAPS AND CONTRIBUTIONS

In Table 2, we present an overview of selected EI and Generative AI surveys. Alongside the surveys that study Edge Intelligence (EI) and Generative AI individually. To the best of our knowledge, only two surveys cover Generative AI-based EI. However, none of the existing

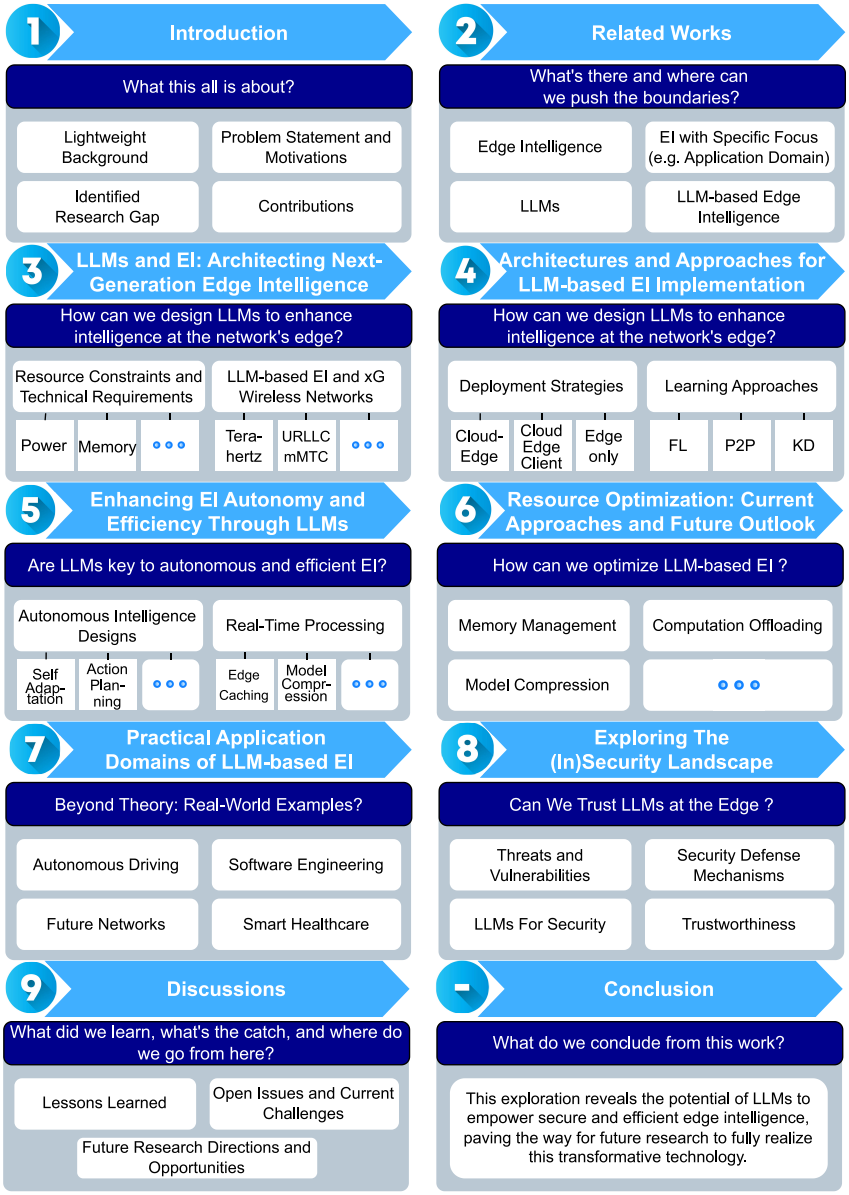

FIGURE 1. Survey Structure.

surveys complement the LLM-based EI, architectures, and applications with a thorough review of security issues from attack and defense points nor trustworthiness issues in Generative AI-based Edge Intelligence.

Fig. 1 illustrates this survey's structure. As illustrated, the purpose of this paper is to fill the gap in the literature by providing a comprehensive and holistic resource to the research community working on EI powered by LLMs. This aims to help researchers, developers, and practitioners in this rapidly evolving field. First, this study explores various aspects of LLM-based architectures. Second, we focus on the optimization and autonomy techniques involved in an LLM-based EI system. Third, we showcase the practical applications of LLM-powered EI across diverse domains. Fourth, we conduct a rigorous security analysis of LLM-based EI systems. Finally, by providing a unified resource and in-depth analysis, this study lays the groundwork for efficient, safe, and responsible development of LLM-powered EI. The main contributions of this study are:

- *In-Depth Exploration of LLM-based EI Architectures:* We systematically analyze cutting-edge architectural paradigms designed for seamless LLM integration at the network edge. This analysis will provide valuable insights into the design choices and trade-offs in crafting efficient and scalable LLM-powered EI systems.
- *Comparative Analysis of Optimization and Autonomy Techniques:* We delve into recent advancements in optimization and autonomy techniques specifically tailored for LLM-powered EI systems. This comparative

analysis will shed light on effective strategies for managing resource constraints, improving accuracy, and enabling intelligent decision-making at the edge.
- *Diverse Practical Applications:* We showcase the versatility of LLM-based EI by exploring its applications across a broad spectrum of domains. This exploration will illuminate the transformative potential of LLMs at the edge in areas like intelligent automation, personalized healthcare, and software engineering.
- *Rigorous Security Analysis:* We thoroughly examine security vulnerabilities inherent to LLM-based EI deployments. This analysis will identify potential attack vectors and explore defense mechanisms to mitigate security risks and ensure the integrity and confidentiality of data processed at the edge.
- *Trustworthiness Considerations for Responsible Development:* We emphasize the critical importance of trustworthiness in LLM-based EI. We outline best practices and guiding principles for these systems' responsible development and deployment, ensuring ethical considerations are addressed throughout the entire life cycle.

The rest of the paper is organized as follows. We present the related works in Section II. In Section III, we systematically explore cutting-edge architectural paradigms designed for seamless LLM integration at the network edge. In Section IV, we explores recent deployment strategies and learning approaches for effective LLM-based EI. Section V review recent advancements in optimization and autonomy techniques tailored explicitly for resource-constrained LLM-powered EI systems. In Section VI, we explores recent resource optimization strategies for LLM-based EI. We examines the delicate balance between maximizing model capabilities and navigating the constraints imposed by edge devices. Furthermore, we offers an outlook on the relationship between LLM-based EI and future 6G networks, anticipating mutually beneficial advancements. In Section VII, we showcase the versatility of LLM-based EI by exploring its diverse applications across a broad spectrum of domains, illuminating its transformative potential. In Section VIII, we thoroughly examine security vulnerabilities inherent to LLM-based EI deployments, identify potential attack vectors, explore corresponding defense mechanisms, and emphasize the critical importance of trustworthiness in LLM-based EI. In Section IX We summarize the key takeaways from this survey, discuss its broader implications, and highlight promising avenues for future research in LLM-powered EI. Finally, Section X briefly concludes the survey.

II. RELATED WORKS

The number of AI services and applications has increased exponentially due to recent developments in Deep Learning (DL) [57]. The proliferation of mobile computing and the Internet of Things (IoT) has linked billions of devices, producing enormous amounts of data at the network edge. Extending AI capabilities to the network edge is essential to realizing the full potential of this massive volume of data [58]. As the need for intelligent processing at the network edge grows, this necessity has given rise to a new multidisciplinary field known as Edge AI, or Edge Intelligence (EI) [42]. In the literature, various recent works have explored the concept of extending AI capabilities to the network edge [42], [43], [44], [45], [47], [48], [50]. These studies delve into the challenges, methodologies, and potential applications at the intersection of AI and edge computing. In the following, we summarize these recent works, elucidating their contributions and insights in shaping the landscape of EI.

A. EDGE INTELLIGENCE AND INTELLIGENT EDGE

The distinction between intelligent edge and edge intelligence is explained by [44]. According to this definition, the purposeful transfer of DL computations from the cloud to the edge facilitates the implementation of a variety of dispersed, trustworthy, low-latency intelligent services. This approach will reduce latency, improve reliability, and enable more efficient execution of intelligent services at the edge. Conversely, an Intelligent Edge is defined by integrating DL into the edge with the aim of dynamic, adaptive edge management and maintenance. The objective is to directly integrate DL capabilities into the edge infrastructure to develop more flexible and responsive edge computing environments.

For instance, Zhou et al. [42] provided a thorough survey of recent research in EI, emphasizing the crucial role of AI in harnessing data at the network edge. The authors highlight Edge Computing's potential to enhance AI with diverse data and application designs. The survey systematically reviews DL model architectures, computing architectures (centralized, decentralized, hybrid), and enabling technologies for EI model training, including Federated Learning (FL), Aggregation Frequency Control, and Knowledge Transfer Learning. The analysis extends to the inference phase, covering techniques like Edge caching and model partition, providing a comprehensive perspective on the evolving landscape of EI models at the network edge. Edge caching, training, inference, and offloading are the four core elements of EI that [45] investigated. The authors thoroughly investigated the present conditions of each of these components. For example, they highlighted three critical problems in edge network caching technologies related to the content, locations, and algorithms. In addition, they also examined several training topologies, acceleration strategies, and optimization methodologies for AI models trained in edge contexts. Additionally, they looked at several model construction and compression techniques that could be used for edge intelligence inference.

From a networking perspective, for example in the context of future 6G, two other paradigms arise: AI on edge and AI for the edge. These paradigms form the center of progress in upgrading network technologies to meet the ever-increasing demand for low-latency, high-efficiency data

processing, and real-time decision-making directly at the edges of the network [59], [60], [61]. Enabling the so called an Intelligent Internet of Intelligent Things [60]. *AI on Edge* refers to the direct deployment of AI algorithms onto edge devices, along with a wide range of intelligence-capable things, such as smartphones. The advantages associated with this approach primarily revolve around bypassing latency associated with data transmission to centralized cloud servers by harnessing the power of local data processing. For instance, in autonomous vehicle systems, edge AI processes sensor data, enabling real-time driving decisions critical for safe and high-performance [62]. AI for the Edge refers to development and optimization of AI algorithms for edge computing environments [60], [61]. Designing lightweight neural networks and efficient inference models that can execute, within the constrained computational and power resources of edge devices, would therefore be necessary [63]. Ericsson [64] and Nokia Bell Labs [65] have led the way in the native integration of AI capabilities within network interfaces–a paradigm called AI native. This paradigm avails dynamic, intelligent network management by infusing AI and ML algorithms directly within the infrastructure of the network. According to [65], integrating AI into network interfaces offers various key benefits. For instance, AI can learn custom waveforms for different frequencies, optimizing spectrum use and adapting to hardware and channel limitations. It can also enhance modulation schemes, pilot sequences, and codes.

B. EI FOR NEXT GENERATION NETWORKING

The investigation undertaken by [47] delved into the intersection of AI and future networks, scrutinizing this convergence from a bidirectional standpoint, encompassing both networking for AI and AI for networking. The authors' primary focus was directed towards virtualization and network intelligence. They emphasized the facilitation of various interplays, particularly highlighting the interplay between model-driven and data-driven approaches for network management. In the same context, the work in [48] offered a thorough overview regarding the development of scalable and reliable EI in the context of 6G. This was achieved by aligning the principles of wireless networks with the various configurations of EI implementations. The authors emphasized that it is imperative to incorporate next-generation network architectures to design a communication-efficient edge AI training system. These architectures should be capable of supporting various edge-learning models and topologies. For instance, the authors suggested integrating new multiple antenna techniques to facilitate the rapid exchange of high-dimensional model updates.

Xu et al. [51] explored the integration of cloud computing, edge computing, and mobile technologies. Their research provides valuable insights into the synergies necessary to deliver effective and responsive AI-generated content services to end-users within the dynamic context of EI-enabled mobile network environments. The authors highlighted some benefits of such synergy, as explained in their work, encompass reconfigurability, customization, efficiency, and augmented security and privacy guarantees. Integrating these technologies enhances the adaptability and tolerability of systems. It contributes to improved operational efficiency and heightened security measures, thereby substantiating the viability and advantages of the proposed framework for delivering AI-generated content services.

C. EI FOR DISTRIBUTED/DECENTRALIZED APPLICATIONS

The deployment of EI in distributed and decentralized applications represents a fundamental change in modern computing paradigms. Such integration enables the greater efficiency and robustness of distributed systems by leveraging local processing capabilities at the edge of the network [66]. A key aspect of this area of research is to optimize application performance and security through EI, leading to innovations in application architecture and system design [67]. Zhang and Letaief [43] provide an extensive analysis of Internet of Vehicles (IoV) systems, highlighting the critical role of mobile EI capabilities. Their study identifies three core functions essential to IoV: perception, simultaneous localization and mapping (SLAM), and high-definition mapping. These functions underscore the necessity of environmental awareness in IoV, where Edge deployments can augment perception by offloading intensive processing tasks to local nodes, thereby reducing latency and improving real-time decision-making. This work shows how Edge nodes perception can be improved by state-of-the-art sensors and more powerful processors to collect and analyze locally environmental data, greatly reducing actual reliance on centralized cloud services. This on-device processing will permit faster response times and more accurate situational awareness, both crucial to the safe operation of self-driving vehicles. Simultaneous localization and mapping, another fundamental function, benefits from EI through enhanced data fusion techniques. By processing sensor data locally, Edge nodes can generate more accurate and up-to-date maps, which are important for navigation and obstacle avoidance in dynamic environments. High-definition mapping further leverages EI by allowing for real-time updates and dissemination of detailed environmental maps, facilitating improved route planning and hazard detection. In addition the authors investigated caching strategies at the edge, providing a detailed examination of content caching locations within the IoV framework. The authors classify caching techniques into three main categories: device-level caching, edge-level caching and cloud-level caching. Within each level, there is its own set of benefits and challenges, with edge-level caching offering a balance between latency reduction and storage efficiency. The authors thoroughly analyze various applications using caching, such as video streaming and software updates, demonstrating how strategic placement of cached content can significantly

improve intelligent edge systems performance and user experience.

Wang et al. [50] explored the incorporation of blockchain technology within the EI framework as a way to tackle both deployment and security challenges. The authors pointed out several limitations of the EI systems, notably limited computing capacity and vulnerability to security breaches, and suggested various blockchain-based solutions to address these concerns, with a focus on decentralized management and model optimization. The authors highlighted various consensus algorithms, including proof-of-work (PoW), proof-of-stake (PoS) and delegated proof-of-stake (DPoS), and discuss their applicability in the context of EI. The authors also highlighted the importance of designing efficient incentive mechanisms to encourage participation in EI networks, ensuring that nodes are rewarded for contributing computational resources and maintaining network security. Smart contracts, another important component in the blockchain, were investigated for their potential to automate and secure transactions within EI ecosystems. The study also discussed the challenges associated with smart contract security, including vulnerabilities and scalability issues, and propose several mitigation strategies, such as formal verification techniques and layer-2 scaling solutions, to enhance the robustness and scalability of blockchain-integrated EI systems. The integration of EI in distributed and decentralized applications offers substantial benefits in terms of performance, security, and scalability. By leveraging local processing capabilities and advanced technologies like blockchain, these systems can achieve higher efficiency and resilience, paving the way for innovative applications and robust infrastructures [68].

D. FL AS AN EI ENABLER

Numerous research works have examined specific AI techniques as EI enablers. For instance, a thorough analysis of FL implementations in Mobile Edge Computing (MEC) systems was conducted in [46]. Notably, the authors examined techniques for lowering communication costs, such as Edge and end computation, model compression, and importance-based updating–which averages only the most essential or pertinent information. The paper also explored more general challenges as FL resource distribution, security, and privacy concerns, and offered possible answers. Additionally, the research investigated a variety of uses of FL for enhancing MEC performance, such as intrusion detection, edge caching and compute offloading, base station association, and IoV applications. To measure FL performance across wireless networks, [66] has established four basic performance metrics. The article analyzes the effects of different wireless parameters on reliability, training loss, convergence time, and energy consumption. These effects include transmission power, computing capacity, and resource allocation. They shed light on the complex interactions between the defined FL performance measures and wireless network properties.

E. EI SECURITY AND PRIVACY

Various works have recently delved into EI's problematic security and privacy challenges. These studies mainly focus on exploring innovative solutions to address the complexities arising from the heterogeneous nature of edge servers, fluctuating communication conditions, and the intricate details involved in ensuring data privacy and system integrity. The work in [49] delved into security and privacy concerns within EI, particularly in the context of 5G and beyond networks. The authors suggested employing blockchain as a security solution to address these issues. They introduced a conceptual framework consisting of six layers: physical, edge, network, cloud, database, and interface. The authors proposed additional Blockchain usages beyond transparency. For example, they suggested that Blockchain can efficiently manage resources by enforcing strict access control restrictions and assigning valid permissions. The study by [69] discussed the security requirements and associated risks of EI, including computing, caching, and intelligence. The research presented various security and privacy concerns related to Edge computing and caching, including poisoning attacks, eavesdropping, and caching location security. The authors also emphasized that EI did not guarantee complete data privacy protection, especially in the presence of honest but curious participants. Additionally, they investigated how the diverse characteristics and capabilities of edge servers impacted the overall security of an edge intelligence system. Furthermore, they explored methods to address security concerns related to heterogeneous edge servers, such as developing adaptive security mechanisms or standards.

The decentralized nature of EI presents several security challenges. Geographically distributed processing across a network of edge devices with heterogeneous processing power, storage capacity, and varying security capabilities may lead to potential vulnerabilities [49]. Furthermore, unreliable network connectivity may disrupt data transmission and increase the risk of data breaches [70]. Another challenge is how to balance the privacy of user data and the efficiency of data processing within the EI system [71]. Blockchain technology seems to be a very promising solution for solving most of the security concerns associated with EI. Blockchain technology emerges as a promising solution for addressing these security concerns in EI. Its core strengths – openness, cryptography, and decentralization – offer significant advantages in securing EI systems [50], [70]. The paper in [50] explores the potential of blockchain technology to address such security The authors explained the ways in which blockchain is able to benefit EI. One of the major advantages that can be stated of blockchain is related to immutable and tamper-proof ledger maintenance. Unlike traditional methods of data storage, blockchain uses a distributed ledger that is in front of every participant in the network. It creates a record of all transactions against it in chronological fashion, making it virtually impossible to manipulate the data. Blockchain also enables secure and

transparent resource management in the EI network. Unlike traditional centralized architectures that rely on a single authority, blockchain operates in a decentralized manner. This inherently distributed approach removes the requirement of a central authority for the management of resources and puts forward a much safer and transparent environment concerning resource sharing. Moreover, incentive mechanisms of blockchain-based systems support efficient computing power allocation and collaborative communication between edge devices. In addition, since the network is decentralized, no single entity is in charge of the network, which gives an added advantage to the security aspect by preventing unauthorized access. Finally, smart contracts – self-executing agreements based on pre-defined code – can further enhance security by ensuring the immutability of the blockchain and guaranteeing the trustworthiness of collaborative inference at the edge. While blockchain technology offers significant potential for enhancing security in EI, further research is needed to address challenges related to scalability and the development of efficient consensus mechanisms suitable for resource-constrained edge devices. By overcoming these hurdles, we can unlock the full potential of blockchain for building a secure and trustworthy future of EI [50].

F. LLM-BASED EDGE INTELLIGENCE

Given the tremendous success that LLMs have shown recently, there is a vision for LLM-based EI systems that are independent and can self-organize, self-adapt, and self-optimize to meet a variety of user needs while utilizing the power of LLMs. Recent studies have investigated and examined this vision from several angles. For instance, Lin et al. [55] investigated LLMs at the edge of future 6G-MEC infrastructures. The authors briefly summarized the industries in which multimodal LLMs are used, like healthcare, and stressed the significance of placing LLMs close to end users. The authors also pointed out the benefits of this deployment strategy while carefully pointing out any potential serious drawbacks, like high processing demands. They suggested involving several methods, such as parameter-sharing inference and parameter-efficient fine-tuning, to address such issues. Xu et al. [52] shed light on resource-efficient solutions for large foundation models, covering practical system designs and algorithmic improvements.

One of the main advantages of LLM-based EI is that it can enhance user-machine interaction. For instance, Shen et al. [72] showed how to coordinate Generative Pretrained Transformers (GPTs) to FL-based EI models. Their work showed how LLMs can potentially help coordinate user interaction and AI model task routing in edge computing environments. Specifically, user requests are first described in natural language. The requests are then passed through GPT for understanding and routing them to the corresponding edge AI models.

III. LLMS AND EI: ENGINEERING NEXT GENERATION EDGE INTELLIGENCE

The evolution of edge computing demands ever-increasing performance, adaptability, and context awareness [22]. In this pursuit, LLMs and EI emerge as promising complementary technologies [51]. LLMs excel at natural language processing and knowledge, while EI focuses on real-time data analysis and decision-making at the edge, enabling faster responses and reduced latency. Integrating these domains holds immense potential for shaping the future of edge computing, unlocking a paradigm shift characterized by personalized, intuitive, and context-aware interactions [55]. This section delves into the synergistic potential of LLMs and EI. We explore deployment strategies and learning approaches that optimize their collaborative learning processes.

A. RESOURCE CONSTRAINTS AND TECHNICAL REQUIREMENTS FOR LLM-BASED EDGE INTELLIGENCE

Edge devices refer to computing hardware that operates at the periphery of the network, close to the source of data generation [44], [45]. Examples include IoT sensors, mobile devices, gateways, and industrial controllers [46]. These devices perform data processing locally rather than relying solely on centralized cloud servers, enabling faster response times and reduced bandwidth usage [48].

1) RESOURCE CONSTRAINTS

The minimum requirements for embedding a certain level of intelligence, particularly LLMs, into edge devices depend on several factors, including computational power, memory, storage, energy efficiency, and connectivity [73], [74], [75]. These requirements ensure that the device can effectively run AI algorithms while maintaining performance and reliability [75].

- *Computational Power:* Edge devices must possess adequate computational capabilities to handle the inference tasks associated with LLMs. This typically involves using low-power microcontrollers or specialized AI accelerators, such as Google's Edge TPU [76] or NVIDIA's Jetson platforms [77]. These components are designed to perform complex computations efficiently. A minimum requirement is generally a quad-core CPU with a processing speed of at least 1 GHz, complemented by an AI accelerator to manage the computational load of running basic LLM models. For instance the work in [78] deployed various types of the MobileBERT model (large and quantized) on a rang of Raspberry Pi devices with a couple of Broadcom CPU models (BCM2837 and BCM2711) of @1.2GHz-@1.5GHz.
- *Memory:* The substantial parameter sizes of LLMs necessitate significant RAM to load and process data effectively. For instance, the LLaMA-13B model can run on a V100 GPU wich comes with a RAM of 32GB of RAM [79]. Insufficient memory can lead to latency issues and degraded performance. However, for small

LLM models, a minimum can be as low as 4 GB. For example, a quantized to 4-bits phi3-mini only occupies about 1.8GB of memory, and ca run on an iPhone 14 with A16 Bionic chip [80].

- *Storage:* Storage is critical for model deployment and data caching, ensuring that the edge device can store the model and additional data required for processing. Large models may require bigger storage capacity. for instance LLama model of 70 billion parameters, requires 150 GB of storage in FP16 format [27]. However, smaller models for Edge devices can need as lower storage requirements, as demonstrated by the work in [78], which used a storage of 32GB μSD to run a small version of the MobileBERT model.
- *Energy Efficiency:* Energy efficiency is paramount for edge devices, which often operate in power-constrained environments. Efficient power management extends operational longevity and reliability. Strategies to enhance energy efficiency include the use of low-power components, dynamic voltage and frequency scaling (DVFS) [81], and energy-efficient AI accelerators [82]. These approaches help minimize power consumption while maintaining computational performance.
- *Connectivity:* Reliable connectivity is essential for edge devices to communicate with other devices and the cloud for data exchange and updates. Ensuring continuous operation requires support for multiple connectivity options, including Wi-Fi, Bluetooth, and cellular networks, with fallback mechanisms in place to maintain communication if the primary connection fails.

2) TECHNICAL PERFORMANCE AND IMPACT ON SECURITY AND RELIABILITY

Local processing at the edge significantly reduces latency compared to cloud-based solutions, offering real-time responses crucial for applications like autonomous vehicles and industrial automation [29]. Edge devices must exhibit consistent performance under varying conditions to ensure reliability [83]. This can be achieved through robust hardware design, efficient cooling mechanisms, and redundancy features [84]. Key metrics for assessing reliability include Mean Time Between Failures (MTBF) and Mean Time to Repair (MTTR) [85], which provide insights into the operational stability and maintenance requirements of the devices. In addition, implementing LLMs on edge devices introduces unique security challenges due to their distributed nature and limited resources. Robust security measures are important to prevent the leakage of data and to ensure the integrity of the device. Critical security concerns on edge devices involve encryption, which requires additional computational resources [86]. Most encryption algorithms tend to be computationally intensive. Therefore, edge devices should be able to maintain sufficient processing power to carry out encryption tasks without influencing other performances. Thus, edge devices need to allocate sufficient processing power to handle encryption tasks without compromising overall performance. Secure boot processes and real-time threat detection systems also play crucial roles in maintaining device security [87]. Evaluating security performance involves metrics such as the number of detected security incidents, response time to breaches, and adherence to security standards.

B. EVOLUTION AND FUTURE PROSPECTS OF LLM-BASED EI

LLMs have seen a remarkable evolution in recent years, with capabilities soaring from rudimentary text generation to complex reasoning and creative outputs [54]. This progress hinges on advancements in Pre-trained Language Models (PLMs). PLMs, firstly attempted by ELMo [9], and enhanced by BERT establishing a "pre-training and fine-tuning" paradigm and enhanced by BERT, establishing a "pre-training and fine-tuning" paradigm. Thanks to the self-attention mechanism within Transformers-based architectures [8]. Self-attention allows the model to efficiently capture long-range contextual dependencies within the text, leading to superior performance in tasks relevant to EI, such as real-time data analysis and anomaly detection [88]. The resulting pre-trained word representations, imbued with contextual understanding, become powerful semantic features for various EI applications. Previous language models were primarily meant for text data modeling and creation [3], but more recent models (*i.e.,* LLMs), such as GPT-4 [15], are geared toward tackling complex issues [14]. Direct deployment of LLMs on edge devices with low resources is not feasible. This challenge is bypassed by different optimization techniques [89], including pruning, quantization, and edge caching [78], [84], [90], and these compressed models are called Small Language Models (SLMs) [80]. SLMs significantly reduce model size and memory footprint, making them ideal for deployment on edge devices with limited resources. Leading the charge in this space are companies like Microsoft with their Phi-3-mini model and Apple's on-device intelligence systems [80], [91]. This push for on-device intelligence is further bolstered by major industry players like NVIDIA [30]. This represents a significant leap forward in bringing powerful AI capabilities directly to the edge.

What can really turn EI into a game-changing opportunity is the ability of devices not only to transmit data but to understand its meaning. Traditional, Shannon information theory-based communication focuses only on the transmission of data, whereas emerging wireless technologies such as 5G and 6G have major priority given to network adaptation according to content and meaning [92]. This shift has fueled the rise of semantic communication–a new approach that would tangle the meaning of messages into the way of communicating [93], [94]. In that respect, semantic communication could dramatically reduce data transmission and bandwidth while increasing communication speed by

knowing the context and meaning of data, without any loss in terms of reliability or accuracy [93], [95], [96]. Actually, this paradigm shift means more efficient and more relevant data transmission, which in general will lead to enhanced EI system performance. Semantic communication ensures only the most crucial insights are processed and shared, boosting both efficiency and precision. Furthermore, semantic communication can incorporate multimodal signals (text, audio, image, video) to create immersive experiences with low latency and high semantic quality [96]. Recent advancements in large AI models, particularly Multimodal Large Language Models (MLLMs), offer solutions to challenges like data variety, ambiguity, and signal distortion [94], [96]. For instance, the work in [96] proposed a Large AI Model-based Multimodal Semantic Communication (LAM-MSC) framework. The MLM-based Multimodal Alignment (MMA) utilizes MLMs to convert between multimodal and single-modality data while preserving semantic consistency. The future of edge intelligence lies in leveraging such technologies synergistically, ensuring scalable solutions that meet the diverse needs of real-time applications. As we navigate this transformative landscape, the evolution and integration of LLM-based technologies are set to redefine the possibilities of edge computing, driving innovations that empower smarter, more responsive systems for tomorrow's challenges.

C. LLM-BASED EI AND XG WIRELESS COMMUNICATION NETWORKS

The next generation of xG wireless communications, encompassing 5G, 6G, and beyond, paradigm shift in connectivity, performance, and intelligence [59], [60], [97]. In this regard, LLM-based EI bring new opportunities for enhancement in network management, security, and service delivery and hence drive further evolution toward smart, adaptive communication systems. In this part we briefly discuss the advancements and requirements of these networks in relation to LLM-based EI.

1) SYNERGY BETWEEN LLMS, EI, AND 6G TECHNOLOGIES

A powerful synergy can be created by the convergence of LLMs, EI, and 6G technologies, enabling a fundamental change in the way intelligent systems operate in edge environments. This synergy capitalizes on the unique strengths of each component - LLMs for context understanding, EI for localized processing and 6G for ultra-fast communication. Benefits and contributions from each component can be outlined as follows:

- *LLMs: Contextual Understanding and Decision-Making:* LLMs are capable of interpreting complex and multi-dimensional data [98]. In edge scenarios, techniques such as model distillation and fine-tuning on domain-specific datasets can be adapted [38], to reduce the computational load of the model while retaining most of their interpretative power. In the context of the synergy, LLMs provide the contextual understanding functions [20], needed to generate insights, make informed decisions, and oversee processes, autonomously. Their ability to synthesize information from diverse sources [99], allows them to guide and refine the operations carried out by EI systems.
- *EI: Real-Time Localized Processing:* EI employs lightweight machine learning models with efficient algorithms optimized for edge devices. FL [87], [100] is one of the techniques that enables EI systems to learn from decentralized data sources without sharing the raw data, thereby enhancing both privacy and efficiency. EI acts as the first layer of processing, handling real-time data collection and analysis. It enables immediate actions to be taken right at the data source [42], which is crucial for latency-sensitive applications. By processing data locally, it minimizes the need for constant communication with centralized systems, leading to reduced bandwidth usage and faster response times [101].
- *6G and Ultra-Fast and Reliable Communications:* 6G technology expected to be delivered with various enhancements, such as extended ultra-reliable low-latency communication (xURLLC), massive machine-type communication (mMTC), and network slicing [48], [102], [103]. These features ensure that data and insights could be delivered with near-nonexistent delays, with extremely high reliability, even in an environment with large densities of connected devices. 6G's ability to support AI-driven network management also has a critical role in optimizing pathways of communication and resource allocation [60]. Its high-speed communication capabilities allow for real-time synchronization between edge devices and central systems, ensuring that the insights generated by LLMs and the real-time actions taken by EI are perfectly aligned.

This synergy can be used, for instance, in smart city transportation systems where an EI system detects a traffic accident that is causing congestion. EI processes real-time data from traffic cameras, connected vehicles, and sensors to assess the situation and identify the need for immediate action. The LLM then analyzes the broader context of the situation, such as historical traffic patterns and current road conditions, and recommends rerouting strategies to alleviate congestion. The 6G network ensures that these recommendations are communicated and implemented in real-time, allowing traffic management systems to dynamically adjust traffic signals and provide route updates to connected vehicles and drivers.

2) LEVERAGING LLMS AND EI IN 6G TERAHERTZ COMMUNICATION NETWORKS

Fast-paced development in wireless communication technologies is pushing us to the sixth generation, 6G, networks that can achieve data rates unprecedentedly high with

ultra-low latency and massive connectivity [59], [60], [102]. Terahertz communications, one of the most essential technologies supporting 6G, shall operate at a frequency range of 0.1-10 THz [104], much beyond the millimeter waves used by the 5G networks. This has the benefits of high data rates, ultra-low latency, and the capacity for supporting dense network deployments–needed for the proliferation of IoT devices. With large bandwidth availability, data rates that the THz bands can support are as high as 100 Gbps [105], providing high-speed signal propagation with minimal delay. However, these advantages come with challenges such as high propagation losses, atmospheric absorption, and the need for advanced materials and technologies for efficient signal generation and detection [104].

LLMs can be very instrumental in improving both performance and the capabilities exhibited by 6G. They can digest enormous amounts of data in real-time to predict spectrum usage patterns and hence optimize spectrum allocation, mitigating high propagation losses and atmospheric absorption issues in THz bands [106], [107]. Moreover, LLMs are able to process environmental data to be used in dynamically adjusting beamforming algorithms for efficient signal directionality that ensures reduced interference [108]. On the other hand, EI plays an important role in 6G due to its stringent latency requirements and the need for real-time data processing [61]. EI can process data at the edge, utilizing the high data rates of THz communication to provide instantaneous feedback and decision-making, which is crucial for applications like autonomous driving and remote surgery [3]. Additionally, EI algorithms can optimize the allocation of computational resources across the network, ensuring that high-bandwidth THz links are efficiently utilized [61]. EI can also deploy localized AI models that are fine-tuned for specific environments and applications, enhancing the overall network performance by reducing the need for centralized processing [109].

To effectively leverage LLMs and EI with 6G terahertz communication, several implementation strategies can be adopted. Implementing adaptive learning algorithms that can adjust to the dynamic conditions of THz communication environments ensures robust and reliable performance [109]. Additionally, investing in research and development of advanced materials and antenna technologies that can support efficient THz signal generation and detection complements the AI-driven edge network optimization efforts [110].

3) 6G'S URLLC AND MMTC FOR LLM-BASED EI

The integration of URLLC into 6G Networks presents a huge leap toward supporting real-time processing capability at the edge for LLMs [59]. URLLC is designed to meet strict requirements of applications that demand high reliability and ultra-low latency, typically on the order of 1 millisecond or less [104]. This feature is highly critical in enabling real-time interactions and decision-making processes requisite for advanced applications like autonomous driving, industrial automation, and healthcare [102], [111].

LLMs require substantial computational resources and efficient data handling to perform real-time language processing and analysis. By leveraging URLLC, edge devices equipped with LLMs can achieve the necessary low-latency communication to quickly exchange data with central servers or other edge nodes [94]. For instance, LLMs can process sensor data in real time to perceive complex driving environments and react correspondingly in an autonomous vehicle use case. URLLC ensures that the transmission of data between the vehicle's sensors and edge processing units comes with minimum delay, which will permit instant decision-making and increase safety and vehicle performance.

mMTC is another important component of 6G, designed to support the connectivity of a vast number of devices simultaneously [112]. mMTC enables the large-scale deployment of intelligent edge devices by providing robust and scalable connectivity solutions [113]. This is particularly relevant for the Internet of Things (IoT) ecosystem, where billions of devices need to be interconnected efficiently. mMTC can supports LLM-based EI large-scale deployment by offering the following capabilities:

- *High Device Density Tolerance:* Part of the mMTC is designed for high device densities so that millions of devices per square kilometer can be connected [113]. This capability is particularly relevant with respect to smart city applications, where numerous sensors, cameras, and other IoT devices need to communicate with edge servers to manage urban infrastructure efficiently [114]. For example, in an LLM-based EI smart traffic management system, it will enable real-time connectivity to traffic lights, surveillance cameras, vehicle sensors, and others, facilitating dynamic optimization of traffic flow and incident management [115].
- *Energy Efficiency:* mMTC protocols are optimized for low power consumption, which should be most important in battery-operated edge devices [115]. Energy efficiency is a guarantee toward a long life and sustainability of large-scale intelligent edge devices deployments [116]. For instance, agricultural IoT applications provide sensors able to be deployed in fields with the capabilities to monitor, through mMTC energy-efficient communication protocols, parameters like soil moisture, temperature, and crop health during extended periods without frequent battery recharging [117], [118].
- *Scalability and Flexibility:* mMTC provides flexible and efficient solutions for scalable connectivity, able to be accommodated in various network situations and requirements [113]. This is primordial in LLM-based EI automation, as the network can become relatively dynamic in terms of devices and applications supported with very heterogeneous communication requirements [114]. Scalability of the network with an increasing number of devices is seamless, and mMTC ensures constant performance and reliability [116].

- *Robustness and Reliability:* mMTC ensures strong and reliable communication even in hostile environments, such as underground or underwater deployments [113]. This robustness is particularly critical to applications in environmental monitoring, wherein edge devices could be deployed in locations that are hard to reach or extreme in climate.

4) SCALABILITY AND FLEXIBILITY OF LLM-BASED EI IN 6G HETEROGENEOUS NETWORK ENVIRONMENTS

Next-generation wireless communication systems, particularly 6G, are anticipated to encompass a highly heterogeneous network environment. This is in relation to the integration of various sorts of network types (cellular, Wi-Fi, satellite, IoT networks.. etc), and deployments (Non-terrestrial, Under-water..etc)- all of which differ much in characteristics and requirements [102]. Under such circumstances, scalability and flexibility implicit in LLM-based EI become very important for slotting in these complex networks efficiently and effectively [55], [110].

Scalability refers to the capability of a system to handle a growing amount of work or its potential to accommodate growth. In the context of 6G networks, scalability of LLM-based EI involves managing an increasing number of devices, vast data volumes, and complex services without degradation in performance [119]. LLM-based EI scalability can be provided with:

- *Distributed Computing Architectures:* One approach to achieving scalability is the deployment of distributed computing architectures [61], [66], [67]. LLMs can be distributed across multiple edge nodes, enabling parallel processing and reducing the computational burden on individual nodes [67], [120]. Given that 6G networks are expected to support distributed learning [55].
- *Hierarchical Edge Computing:* A hierarchical edge computing model can enhance scalability by organizing edge nodes in a multi-tier architecture [121]. In this model, primary edge nodes handle immediate, local processing tasks, while secondary nodes manage more complex and aggregated data processing. This tiered approach allows the network to scale efficiently by distributing processing loads according to the capabilities and proximity of edge nodes [122].
- *Federated Learning:* Federated learning has emerged as a potential for scaling LLM-based EI [36], [123], [124]. It can enable many edge devices to train a global model collaboratively without having to share the underlying raw data to assure privacy and reduce large data transmission [100]. This would be extremely useful in scenarios when such huge amounts of data are generated at the edge. Federated learning empowers edge devices to build high-quality and robust global models while remaining efficient in the local management of computational resources [39], [125].

Flexibility in LLM-based EI refers to the system's ability to adapt to varying network conditions, diverse application requirements, and evolving technological landscapes [55]. In the context of 6G, this can be provided with:

- *Adaptiveness:* Flexibility can be achieved through the deployment of adaptive AI models capable of adjusting their parameters and algorithms based on real-time network conditions [48]. A recent study by Chen et al. [126] introduced an adaptive layer splitting framework to enhance the deployment of LLMs in edge computing environments, significantly contributing to scalability and flexibility.
- *Modular Network Architectures:* Designing modular networking architectures will allow for independent upgrading or reconfiguration of different networking elements [127]. For instance, edge nodes will have modular AI accelerators that might be replaced with better ones in case of more advanced hardware emerging, keeping your network flexible and future-proof [128].
- *Cross-Layer Optimization:* Another way for flexibility is cross-layer optimization methods [129], where LLMs are utilized to optimize interactions across different layers of the network stack [107].
- *Context-Aware Computing:* context-aware computing, in which LLMs are able to enhance themselves as entities through the use of context information [55], such as user preferences, environmental conditions, and device capabilities to tune their processing and decisions.

5) AI-DRIVEN NETWORK MANAGEMENT FOR 6G NETWORKS USING LLM-BASED EI

In recent years, the convergence of AI and edge computing has enhanced network management, paving the way for innovative approaches in managing next-generation networks such as 6G [48]. LLMs integrated with EI present a compelling paradigm for enhancing AI-driven network management due to their advanced capabilities such as adaptive learning [110]. This also includes:

- *Adaptive Network Slicing and Resource Management:* 6G networks are expected to host a wide array of applications that come with very heterogeneous Quality of Service (QoS) requirements [102]. This can be further enhanced in network slicing using LLM-based EI through dynamic resource allocation based on real-time network situational awareness and sensing of user demands [48], [97]. For instance, the work in [126] utilized a model-based reinforcement learning (MBRL) approach to determine optimal splitting points across edge and user equipment (UE) for 6G networks, thereby balancing inference performance and computational load under varying network conditions.
- *Predictive Maintenance and Fault Management:* The predictive capabilities of the LLMs can be effectively used to predict network faults ahead of time, before these have any impact on service quality [109]. To that effect, LLM-based EI systems could aid in the analysis of vast amounts of operational data to identify patterns indicative of network issues and trigger preemptive

maintenance activities, such as in [130], ensuring higher network reliability with reduced downtime.

- *Intelligent Traffic Management:* 6G will need to accommodate such unparalleled data volumes and new types of traffic [102]. With contextual understanding and real-time processing capabilities, LLM-based EI may contribute much to traffic management [109]. In that respect, LLM-based EI will analyze the network traffic pattern to predict congestion points [110], thus achieving dynamic traffic routing for load balancing and enhanced network efficiency and user experience.

6) ADVANCED SECURITY FEATURES IN 6G NETWORKS AND THE ROLE OF LLM-BASED EI

The security architecture in 6G will be much more complex and advanced than in previous generations [131], with cutting-edge technologies that enable them to safeguard against the ever-evolving cyber threats [132]. Such advancement includes Intelligent Native Security [131]. LLMs-based EI can play a role in utilizing such advanced security features, enhancing threat detection driven by AI and ensuring strong protection in edge deployments [132]. This includes:

- *Quantum-Resistant Cryptography and Secure Communication:* Quantum-resistant cryptographic algorithms are envisioned to be applied in 6G networks as a countermeasure against the potential threats originating from quantum computing [71], [133]. This could be leveraged by LLM to set up a secure communication path between edge devices and the core network [134]. For instance, LLMs can for example, on the fly, select the most appropriate [135] quantum-resistant cryptographic algorithms with respect to the threat landscape and computational resources available at a time. In addition, LLMs can also manage cryptographic keys using state-of-the-art key exchange protocols like Quantum Key Distribution [136].
- *Blockchain for Secure Identity Management and Transaction Verification:* Blockchain technology will play a very important role in 6G networks, especially relating to secure identity management and transaction verification [137]. LLMs will interact with blockchain protocols for edge device and data transaction integrity and authenticity [138]. LLMs can leverage blockchain to create a decentralized and tamper-proof identity management system [139], where each edge device can have a unique digital identity recorded on the blockchain, which LLMs can verify before allowing access to network resources. In addition, LLM-based EI can be used validate transactions against blockchain records [140], ensuring that all data exchanges and transactions are legitimate and have not been tampered with.
- *AI-Driven Threat Detection and Response:* AI-driven threat detection mechanisms for future networks are essential for identifying and mitigating security threats in real-time [71]. LLM-based EI can enhance these mechanisms through sophisticated data analysis and real-time decision-making capabilities [15]. By continuously learning from new data, LLMs can adapt to evolving threats and improve detection accuracy [88]. Also, LLMs can process unstructured data sources, such as security logs, threat reports, and other intelligence data [79]. By extracting actionable insights from such sources, LLMs can update threat detection models and preemptively counter emerging threats [141].

IV. ARCHITECTURES AND APPROACHES FOR LLM-BASED EI IMPLEMENTATION

In this section we explore the recent strategies and approaches for effective LLM-based EI deploying and learning.

A. DEPLOYMENT STRATEGIES

The adoption of layered deployment designs for LLM-based EI emerges as a fundamental paradigm in the literature [72], [142], [143], [144], [145]. In this part, we explore three specific classes within this strategic framework: Edge-only, Cloud-Edge, and Cloud-Edge-End deployments. By dissecting the distinctive attributes of each class, we aim to describe the intricate orchestration required to harness the full potential of LLMs in the context of EI.

In Fig. 2, we illustrate deployment strategies inspired by various works on LLM-based EI [18], [142], [145]. In the client-edge-cloud strategy, lightweight personalized LLM models are housed in clients like smartphones. At the edge layer, more powerful models act to relay knowledge, handling tasks such as data pre-processing and caching. The most powerful LLMs, specializing in complex tasks and model training, are kept in the cloud. This strategy maximizes resource usage, providing clients with routine tasks that simplify complex work for offloading. Additionally, personalizing client models and edge caching creates more tailored and faster user experiences. In the cloud-edge deployment architecture, such as in [142], the cloud layer houses a powerful "Responsive Large Model," a large and complex LLM trained on comprehensive datasets. This model excels at handling complex tasks and can be viewed as the root of the entire system. The edge layer, designed for real-time user interaction, utilizes a "Fine-tuned Model." This smaller, less computationally expensive version of the Responsive Large Model is specifically designed for efficient operation on resource-constrained devices. It receives "Concise prompts" (simpler instructions sent by the clients) and processes them to the cloud after prompt elaboration. Communication is facilitated by an "API Gateway," which acts as a central hub for routing requests and responses between the cloud and edge devices. "API Calls" represent user interaction points, where users can submit queries or instructions to the system. For the edge-only deployment architecture, the entire system operates solely on the edge device, without relying on a cloud component. Training

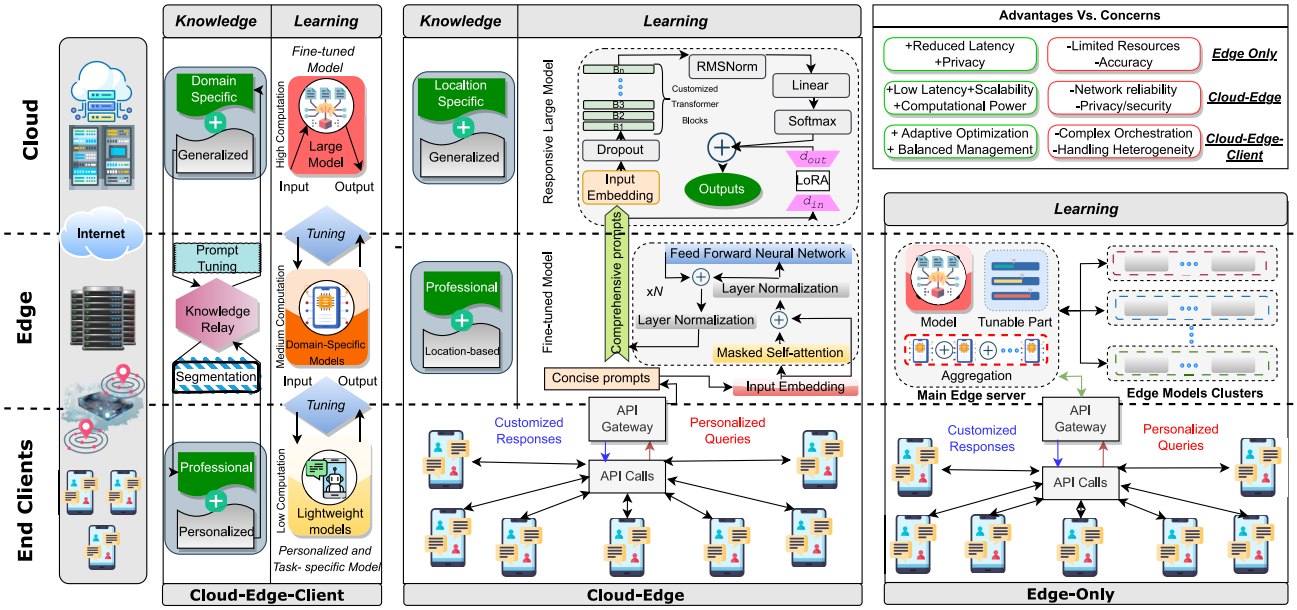

FIGURE 2. Illustration of different LLM-based EI deployment strategies.

distributed edge models can be done through techniques such as FL and KD, as demonstrated in [39], and clients can access them using API calls, similar to the previous scenario.

1) CLOUD-EDGE

In this deployment, the LLM-based EI model is distributed between the edge devices and the cloud. The model might be split, with some processing happening on the edge and more intensive tasks on the cloud. For example, the study in [143] introduced an edge-cloud LLM service. This service focuses explicitly on communicating activation tasks solely between edge and cloud instances, aiming to enhance data privacy, provide a balanced workload distribution, and achieve significant communication reduction, reaching up to 95%. In LLM-based AI, collaboration between Edge and Cloud resources has been investigated for language modeling and task-solving. One study that used this cooperative method for Automated Program Repairing (APR) is described in [144]. By offering code patches for found flaws, LLMs were shown to be quite helpful to developers. The study's suggested method used ChatGPT and Bard LLMs, limited to fewer than sixty instances of Ansible script revisions. In 70% of the sampled cases, the system produced useful results, indicating its usefulness despite the short dataset. Other works focus on optimizing such deployment scenarios. For instance, the work in [142] places smaller models at the edge and larger ones in the cloud to enhance the efficiency and customization of content generation services. In addition, edge-based LLMs had a role in smart network management and orchestration, enhancing the system's ability to understand user preferences and optimize content distribution.

- *Benefits Vs. Concerns:* The Cloud-Edge deployment combines the low latency benefits of edge processing with the scalability and computational power of the cloud. This allows for more complex models and centralized management. However, reliable network connectivity is required to circumvent uplink/downlink knowledge interruption between the cloud and the edge. In addition, some concerns are related to data privacy and security, especially when sending data to the cloud. Examples of such concerns include potential data loss, data leaks, or misuse, due to a breach or unauthorized access mechanisms [146]. And real-world examples highlight such weaknesses. For instance, the OpenAI Web Cache Deception vulnerabilities from 2023 and 2024 [147], permitting any account takeover, through one simple HTTP GET request, to store users access tokens in edge caching servers. In particular, such incidents bring up critical concerns about user data protection, especially, for sensitive details stored in chat histories. Given the substantial volume of data flowing through the cloud, it is crucial to take all precautions to prevent this information from unauthorized access.

2) CLOUD-EDGE-CLIENT

This deployment strategy involves distributing the LLM-based EI model across the cloud, edge, and end devices. It's a more comprehensive approach where the model adapts its deployment based on the specific requirements of each component. An LLM-based cloud-edge-end framework is presented in [145]. This framework is primarily meant for task inference and generative AI model optimization. As a bridge between the cloud and end terminals, the Edge model facilitates bidirectional knowledge transfer and acts as a data-free knowledge relay. The end layer functions as a data source to facilitate continuous data production

for lightweight model training and task inference. The experimental results demonstrated that pretraining greatly enhances classification accuracy for edge/end client models. Comparative analysis showed that fine-tuning based on a pre-trained model significantly outperforms direct fine-tuning on clients' data, achieving a first-epoch accuracy of 96.80%, compared to the convergence of 57.00% without pre-training. In the same architectural context, the work in [72] provided a GPT-based autonomous EI. This system aims to orchestrate edge AI models to fulfill individual user preferences and generate customized model training codes simultaneously. The primary objective is to enhance autonomous AI and personalized user experiences while maintaining data privacy in the cloud-edge-client ecosystem. The evaluation demonstrated commendable outcomes in assessing the EI Model coordination task with the GPT-3 175B IT model. Utilizing a Jetson Nano board as the client, a GeForce RTX 4090-equipped edge server, and Microsoft Azure cloud, the system achieved an accuracy rate of 84.44% along with a latency of 0.58 seconds.

- *Benefits Vs. Concerns:* The Cloud-Edge-Client deployment offers a flexible and adaptive system, optimizing the use of resources based on the processing needs. In addition, it provides a balance between latency, privacy, and computational power. Nevertheless, the complexity of orchestration introduces significant challenges. Various orchestration methods have been proposed, with the primary objective being the efficient and optimized management of resources, as highlighted by [148]. These mechanisms generally fall into three control topologies: centralized, decentralized, and distributed [149]. In the centralized topology, a single central node governs the overall orchestration and thus affords a holistic view of the dynamic and distributed infrastructure [148], [150]. The decentralized model acts like a network of communities, wherein each group has a leader (Fog Orchestrator Agent (FOA) [148]) that networks with other leaders according to predefined protocols [151]. The distributed model is like a democratic assembly where every node makes decisions so that the whole infrastructure is managed mutually [152]. It is important for Cloud-Edge-Client orchestration strategies to ensure that the deployment and operation of LLM models are managed across the cloud, edge, and client devices to guarantee that there is efficient task distribution and data flow [67]. Orchestration functionalities that can help in such situations include resource scheduling and communication management [148]. Resource scheduling involves optimizing the allocation of available resources to meet service requirements based on existing scheduling policies [153], incorporating techniques like computation offloading [154]. Communication management addresses the challenges posed by resource-constrained and heterogeneous devices within unstable networks in decentralized and distributed environments. For example, [67] introduces a Neural Pub/Sub architecture that enables subscription to inferences and the decomposition of ML operations into distributed pipelines using event-based communication. Further, orchestration mechanisms must balance latency, privacy, and computational power. For instance, a robust resource management functionality within the orchestration mechanisms can be of great assistance [155], [156]. It also has to ensure fault tolerance and reliability. A sound orchestration system must efficiently manage the occurrence of possible failures or resource limitations while ensuring model availability and reliable service delivery. For instance, the centralized topology discussed earlier may fall short if the centralized orchestrator goes down [148]. Proposed solutions for such problems include the use of replicated nodes of the centralized orchestrator [157].

3) EDGE-ONLY

The LLM-based EI model is deployed exclusively on edge devices in this deployment. Edge devices could include IoT devices, mobile devices, or other local computing resources. An example of such deployment is Confidant [18], a collaborative training framework for fine-tuning LLMs on edge mobile devices. The model is divided into multiple sub-models strategically distributed across edge devices. Piper parallel training and dynamic model partitioning techniques accelerate the training process. Confidant exhibits a noteworthy speedup of 3.84x through parallel pipeline training compared to single-phone training and a substantial memory reduction of up to 45.3%

- *Benefits Vs. Concerns:* Some advantages of the Edge-only deployment include reduced latency, as processing happens closer to the data source. Also, Privacy concerns might be addressed as data processing occurs locally. However, limited computational resources on edge devices might constrain the model size or complexity.

B. LEARNING APPROACHES

Collaborative learning strategies in the context of LLM-based EI represent approaches to distributed and decentralized AI systems. Federated and Peer-to-Peer (P2P) edge learning are two notable courses in this category. However, various frameworks for collaborative learning based on LLMs are under exploration, including Knowledge Distillation. In addition, other hybrid learning approaches are investigated, such as the work outlined in [158], where the authors employed an architecture where users can train the initial model layers while servers handle subsequent layers. This system aims to minimize network resource consumption and ensure stable LLM performance. In Fig. 3, we present an illustration of various LLM-based EI learning approaches inspired by diverse works, including [120], [125], [159].

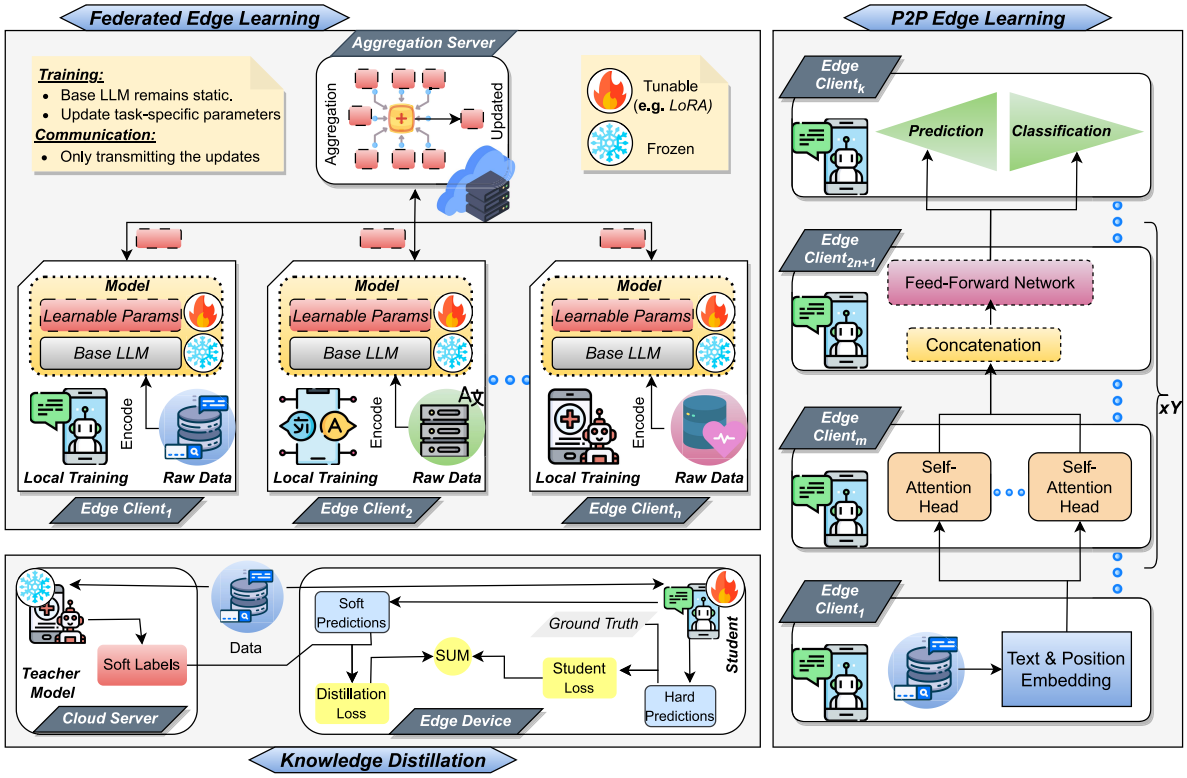

FIGURE 3. Illustration of different LLM-based EI learning approaches.

1) FEDERATED EDGE LEARNING

FL is a distributed machine learning paradigm that allows parties to collaborate on model training while protecting their privacy by not exposing their data [160]. FL is recently used to address two primary challenges associated with LLM implementations [36], [39], [123]. The first challenge involves the substantial computing resources required for training, which may be prohibitive for parties with limited computational capabilities. The second challenge pertains to the significant volume of data required for training, with an additional consideration for data privacy. FL can solve the above challenges by enabling collaborative model training without sharing extensive data and computational resources centrally [125].

FedBERT [124], one of the first FL-based language model pre-training frameworks, operates with clients updating their Embedding layers, followed by the aggregation server updating the Transformer layers. This framework allows each client the flexibility to fine-tune the model for specific tasks. Also, recent efforts, such as FedLLM [39], introduce efficient federated training methods like Parameter-Efficient Fine-Tuning (PEFT) for FL-based LLM training. FedLLM includes diverse federated LLM training methods, like FedHomoLLM for homogeneous training and FedHeteroLLM for heterogeneous configurations. Additionally, it incorporates federated offsite-tuning and federated co-tuning methods tailored for Knowledge Transfer strategies. Similarly, FedIT [123] is a privacy-preserving FL-based LLM instruction-tuning framework, leveraging local instructions from end devices as a data source. The objective is to enhance the generalizability of LLMs for specific tasks. This framework assigns a dedicated LLM to each client, utilizing local instruction datasets to update a lightweight trainable adapter incorporated into pre-trained model weights. The updates are communicated to an aggregation server, which repeats until convergence. Other studies emphasize a hardware-centric perspective when exploring FL for LLM-based EI. For instance, the work in [36] assessed existing capabilities of edge-related hardware configurations and their potential for FL involving LLMs.

- *Benefits Vs. Concerns:* FL appears promising in tackling challenges associated with centralized training; however, it introduces its issues, including data heterogeneity and clients' unreliability [161]. Also, the study in [36] identifies memory bandwidth and network connection as limiting factors for efficient training with FL on embedded hardware. Legacy FL techniques (e.g., FedAvg [160]) struggle on resource-constrained devices due to their high computational demands [162], limiting accessibility to LLMs from edge devices. Addressing this, Tian et al. [124] proposed using federated split learning, strategically splitting the model for training on separate devices, significantly reducing the computational burden and enabling broader deployment of language models on the edge.

2) P2P EDGE LEARNING

This learning paradigm involves training models on edge devices without relying on a central server or aggregator. In recent years, collaborative LLM training has noticed a shift towards decentralized strategies, fostering a more open and inclusive research ecosystem [120], [172]. Pioneering work by [173] unlocked the potential of model parallelism for large model training, paving the way for subsequent advancements. Notably, [174] and [175] addressed the challenge of heterogeneous devices with limited bandwidth, enabling billion-scale training in diverse computing environments [172].

The study in [120] examined the possibility of distributing LLMs across a vast network of decentralized edge devices, specifically consumer-level GPUs. The paper focuses on optimizing the deployment of model execution Directed Acyclic Graphs (DAGs) on devices with limited memory through sub-DAG segmentation. It enhances fault tolerance by incorporating backup nodes within computing providers and proposes strategies like partitioning and pipelines to reduce networking costs. Additionally, the paper addresses hardware performance variations through predictive models for computation and networking costs, facilitating efficient task scheduling on devices. The performance analysis of the proposed system using 50 RTX 3080 and 4 H100 GPUs underscores a substantial difference between throughputs (comparable) and prices (significant).

The focus of the study in [172] was to identify the essential components required for a robust and efficient decentralized training infrastructure. The authors emphasized three key elements: protecting privacy, spotting intentional tampering between training stages, and quickly recovering from failures. Several decentralized designs have been investigated in other works for LLM training. For example, the study in [176] offers a viewpoint of a decentralized LLM that can function inside a blockchain-based transaction system.

- *Benefits Vs. Concerns:* P2P edge learning promotes scalability, reduces latency, and enhances privacy since the raw data never leaves the edge device. It is well-suited for scenarios where real-time decision-making is crucial. However, models on edge devices need to be periodically synchronized to ensure consistency, and achieving synchronization can be challenging in dynamic or resource-constrained environments. Also, in contrast to training LLMs at cloud infrastructures equipped with powerful GPUs, embedded devices pose notable challenges for P2P edge learning, including constraints in memory, fault tolerance issues, rescheduling complexities, and low network bandwidth [120], [172].

3) KNOWLEDGE DISTILLATION

Knowledge distillation empowers transferring learned knowledge from complex models to smaller counterparts [177]. This learning paradigm is investigated for improving the generalization capability of LLMs in various tasks [38], [159], [178] (e,g, software engineering [179]), and for facilitating efficient deployment on resource-constrained platforms [180]. Multi-model knowledge distillation transcends the paradigm of single-model distillation [181], [182], extending its capability to encompass the extraction of collective wisdom from ensembles of models [54], [180], [183]. By generating compact models through distillation, this approach unlocks the potential for deployment on edge devices with limited computational resources [180], [183]. For instance, the framework in [183] leverages knowledge distillation from multiple LLMs, achieving up to 12.50% performance improvement in contrast to fine-tuning while reducing model size by up to 98.9%.

- *Benefits Vs. Concerns:* While Knowledge distillation presents an attractive avenue with low-cost inference and inferior memory usage for LLM-based EI devices, various challenges must be considered [184]. For instance, recent strategies that leverage student-generated outputs to address the training-inference mismatch while enhancing performance incur substantial computational burdens [185]. Also, API distillation, which utilizes outputs from LLM APIs to train smaller models [178], presents various concerns, particularly when the original model remains inaccessible and restrictions are placed on how predictions can be utilized [54].

V. ENHANCING EI AUTONOMY AND EFFICIENCY THROUGH LLMS

The ongoing quest to empower EI with autonomous capabilities encompasses areas like self-resilient task offloading, mobile autonomous systems, and Deep Reinforcement Learning (DRL)-based self-learning agents [42], [101]. However, navigating this trajectory toward an autonomous EI necessitates acknowledging and mitigating significant challenges. Notably, existing solutions for autonomous deployment and resource allocation within EI often rely on computationally expensive, iterative learning-based AI algorithms, potentially hindering scalability and limiting deployment efficiency [101]. Moreover, edge model selection challenges EI systems, demanding a delicate balancing act between achieving desired accuracy and maintaining a model size conducive to rapid deployment under resource constraints edge devices [101]. As EI ventures into increasingly complex and dynamic environments, the demand for robust autonomous systems and optimized performance escalates [42]. In this exigent landscape, LLMs emerge as potential solutions, offering effective capabilities [14], including in-context learning [186], which hold promise for addressing these critical challenges. This section delves into the transformative potential of LLM in empowering EI systems with these very attributes. By exploring advancements in autonomous intelligence design and optimization

TABLE 3. Selected works on autonomy designs for enhancing LLM-based EI.

Technique	Work	Year	Main Contribution	Advantages	Limitations
Self Adaptation	[163]	2023	LLM-based multi-agent system with self-governing and self-adaptation capabilities	+ Possible enhancements in decision-making throughout complex tasks.	-No concrete measurements or data are provided to assess the system's performance.
	[164]	2023	Addresses the problem of Degeneration-of-Thought with a Multi-Agent Debate scheme	+Reasoning accuracy of 36% and superiority in Commonsense Translation with 79.5%	-The study does not explicitly address the computational cost and scalability.
Action Planning	[165]	2023	Hybrid rule-based LLM-empowered autonomous vehicles planner	+ 11% reduction in unsafe driving cases +Up to 81% success with no explicit training	-The system employs a text-only input -Real-time decision-making constraints
	[166]	2023	+Tackles the cost/quality of data challenge in agent planning by few-shot planning.	+Reduced data dependency (<0.5%) +Achieving expert-level planning.	-The system employs a text-only input -Fails to handle heterogeneous data directly.
	[167]	2023	Autonomous motion planning with LLMs-based precise trajectories generation.	+Handle heterogeneous data inputs +Transparent and interpretable decision-making.	-Model inference timing uncertainty raises concerns for practical use.
	[168]	2022	Embodied agent planning with PLM-based in-context learning and real-time feedback	+Robust zero-shot performance +Errors analyze capabilities	-The system is dependent on the quality of reports generated by the Reporter module
Goal-Directed Tasks	[169]	2023	Synthetic dialogue data for zero-shot goal-directed conversation LLMs via offline RL	+Smaller models deployments +Higher success rate (30%) compared to GPT-3.5	-Manual prompt dependence hinders full task autonomy
	[170]	2023	Autonomous software development through goal-directed LLMs dialogues	+Minimal cost (under 1$) and super fast (>7min) for the entire process+	Imperfect code generation in complex projects - LLM biases influence code pattern
Self Organizing	[171]	2023	Self-healing for LLM training for efficient detection and recovery from failures.	+Up to 1.9x faster training +Streamlined failure recovery +Cost-aware scheduling	-Complex scenarios might still require manual intervention for diagnosis and resolution.
	[33]	2023	Leverages LLMs for autonomous network management through intent-driven control	+continuous cycle of learning and optimization, leading to self-organizing networks	-No concrete experiments provided to assess the system's performance

techniques, we unveil how LLMs can unlock a new era of autonomous and efficient edge applications.

A. AUTONOMOUS INTELLIGENCE DESIGNS

Autonomous Systems (AS) represents a sophisticated approach to equipping machines with complex cognitive abilities [187], that progressively build upon various intelligence levels, starting from basic reflexes and reactions to following commands, adapting to their environment, and ultimately achieving autonomous decision-making [188]. The rapid advancements in LLMs have fueled investigations on their potential to achieve *autonomous intelligence* [189], [190], [191], [192], [193], which stands for AI-enabled software solutions to eliminate human intervention [187], in scenarios traditionally requiring human-like reasoning and acting, such as driving [192].

LLMs are trained on vast datasets, allowing them to perform a remarkable variety of natural language tasks like translation, question answering, and dialogue without task-specific fine-tuning. For instance, Meta's LLaMA models leverage publicly available datasets for pre-training. This includes a substantial contribution (67%) from CommonCrawl (3.3 TB) and supplemented by datasets like C4 (783 GB), Github (328 GB), Wikipedia (83 GB), Books (85 GB), ArXiv (92 GB), and StackExchange (78 GB) [79]. PaLM 2 [194] boasts a substantially larger and more diverse pre-training corpus compared to PaLM [195]. This corpus, exceeding 780 billion tokens, encompasses Web-scraped documents, books, math expressions, code, and conversational data, reflecting the vast spectrum of natural language applications [194].

However, a fundamental limitation is that they still require human prompts and interaction to function. In addition, their logical consistency crumbles in complex reasoning, limiting their ability for in-depth and interactive evaluation, thus hindering performance in intricate, interconnected tasks [196]. This has led to growing interest in autonomous intelligence for LLMs - enabling these models to set their own goals [197], adapt themselves [163], and learn/act without human oversight [33], [198]. The promise is that autonomous LLMs could empower virtual assistants, scientific research, and other applications [72]. While reducing risks from uncontrolled model behaviors [199].

Several approaches are being explored to develop autonomous intelligence for LLM-based solutions, including EI, from various operative standpoints, as presented in Table 3. Recognizing the limitations of LLMs, researchers explore multi-agent systems where collaborative, specialized agents tackle complex tasks through "*cognitive synergy*" [163], [191], [193], [198], [199], [216]. The concept leverages the power of specialization, akin to a well-coordinated team. Each agent focuses on its unique strengths, resulting in an efficient and effective whole [191]. This diverse set of specialists resembles a team with varied expertise, tackling various challenges by utilizing different problem-solving approaches [193]. This allows the system to adapt to diverse inputs, outputs, and processing needs, enhancing its versatility and effectiveness.

EI offers a transformative approach to LLM-based autonomous intelligence [163], [191]. By enabling decentralized edge data processing (e.g., on-device inference [34], [204]) and autonomous multi-agent collaborations [163], [217], while reducing cloud-based designs challenges such as latency, privacy concerns, and computational load [55]. This facilitates cognitive synergy within multi-agent systems, leading to robust, adaptable, and efficient AI solutions capable of real-time decision-making in diverse conditions [35].

1) SELF-ADAPTATION

Building flexible and adaptable multi-agent systems is challenging yet crucial for tackling complex tasks for LLM-based autonomous EI. Self-adaptation helps these systems overcome this complexity by enabling agents to monitor and adjust themselves based on specific goals or priorities [218]. This allows agents to optimize resource usage, tolerate faults, and achieve self-optimization, making agents more suitable for dynamic environments. The authors in [163] used GPT-4 for an LLM-empowered multi-agent autonomous system, allowing agents to monitor, analyze, plan, and execute self-adjustments effectively. This integration empowered agents with enhanced communicative capabilities, allowing agents to tackle complex tasks, react intelligently to evolving situations, and operate seamlessly. The study in [164] introduced a multi-agent debate framework that forces LLMs to engage in dynamic, thought-provoking debates, fostering critical analysis and dismantling misleading concepts. The authors' experiments show that their Multi-Agent Debate (MAD) framework significantly surpasses baseline methods such as GPT-3.5-Turbo and GPT-4 in tasks such as common-sense machine translation and counter-intuitive arithmetic reasoning. For instance, using MAD, this resulted for GPT-3.5-Turbo in an 82.0 COMET score and a BLEURT score of 70.9, together with a human evaluation score of 3.78 in translation tasks and an arithmetic reasoning task accuracy of 37.0%, whereas GPT-3.5-Turbo results in an accuracy of 26.0%. Furthermore, the MAD framework decreased bias from 29.0 to 24.8 and increased diversity from 19.3 to 49.7 of the responses for the Degeneration-of-Thought problem.

2) ACTION PLANNING

Researchers are exploring the potential of employing LLMs to help in this matter. For instance, the work in [166] proposed a method for embodied agent planning that addresses limitations in data efficiency and task adaptability. It leverages few-shot learning with LLMs, enabling agents to master new tasks with minimal data. Also, the work in [165] proposed a solution for robust self-driving vehicle planning by combining the strengths of traditional rule-based approaches with the powerful common-sense reasoning capabilities of LLMs. This hybrid approach demonstrates significant potential for navigating complex driving situations and achieving superior performance compared to existing methods. The work in [167] used LLMs for detailed numerical reasoning in motion planning. The authors' experiments on the large-scale nuScenes dataset demonstrate that GPT-Driver significantly outperforms state-of-the-art methods, achieving an average L2 error of 0.84 meters and a collision rate of 0.44%, and shows superior generalization ability, achieving a 1.20 meter L2 error with only 10% of the training data compared to UniAD's [219] 1.80 meters.

The study in [168] proposed a system to improve autonomous agent reasoning through LLMs. Their system follows a three-part architecture: planning, acting, and reporting. The planning module utilizes a PLM to process information and issue instructions. The reporter module feeds information back to the PLM, forming a continuous loop where past actions and outcomes inform future decisions. This closed-loop design fosters enhanced learning and adaptation within the autonomous system. The experiments of the authors demonstrate that their Planner-Actor-Reporter system can deal with challenging reasoning tasks on embodied environments due to large-scale language models. For example, on the secret property conditional task suggesting the correct object to pick up, their Planner-Actor-Reporter system with a 70B parameter LSLM had a success rate of 96%, while the baseline 7B parameter model had only a 58% success rate.

3) GOAL-DIRECTED TASKS

LLMs and Reinforcement Learning (RL)'s unique strengths, language understanding, and goal-oriented learning can synergistically empower autonomous LLM-based EI systems with dynamic decision-making, proactive problem-solving, continuous learning, and explainable actions [197]. The work in [169] departs from the conventional paradigm of directly utilizing LLMs as goal-directed dialogue agents. Instead, it proposes a data-driven framework that leverages LLMs as generators of synthetic dialogue data. The paper introduced an "imagination engine" capable of producing task-specific, realistic, and behaviorally diverse dialogue scenarios. These simulated interactions serve as the training ground for a dedicated dialogue agent utilizing offline RL. The authors illustrate by experiments that the approach proposed strongly outperforms the traditional methods of goal-directed dialogue tasks in reinforcement learning-based training on imagined conversations. For example, in one instruction task, it achieved a user satisfaction score of 4.2 out of 5 where the baseline GPT model scored only 2.4. Their method achieved a one-shot success rate for this preference elicitation task of 44%, while the GPT baseline did it in 18% of the cases. Other than that, RL-trained agent ensured conciseness and efficiency of dialogues: it generated an average of 43 tokens per utterance compared to 118 by the baseline. These results show how much RL on synthetic data produced by LLMs has been effective for optimizing multi-turn conversational goals.

In [170], the authors introduced an approach that unifies disparate software development models via goal-driven LLM-based dialogues led by autonomous agents. This conversational workflow streamlines tasks across software development stages, achieving completion in under 7 minutes at minimal cost while proactively identifying vulnerabilities. The experiments by the authors depict that the quality of software development can significantly be improved if their model, ChatDev, is used instead of conventional techniques. ChatDev results in a completeness score of 0.5600, executability score of 0.8800, consistency score of 0.8021, and an overall quality score of 0.3953, which are all better than the respective quality scores demonstrated by GPT-Engineer, a completeness score of 0.5022, an executability

score of 0.3583, a consistency score of 0.7887, and a quality score of 0.1419, and MetaGPT, a completeness score of 0.4834, an executability score of 0.4145, a consistency score of 0.7601, and a quality score of 0.1523. In pairwise evaluations, ChatDev's solutions were preferred 77.08% of the time by GPT-4 and 90.16% by human evaluators over GPT-Engineer, and 57.08% by GPT-4 and 88.00% by human evaluators over MetaGPT. These results highlight ChatDev's superior performance in generating complete, executable, and consistent software solutions through effective multi-agent communication and collaborative problem-solving.

4) SELF-ORGANIZING

In this class, multi-agents are self-driven agents. The work in [193] describes these advanced entities actively learning and adapting their behaviors in real-time based on environmental cues, self-organizing without relying on predefined rules or fixed mechanisms. This autonomy empowers them to architect their operations. While a foundational framework might be provided, it serves as a modifiable substrate for the LLMs themselves, enabling them to not only operate within it but also evolve and self-govern by creating their own rules and structures. Examples of such efforts include Unicron [171], a self-healing manager for LLM training, and introduced domain-specific in-band error detection for real-time identification and dynamic cost-aware planning for optimal reconfiguration. The authors' experiments underline the tremendous enhancement of Unicron if failures are handled in large-scale language model training. On a 128-GPU distributed cluster, Unicron achieved up to 1.9 times higher training efficiency compared with leading methods, significantly reducing recovery costs caused by failures and boosting reliability. Under different failure scenarios, Unicron could predictably provide higher training efficiency and decrease a lot of downtime. For instance, it reduced the average transition time to 16 minutes to resume training from failures, while baseline methods took 30 minutes. Also, in [33] an LLM architecture leveraging multi-modal learning, proposes an LLM-based autonomous intelligence in telecommunications network management and control, promising optimal resource allocation, proactive traffic control, dynamic network adjustments, and real-time anomaly detection.

B. REAL-TIME PROCESSING

For applications demanding instantaneous responses, the real-time performance of LLM on resource-constrained edge devices becomes paramount. Here, we explore real-time processing techniques, including stream processing and latency reduction mechanisms. By analyzing their impact on performance, we emphasize the necessity of these techniques in enabling responsive LLM-based EI applications.

1) EDGE CACHING

LLM performance challenges like cost and speed can be mitigated by model edge caching solutions like GPTCache [84]. This open-source tool stores pre-computed responses, acting as a speedy middleman that delivers answers 2-10x faster than directly calling the LLM. GPTCache empowers developers to build cost-effective, real-time, and performant LLM applications without costly upgrades or infrastructure investments by reducing expensive API calls and ensuring stable performance. Prompt caching offers an approach to accelerate LLMs by capitalizing on inherent redundancies in user prompts. Precomputing and reusing the "attention states" of frequently occurring text segments significantly reduce processing time while maintaining accuracy [200]. Beyond pre-computed response caching, *neural caching* [201] leverages a student model trained on LLM responses. The authors have demonstrated through their experiments that the application of active learning-based policies, such as Margin Sampling and Query by Committee methods, significantly improves performance in the setup of neural caching. According to their results, former techniques have time and again turned out to outperform baseline approaches on a range of datasets and budgets. For example, on the ISEAR dataset, Margin Sampling went on to return an average online accuracy of 0.666 as compared to that achieved by the random selection baseline, which was 0.640. Similarly, Query by Committee had an online accuracy of 0.656. For RT-Polarity, Margin Sampling obtained an accuracy of 0.896 against the random baseline at 0.886. The results of the study demonstrate how such selection strategies allow for the optimization of API calls to large language models while maintaining a high model accuracy for classification tasks.

In addition, consider the case where a person has little to no connectivity, and yet, he or she expects something from a virtual assistant like Google Assistant or Siri. The edge LLM caching could further be holding pre-computed responses to frequently asked questions or common commands. This would allow the assistant to function efficiently to a limited degree without Internet connectivity [84].

2) TRANSACTIONAL STREAM PROCESSING (TSP)

is processing continuous data streams in real-time while ensuring data consistency, like reliable bank transactions. Transactional Stream Processing (TSP) can empower LLM-based EI with real-time data analysis at the device level, enabling swift decision-making, minimized latency, enhanced privacy, and optimized resource utilization. For instance, TStreamLLM's [202] Stream Processing component, acting as a proactive dispatcher, efficiently manages high-velocity data through filtering, aggregation, and compression, enabling real-time LLM updates and user inference with transactional guarantees, facilitating seamless adaptation and interaction. Also, efficient streaming introduced in StreamingLLM [203] is designed to help the model understand longer sequences without extra training. The experiments provided by the authors have shown that the proposed framework can be used to improve performance and make LLMs more lightweight for streaming applications.

For instance, they showed that StreamingLLM enabled not only the handling of Llama-2, MPT, Falcon, and Pythia with as many as 4 million tokens but without any fine-tuning at all. Introducing an attention sink and the states of key or value from the initial tokens, this framework lightens the window-attention limit and enables language modeling with stability. This provides StreamingLLM with up to 22.2× speedup over the sliding window recomputation baseline for long-text processing in streaming setups, underscoring efficiency and scalability.

3) ON-DEVICE EDGE INFERENCE OPTIMIZATION

on-device inference optimization aims to achieve enhanced privacy, real-time responsiveness, and resource efficiency by tailoring complex language models for execution directly on edge devices. This local processing approach mitigates privacy concerns surrounding data transmission, minimizes latency for immediate responses, and optimizes resource allocation. While on-device inference enhances privacy and responsiveness, their parameter sizes remain challenging for resource-constrained edge devices. To address this challenge, EdgeMoE [34] introduces an on-device inference engine that strategically partitions the model across device memory and external storage, minimizing memory footprint. The authors demonstrated that, compared to other solutions, EdgeMoE substantially improved performance and memory efficiency of mixture-of-experts-based large language models on edge devices in experiments conducted on two platforms: Jetson TX2 and Raspberry Pi 4B. Besides, EdgeMoE showed the speedup from 2.63× to 3.01× in per-token inference on Jetson TX2 and from 4.49× to 5.43× on Raspberry Pi 4B against the IO-EXP baseline. These improvements are credited mostly to expert-wise bit-width adaptive inference and in-memory expert management techniques in EdgeMoE. LLMCad [204] tackles the previous challenge collaboratively. A smaller, faster LLM generates potential text sequences, while a larger, more accurate LLM verifies and corrects them. This duo, combined with clever techniques like simultaneous validation and speculative generation, enables LLMCad to generate text up to 9.3 times faster than existing methods. PipeLLM [205] leverages diverse device capabilities in edge environments by distributing LLM components and enabling parallel execution, accelerating LLM inference in the edge.

4) 6G ULTRA-LOW LATENCY FOR LLM-BASED EI

6G aims to achieve end-to-end latency as low as 1 millisecond, which is significantly lower than the sub-10 millisecond latency targeted by 5G [104]. Such latency reduction is crucial for critical real-time applications, in which immediate data processing and response are essential. Within the context of LLMs at the edge, ultra-low latency ensures that data from sensors, user inputs and other devices at the edge can be processed almost instantaneously, facilitating rapid decision-making [156]. A reduction in latency can provide users with a seamless experience for applications such as augmented reality (AR), virtual reality (VR) and real-time translation services [220]. This is particularly important for applications requiring real-time language processing and response generation, where rapid decision-making is necessary, such as autonomous driving [192], [221]. In addition, the expected increase in 6G bandwidth (up to 100 Gbps [222]) will make it easier to process the large datasets typical of LLM operations. This enables the transmission of large volumes of data between peripheral devices and servers without bottlenecks, allowing more efficient data processing and faster decision-making based on comprehensive data analysis. Furthermore, the capability to deliver high-resolution video and audio data quickly is crucial for applications involving multimedia processing [223]. For instance, in the field of telemedicine, real-time analysis of high-resolution medical imaging data by LLMs can lead to more accurate and faster diagnoses, improving decision-making capabilities in critical health situations [224]. Advanced network slicing capabilities will be available with 6G [97], [225], allowing the creation of virtual networks tailored to the demands of specific applications. In mission-critical LLM applications such as autonomous driving or industrial automation, guaranteed QoS ensures reliable and consistent performance, which is crucial for real-time decision-making.

VI. RESOURCE OPTIMIZATION: CURRENT APPROACHES AND FUTURE OUTLOOK

Deploying LLMs on Edge devices presents unique challenges due to their inherent resource constraints, such as limited memory, processing power, and battery life. Optimizing these resources is crucial for realizing the full potential of LLM-based EI while ensuring low latency and robust performance. This part explores various recent approaches to resource optimization, examining the balancing act between model capabilities and device limitations. In addition it provides and outlook of the expected positive feed-back loop between LLM-based EI and future 6G. In Table 4 we highlight some of these efforts.

A. MEMORY MANAGEMENT

While increasingly powerful, Edge devices still struggle to accommodate the memory demands of LLMs. This mismatch severely constrains LLM performance at the edge, hindering LLM-based EI potential. vLLM [206] tackles this challenge of memory constraints in distributed LLM serving by leveraging *PagedAttention*, an attention mechanism inspired by virtual memory. This enabled extremely low memory waste, resulting in a substantial 2-4x improvement in throughput. SparQ Attention [207] decreases the memory necessities by leveraging an approximation of attention scores using a carefully selected subset of query and key vector components during every inference stage. The authors' experiments demonstrate that SparQ Attention significantly improves memory bandwidth efficiency during inference without compromising accuracy. Evaluated on models such as Llama 2, Mistral, and Pythia across various tasks, SparQ

TABLE 4. Selected works on optimization strategies for LLM-based EI efficiency.

	Technique	Work	Year	Main Contribution	Advantages	Limitations
Real-time Processing	Edge Caching	[84]	2023	Semantic caching for LLM responses (storage and retrieval) through embedding models	+Low latency +Reduce expenses by minimizing costly LLM API calls	-Insufficient granularity in recognizing true cache hits versus false positives.
		[200]	2023	LLM inference speedup leveraging prompt-specific attention state caching	+Accelerates LLM inference with latency reduction (8x-60x)+ Preserving accuracy	-Memory overhead scales with the number of cached tokens
		[201]	2023	Introduces neural caching for leveraging smaller student models with active learning	+Reduces expensive API calls, compared to embedding-based techniques (e.g., [84])	-Student model performance affected by noisy LLM labels in complex neural caching.
	Transactional Stream Processing	[202]	2023	LLM management with Transactional Stream Processing (TSP) integration for real-time updates and concurrent usage	+Optimizes multi-device LLM access for efficient concurrent processing	-Lacks experimental validation, leaving open questions about real-world performance.
		[203]	2023	Enabling LLM streaming to up to 4 million tokens without pre-training	+ Delivers up to 22.2x (per token) faster performance for streaming LLMs	- Might still necessitate more memory than available on some edge devices
	Edge Inference	[34]	2023	LLM inference optimized for edge via tiered storage and dynamic weight loading	+Achieves 2x faster inference on Raspberry Pi 4B + Accuracy loss of $\leq 5\%$	-Reducing expert numbers to fit memory constraints can compromise effectiveness.
		[204]	2023	Collaborative LLM architecture for on-edge device inference with autonomous fallback	+Achieves up to 9.3x faster token generation (e.g., 0.86s/token on Xiaomi 11)	-The "generate-then-verify" approach might negate some of the efficiency gains.
		[205]	2023	Accelerates on-edge devices inference by workload distribution and slicing	+Speeds up on-edge device inference with up to 2.86x on heterogeneous settings.	-Trade-off between speed and accuracy remains unclear compared to [34], [204]
Resource Optimization	Memory Management	[206]	2023	OS-inspired memory management attention mechanism for LLM serving	+ Up to 4x throughput augmentation with negligible memory overhead in LLM serving	-Small block swapping slows LLM serving due to high CPU-GPU data transfer load.
		[207]	2023	Memory-optimized LLM inference via selective key fetching-based attention	+Up to 8x reduction in memory bandwidth requirements + Maintains accuracy	-Incurs minimal memory capacity savings due to in-memory cached values.
		[208]	2024	Hardware-aware faster inference through memory-optimized data transfer and access	+ Running models x2 the size of the DRAM + Up to 25x inference and 9.32x latency	-Rigorous power/thermal analysis is essential for on-device LLM deployment.
	Model Compression	[41]	2023	LLM pruning leveraging input activations and weight magnitudes	+Inference speedup of 1.24× +Pruning speed gain of up to 99.59% compared to [209]	-Introduced trade-off between speed and accuracy for OPT models [210]
		[78]	2023	FlatBuffer-enabled MobileBERT for efficient resource-constrained device deployment.	+Achieves a 160x reduction in model size for efficient on-edge device deployment	-Introduces a trade-off with a 4.1% accuracy drop compared to [13]
		[23]	2023	Approximate second-order information-based one-shot weight quantization method	+Up to 4.5x inference speedup on cost-effective GPUs +Quantize GPT (175b) in 4H	+No speedup of the actual multiplication operation +No activation quantization
		[211]	2023	Activation-guided quantization with token pruning and hardware-aware optimization	+Edge speedup up to 2.6x with overall acceleration ratio of up to 2.55x +low latency	-Missing accuracy analysis limit understanding of performance trade-offs, such in [78]
		[24]	2023	Identifies and protects the critical model weights, enabling efficient compression.	+ up to 3.3× inference speedup + Enable a 70B model to run on a mobile GPU.	-Focus on traditional accuracy metrics only -Focuses on low-bit integer quantization
	Computation Offloading	[35]	2023	LLM-powered offloading using embeddings and active learning with expert feedback	+Inference takes 0.06s vs. 1.3s for traditional methods [212], offering a 21.67x speedup.	-Edge deployment limited by LLM size, processing power, and latency requirements
		[156]	2023	Rewardless active inference for LLM offloading and resource allocation	+Task completion rate of nearly 99% +Outperforms mainstream DRLs [213], [214]	-High training resource demands limit edge device deployment and applicability
	Bandwidth Efficiency	[142]	2023	Strategically edge-positioned LLMs for AI-driven, self-optimizing network management	+Predict user preferences for proactive resource allocation, outperforming LSTM	-Limited experimental validation and comprehensive evaluation, compared to [215]
		[215]	2023	Cost-efficient network architecture for collaborative edge-supported LLM training	+Reduces LLM training cost by up to 75% +Integrates mechanisms to handle failures	-Trade-off between scalability and performance gains, due to complex management

Attention achieved up to 8× savings in attention data transfers. The work in [208] presented an approach that leverages flash memory. Their method strategically stores LLM parameters in a flash and selectively transfers them to DRAM only when needed. By employing windowing and row-column bundling techniques optimized for flash memory characteristics, the technique resulted in a remarkable boost (20-25x) in inference speed. The work in [40] proposed MemGPT, a virtual context management system inspired by hierarchical memory systems in traditional operating systems. MemGPT manages different storage tiers to effectively extend the context window of LLMs. This system was evaluated in two domains: document analysis, where it successfully processed large documents beyond the LLM's native context window, and multi-session chat, where it enabled conversational agents to remember, reflect, and adapt over long-term interactions. This approach significantly enhances LLM performance in scenarios requiring extended contextual understanding.

B. MODEL COMPRESSION

Compression techniques offer promising solutions to address the need for efficient real-time deployment of LLM-based EI. Methods like pruning and quantization aim to reduce model size and computational demands. Pruning strategically removes redundant connections or neurons, while quantization lowers data precision [90]. Both approaches can be used independently or combined, with trade-offs depending on the specific application and the desired balance between size, speed, and accuracy. Also, enhancements to such techniques have been introduced for LLMs. For instance, a pruning method, termed Wanda, delivered efficient compression without retraining or costly weight updates is presented in [41]. The study employs a per-output analysis that ensures only truly unimportant connections are pruned, leading to a more efficient and accurate compressed model. Wanda gave results better than all traditional magnitude pruning methods and remained very competitive with SparseGPT on evaluation across a number of benchmarks with the LLaMA

and LLaMA-2 models. For example, Wanda yielded an average zero-shot accuracy of 54.21% at 50% unstructured sparsity, compared to 46.94% for magnitude pruning and 54.94% for SparseGPT on LLaMA-7B. On the other hand, with respect to perplexity, this is the result on the WikiText validation set Wanda obtained for LLaMA-7B: 7.26, far better than what magnitude pruning achieved, which turned out to be 17.29. This proves just how significantly effective Wanda is at preserving performance while merely reducing computational complexity by massive amounts. The work in [78] leveraged the optimized FlatBuffer format to shrink massive models by a staggering 160x, making them agile enough for real-time edge device execution. Remarkably, this dramatic size reduction comes at a minimal cost: a mere 4.1% accuracy. Another framework in [211] tackles this problem using both techniques, introducing an activation-aware token pruning with an activation-guided quantization strategy. This approach achieves a remarkable 2.55x speedup in LLM inference on edge devices while maintaining accuracy. *Extreme quantization* has been proposed by recent works, enabling 2-bit or ternary levels [23], [24], [28]. Extensive analysis conducted by [229], encompassing over 35,000 experiments, demonstrated that 4-bit precision is generally the optimal choice for balancing model size and zero-shot accuracy in LLMs across various architectures and parameter scales. Recent works, such as [32], introduced 1.58-bit LLM that demonstrates performance parity with full-precision models while significantly reducing latency, memory footprint, throughput requirements, and energy costs. The authors of this paper prove through experiments that their single-bit large language model variant, BitNetb1.58, significantly reduces memory and computational costs while remaining at par with full-precision model performance. For example, on the WikiText2 dataset, BitNets b1.58 reached a perplexity of 9.91 for model size 3B. Contrasted to that, however, the memory usage on this setting was 3.55 times smaller on GPU; similarly, inference latency was 2.71 times faster compared to LLaMA 3B. In addition, major industry players, such as NVIDIA, are also actively supporting the development and implementation of these methods. For instance, the NVIDIA RTX A6000 GPU (48 GB of VRAM) enables 4 bits-quantized LLaMA2 70B at 14 tokens/s [30].

The push for on-device intelligence in smartphones is driving the development of quantized language models (QLMs). These models, like Microsoft's Phi-3-mini [80], offer the benefits of powerful AI capabilities while remaining compact and efficient for mobile devices. Apple's on-device intelligence system announced at WWDC 2024 [91], exemplifies this trend. By incorporating quantized models, these systems can deliver features like personalized voice assistants, on-device language translation, and smart text summarization – all without relying solely on cloud processing [78].

As research on quantization techniques remains at the forefront of enabling LLMs at the edge for empowering EI, we present a comparison between Post-Training Quantization (PTQ) and Quantization-Aware Training (QAT) for LLM-based applications in Table 5. While PTQ techniques have garnered significant research attention, QAT for LLMs has also shown promising initial results, as exemplified by the LLM-QAT approach proposed in [228].

C. COMPUTATION OFFLOADING

MEC leverages computation offloading to empower resource-constrained edge devices by delegating demanding tasks to powerful cloud or server resources [232]. However, traditional offloading architectures stumble upon limitations due to the heterogeneous nature of devices, limited environmental information, unpredictable task performance, and computational inefficiencies [233]. LLM-based offloading presents a possible solution that harnesses the strengths of LLMs to navigate these complexities. The work in [35] employed learnable vector representations, an asymmetric encoder-decoder architecture, and active learning to enable intelligent and adaptable offloading decisions in dynamic environments. Existing Deep RL offloading solutions, while helpful for the matter, suffer from data inefficiency, latency insensitivity, and inflexibility to workload shifts. The work in [156] tackles this by proposing a rewardless guidance algorithm embedded in active inference. This approach steers offloading decisions and resource allocation for LLM inference in cloud-edge networks. Real-world data from a GPTJ-6B LLM not only outperformed DRLs-based solutions but also improved data utilization efficiency and adapted effectively to dynamic task loads. The authors' experiments show that the proposed LAMBO framework significantly improves offloading decision-making and resource allocation in MEC systems. In simulations with 4 MEC servers and 50 user equipments (UEs), the large-scale asymmetric encoder-decoder (AED L) model in the LAMBO framework achieved the lowest task latency and energy consumption compared to other methods. Specifically, AED L reduced the total task latency to approximately 0.06 seconds and energy consumption substantially, outperforming traditional methods like DE, DNN, and medium-scale AED models. These results demonstrate the LAMBO framework's effectiveness in optimizing MEC tasks, leveraging reinforcement learning and active learning to adapt to dynamic environments efficiently.

D. BANDWIDTH-EFFICIENT COMMUNICATION

Large Generative models are envisioned to transform future networks, particularly within self-evolving network architectures [234]. These powerful models, trained on diverse network data, can be flexibly adapted to perform various tasks, eliminating the need for dedicated AI models for each scenario [235]. For example, by embodying an AI-native architecture that optimizes communication links based on performance [142]. The work in [215] challenged the established any-to-any network paradigm for LLM training. These authors' experiments prove that their proposed

TABLE 5. Comparison between PTQ and QAT for LLM-based EI.

	Post-Training Quantization (PTQ)	Quantization-Aware Training (QAT)
Description	Converts pre-trained LLM model weights and activations to a lower precision format after training.	Trains an LLM model with the quantization scheme, allowing the model to adapt to the lower precision representation during training [226].
Advantages	-Broad applicability: Applicable to various pre-trained LLM models [23], [24]. -Simpler implementation: minimal modification to existing training pipeline [227]. -Faster deployment: due to pre-quantized model availability (GPTQ [23] quantize GPT-175b in approximately 4 GPU hours), and support from major hardware manufacturers	-Higher accuracy: Achieves better accuracy for the same bit-width compared to PTQ [227]. -Improved robustness: This leads to a more robust quantized model because it relies on backpropagation or regression, while there is no retraining for PTQ [24]. - Higher quantization levels with negligible performance degradation [228].
Limitations	-Potential accuracy degradation: May require fine-tuning to recover accuracy, especially at lower bit-widths [24]. -Accuracy-efficiency trade-off: May not achieve the optimal balance between model size and accuracy at all bit-widths [229]. -May not be suitable for all models: e.g., models with complex architectures or dynamic quantization needs [24].	-Increased complexity: Requires modifying the training process, introducing additional complexity and computational cost [227]. -Potential need for fine-tuning: May still require fine-tuning after quantization for optimal performance. -Expensive training and/or fine-tuning, especially for larger models (e.g., LLMs) compared to PTQ. -Many LLMs are trained in multiple phases. This complex process would be hard to do with QAT [228].
Applications	-Suitable for latency-critical applications where model size and deployment speed are significant concerns. -Can be used for deploying LLMs on edge devices with limited storage and processing power (e.g., AWQ [24] allow deploying 13-B LLM at 30 tokens/s on an 8GB GPU).	- Ideal for high-performance models deployments where accuracy and efficiency are paramount (e.g., time-sensitive edge-based healthcare application). - Useful for optimizing the efficiency of performance-critical LLMs, especially for real-time applications [230]. - Suitable for multi-context/long inputs applications [228].
Recent Trends	-Extreme Compression: by exploring the trade-off between achieving highly compressed quantized LLM representations using minimal bit-widths and maintaining their ability to perform sophisticated linguistic reasoning tasks [27], [28], [231]	-Data-Free distillation: given the hard problem of using QAT for LLMs, the work in [228] proposed using KD and KV cache quantization as enabler for quantizing LLMs in a QAT fashion. By achieving a 4-bit LLM QAT, this work is paving the way toward lightweight, fast, and accurate LLMs at the edge.
Memory Footprint Reduction	-Achieve significant memory footprint reduction compared to full-precision models (e.g., [28], achieved a memory footprint reduction by up to 8x).	- Generally leads to larger memory footprint reduction than PTQ due to potentially higher achievable compression ratios [228].

network architecture reduces costs substantially without affecting performance. By analyzing the traffic patterns of training dense LLMs like MegatronLM, they show that their architecture reduces the network cost by 37% to 75% compared to traditional Clos networks.

1) 6G ENERGY-EFFICIENT PROTOCOLS AND SUSTAINABLE DESIGN PRINCIPLES

6G technology promises a wealth of energy-aware protocols and sustainable design principles that can be exploited to improve the energy efficiency and sustainability of LLM-based EI.

For instance, 6G networks are engineered to use dynamic spectrum sharing techniques [103], which optimizes the use of available spectrum and reduces energy wastage [236]. This allows more cost-effective communication between peripheral devices and central servers, minimizing the energy spent on data transmission [237]. Among the quantitative performance metrics that can be used to benchmark the efficiency of the proposed solutions for dynamic spectrum sharing are: spectral efficiency in (bits/s/Hz), as used in [238], [239], which measures the effectiveness of spectrum utilization by quantifying the amount of data transmitted over a given bandwidth. The interference level in (dB), as used in [239], [240], which evaluates how well the system mitigates interference between users sharing the spectrum. Besides, fairness index, used in [239], [241], it measures how much resources are being equitably distributed to the users.

The fairness index nears 1 is an indication of fair resource allocation. Also, latency in ms, as used in [239], refers to the delay introduced by spectrum allocation processes.

In addition, 6G features adaptive transmission power control [242], which adjusts power levels according to the quality of the communication channel. This means that by transmitting data at optimum power levels, energy consumption is minimized, particularly for LLM operations requiring frequent data exchanges. Examples of quantitative performance measures for assessing the effectiveness of proposed solutions in terms of adaptive transmission power control include: power efficiency (W/bit) [243], which represents the ratio of the transmission power used to the amount of data transmitted. And, the energy consumption per Bit [243].

Furthermore, 6G will introduce advanced sleep mechanisms [244] that allow devices to enter low-power states when they are not actively processing data. One of the most important performance metric used is the energy savings (%), which represents the percentage reduction in energy consumption due to sleep mechanisms compared to continuous operation [244]. For LLM, this means that edge devices can conserve power during periods of inactivity and only wake up when required to perform calculations or transmit data.

Moreover, 6G networks rely on intelligent resource orchestration techniques that allocate computing and communication resources strictly according to energy efficiency criteria [245]. Such techniques can be assessed based on the results delivered according to various metrics, including, Resource Use Efficiency (RUE), as used in [246], representing the ratio of utilized resources to available resources alongside QoS, as used in [247]. This guarantees that LLM tasks are allocated to the most energy-efficient nodes, optimizing overall energy use. Another advantage is the ability to dynamically allocate resources according to real-time needs [248]. With 6G, edge nodes can better handle workloads, guaranteeing that LLMs receive the computing power and data throughput they need when they need it, optimizing overall performance and accelerating decision-making processes.

2) LEVERAGING 6G TO ENHANCE TRAINING AND DEPLOYMENT OF COMPLEX LLMS AT THE EDGE

6G's higher data rates will enable fast transmission of huge datasets for training complex LLMs. For instance, complex LLMs like GPT-4 [15] or Llama [249] can have a volume of tens of terabytes of training data. Therefore, it would allow quick transfer of these datasets between central servers and distributed edge devices, hence reducing latency for data synchronization and preprocessing [250]. This becomes instrumental in real-time updating and optimization of models, putting them in a better shape, hence improving their performance where necessary and adaptability to dynamic environments [36]. Low latency is certainly one of the defining features that characterize 6G, and it will become very important in applications that require real-time processing and decision-making [102]. For LLMs, this ultra-low latency means that inference can be done nearly instantaneously [55]. This is important in applications such as autonomous driving [62], where split-second decisions are crucial for safety [192]. Furthermore, reduced latency improves user experience in interactive applications like virtual assistants [251] or AR [252], which thrive on immediate context-aware responses from LLMs. Edge caching and computation offloading will become more efficient and dynamic with 6G [233]. LLMs, which require substantial computational resources [15], [249], can offload intensive tasks to nearby edge servers [156], balancing the computational load and enhancing the energy efficiency of edge devices [154]. Real-time analytics [253], will optimize offloading strategies, ensuring that edge devices handle tasks within their capabilities while heavier computations are processed by more powerful edge servers or the cloud. This adaptive model deployment and resource-aware offloading will be critical for maintaining high performance and energy efficiency in various edge applications. FL, powered by the advanced infrastructure of 6G [102], will support hierarchical learning frameworks whereby local clusters of edge devices train models in isolation but periodically aggregate at higher-tier edge servers [254]. This allows for a hierarchical approach to guarantee scalable and efficient training processes by utilizing the high data rates and low latency offered by 6G for frequent and secure model updates.

The transformative potential of 6G in supporting LLM-based EI can be illustrated through various innovative applications. In autonomous vehicles, real-time processing of sensor data and environment mapping using LLMs, enabled by 6G's high data rates and low latency, will significantly enhance navigation and safety [192]. In smart healthcare, personalized medicine and real-time patient monitoring can be achieved with edge-based LLMs that analyze patient data locally and share insights globally through secure 6G connections, enabling rapid and accurate medical interventions [224], [255]. AR applications will benefit from seamless integration with LLMs [220], providing real-time language translation and contextual information delivery, powered by the low latency and high throughput of 6G networks. Industrial automation will be enhanced by deploying LLMs for predictive maintenance and process optimization in smart factories [130], leveraging 6G's enhanced connectivity and edge computing capabilities to maintain high efficiency and uptime. Furthermore, disaster response efforts can utilize LLMs for real-time analysis of satellite imagery and sensor data, enabled by 6G's reliable connectivity and fast data transmission, to coordinate timely and effective response strategies.

3) OPTIMIZING LLM-BASED EDGE INTELLIGENCE FOR EMERGING 6G APPLICATIONS AND TECHNOLOGIES

The integration of LLM with EI in the context of upcoming 6G developments such as holographic [256] and semantic

communications [92], [94] presents unique opportunities and challenges [95], [256]. Unprecedented capabilities of 6G networks, including ultra-high data rates, ultra-low latency and enhanced connectivity, create an environment in which LLM-based EI can thrive, enabling such sophisticated applications that were previously unattainable. The goals for 6G holographic communication realized are immersive, three-dimensional experiences in real-time that require extremely low latency and high bandwidth [257]. The huge challenge lies in how much data is transmitted in holographic data with high fidelity [258]. LLM-based EI can therefore optimize these through edge caching and smart data compression techniques. For instance, running LLMs at the edge can enable predictive analytics [259] that project demand by a user and prefetch holographic data segments to reduce latency. Moreover, LLMs can be trained for contextual understanding [260] to compress holographic data without any loss in quality, hence further reducing the bandwidth requirement. Semantic communication, another cornerstone of 6G [92], [95], focuses on transmitting the meaning rather than the raw data [93]. This shift reduces the amount of data transmitted by focusing on the essential information required for understanding. LLM-based EI can enhance semantic communication by employing advanced NLP techniques to interpret and summarize data contextually [96]. For example, in a smart healthcare scenario, instead of transmitting raw sensor data, the edge device can use LLMs to interpret patient vitals and communicate critical health insights to medical professionals. This approach not only reduces data volume but also ensures that the transmitted information is highly relevant and actionable. Despite these advances, there are still a number of challenges to overcome. One such challenge concerns is the computational complexity and power consumption associated therewith, which can be very high for resource-constrained peripheral devices. To overcome this, innovative approaches such as model pruning, quantization and the development of lightweight LLM architectures suitable for edge deployment are required.

These developments are not without their own challenges, however. One of these is the computational complexity and power consumption of LLMs, potentially making them cost-prohibitive for resource-constrained peripheral devices. To remedy this, it is necessary to adopt innovative approaches such as model pruning, quantization and the development of lightweight LLM architectures suitable for edge deployment, without big accuracy loss [27], [227]. Furthermore, due to the dynamic and heterogeneous nature of edge environments, there is a need for robust and adaptable LLMs, capable of learning and adapting in real time [135], [261]. This might involve the integration of reinforcement learning algorithms that enable LLMs to dynamically optimize their performance based on real-time feedback from the environment [108], [218]. One example is a peripheral-based intelligent transport system in which LLMs continuously adapt to changing traffic and environmental conditions to optimize route planning and traffic management [262], [263].

VII. APPLICATION DOMAINS OF LLM-BASED EI

Despite LLMs' impact in natural language processing, their immense computational requirements often pose deployment limitations on cloud-based systems, hindering their potential for real-time, localized applications. Here, the field of EI offers a transformational solution. Integrating LLMs with resource-constrained devices at the network edge unlocks promising directions for decentralized, context-aware AI applications across diverse domains. This section delves into the captivating landscape of practical applications enabled by LLM-based edge intelligence. We shift the focus from theoretical advancements to tangible use cases, providing a critical analysis of how LLMs are being harnessed to empower intelligent devices at the edge. Our exploration encompasses a range of industries and scenarios, from autonomous driving to decentralized healthcare systems delivering real-time diagnostics, as illustrated in Fig. 4. Through these concrete examples, we aim to illuminate LLM-based EI's significant impact and practical value, highlighting its potential to reshape various aspects of our lives. As shown in Table 6, LLMs combined with EI offer significant advantages, pushing the boundaries of traditional AI approaches. We explore how this integration can enhance AI solutions in Autonomous Driving, Future Networks, Healthcare, and Software Engineering.

A. AUTONOMOUS DRIVING

The incorporation of AI into autonomous vehicles offered significant advancements. Notably, AI facilitates enhanced safety by enabling precise object detection and real-time decision-making, ultimately reducing traffic accidents [264]. Additionally, AI empowers autonomous vehicles to adapt to various driving conditions, optimize path planning and motion control, and anticipate traffic movements [265]. These capabilities collectively improve traffic flow, reduce congestion, and create a more sustainable transportation ecosystem.

While significant resources have been directed toward autonomous vehicle development, large-scale commercialization remains languid. This is attributed to various challenges, including safety concerns in handling unforeseen scenarios and the absence of social intelligence in current autonomous decision-making systems [199], [266]. The argument for the critical role of proactive reasoning in autonomous systems is bolstered by considering the limitations of purely reactive approaches, mainly represented by rule-based and learning-based approaches [267]. This reactive approach can lead to delayed or inappropriate actions in edge cases, potentially resulting in accidents or system failures [199]. For instance, imagine an autonomous vehicle

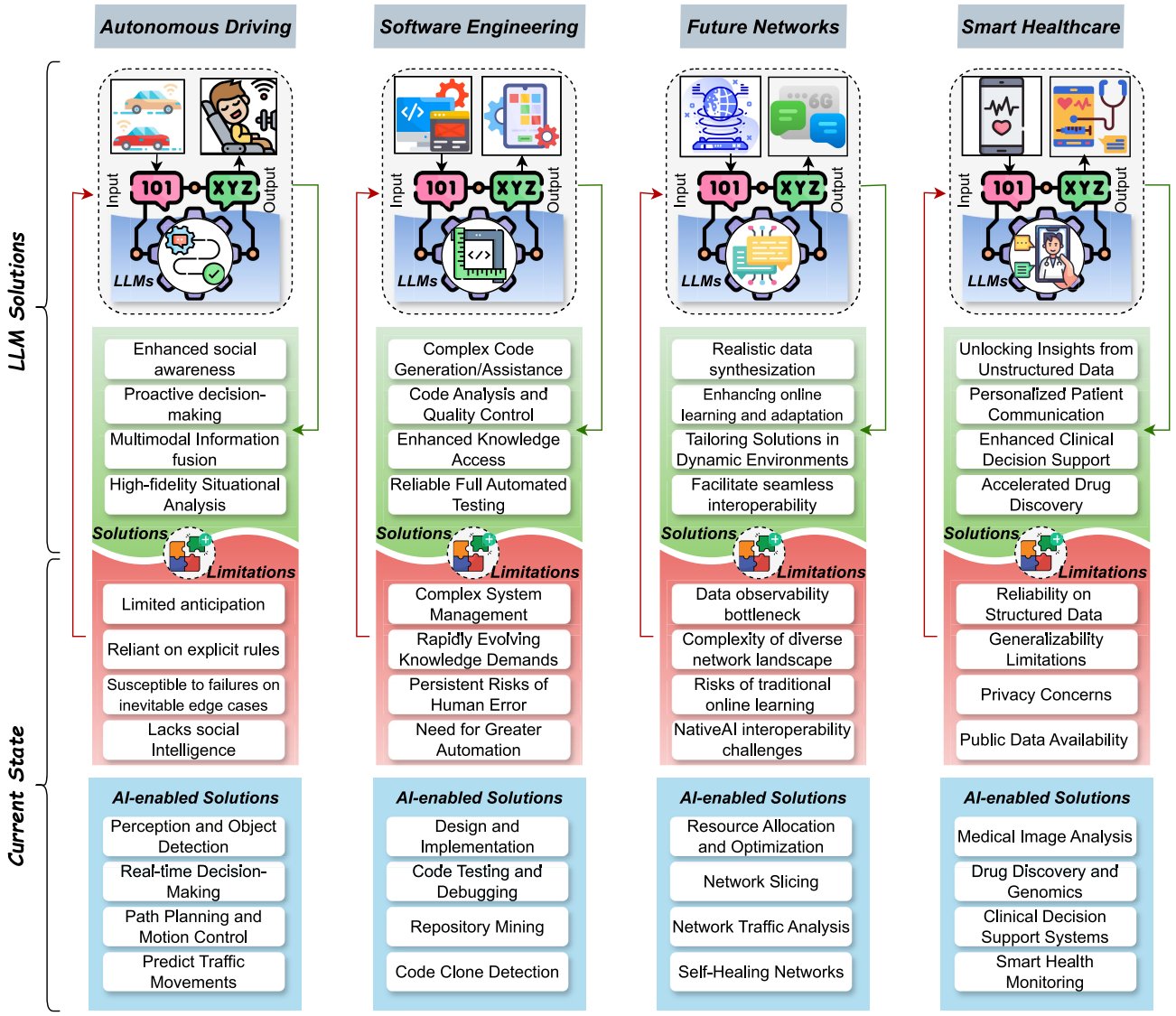

FIGURE 4. LLM-based EI to address limitations in different practical application domains.

encountering a sudden downpour significantly reducing visibility. A purely reactive system might not adjust its speed or braking behavior promptly due to the unexpected nature of the event.[1] However, a system equipped with proactive reasoning capabilities could anticipate the increased stopping distance on wet roads and take preventative measures to ensure safety [266]. This highlights the necessity of moving beyond reactive systems and embracing proactive approaches that enable autonomous systems to not only react to their environment but also reason, anticipate, and adapt – qualities that are crucial for navigating the complexities of the real world.

In addition, the current limitations of autonomous vehicles extend beyond handling edge cases and encompass their inability to comprehend the social context of driving [199], [268]. Most decision-making systems view driving solely from a mechanical perspective, prioritizing strict adherence to traffic regulations and treating interactions with the environment as purely physical maneuvers [266]. This approach disregards the crucial role of social cues and intentions in governing safe and predictable driving behavior. However, effective communication and understanding of intentions (or social intelligence [269]) are paramount for successful interaction on the road. Imagine an autonomous vehicle approaching a roundabout. A strictly rule-based system might prioritize its right of way, potentially causing a collision with a driver who employs a courteous "yield-to-the-right" strategy in congested situations. However, with the ability to understand and adapt to social norms, the autonomous vehicle could recognize the unwritten rule and yield, leading to a smoother flow of traffic and

[1] https://www.tesla.com/ownersmanual/model3/en_us/GUID-E5FF5E84-6AAC-43E6-B7ED-EC1E9AEB17B7.html

TABLE 6. Applications that can leverage LLM-based EI alongside AI.

Application	AI based Cloud Solutions	Limitations	LLM based Solutions	Benefits of EI	SW/HW and Connectivity Requirements
Autonomous Driving	- Computer Vision (e.g., Object detection, lane recognition, localization, Traffic sign detection, ..etc. [276]) - Deep Learning (e.g., Traffic prediction, Path planning, Scene perception .. etc [264])	- Relies on large, labeled datasets for specific scenarios [277]. - Struggles with unexpected situations or complex environments [192], [273]. - High latency due to cloud communication [273].	- Leverage LLMs on the vehicle itself to analyze unstructured data like traffic reports, weather forecasts, and sensor data in real-time for adaptation [192], [273]. - Generate natural language descriptions of surroundings for improved on-board decision-making with lower latency [266], [274]. - Enable proactive route adjustments based on real-time traffic conditions and potential hazards [192], [262].	- Faster response times to critical events [34]. - Improved safety through on-device hazard prediction [29]. - Continued operation even with limited or no connectivity [27].	-SW [192]: Autoware.AI [278], Tesla Autopilot [279], Baidu Apollo [280]..etc -HW [221]: e.g A100 GPU, NVIDIA RTX 3090 Ti -Connectivity: Ultra-low latency (1 ms) [281] -Reliability: Redundant communication links (e.g., C-V2X [282]).
Software Engineering	- Design (e.g., source code representation), implementation (e.g., programming), debugging (e.g., bug localization), maintenance (e.g., repair), and management (e.g., repository mining) [283], [284]	- Difficulty in understanding complex code structures and identifying subtle bugs [285]. - Generated code may lack efficiency or readability [286]. - Cloud-based analysis can be slow for large codebases.	- Integrate LLMs directly into development environments to analyze code for logical errors [287], potential security vulnerabilities [288], [289], and code style consistency by grasping the program's intent in real-time [290]. - Generate natural language documentation that aligns with the code's functionality on the developer's machine [291]. - Suggest code improvements and optimizations based on best practices and real-time analysis [292]. - Faster and autonomous code generation and testing cycles [293], [294].	- Maintain code privacy, while enhance latency [32]. - Increased developer productivity with real-time code analysis and suggestions.	-SW [170]: (Jupyter, Visual Studio Code with AI plugins, GitHub Copilot, CI/CD tools like Jenkins, Docker.) -HW: CPU/GPU (Intel i9 or AMD Ryzen 9 processors with NVIDIA GPUs (e.g., RTX 3080)), RAM (Minimum 16 GB DDR4) -Connectivity: Latency: Low latency (<20 ms)
Future Networks	- Deep Learning (e.g., Traffic prediction, traffic classification, traffic routing [87], [295]) - Deep Reinforcement Learning (e.g., Resource allocation, ..etc [295])	- Limited ability to handle unforeseen network demands [296]. - Difficulty in optimizing for multiple network objectives simultaneously [295]. - High latency for decision-making in cloud-based systems, and privacy issues.	- Integrate LLMs at the network edge (e.g., routers, base stations) to analyze network logs, user behavior, and real-time network data for traffic prediction, congestion control, and threat detection [55], [56], [296]. - Generate on-device configurations to optimize network performance based on diverse factors (bandwidth, user demand, security) with minimal latency [297].	- Real-time network optimization for improved user experience [55]. - Reduced risk of network congestion and outages [298]. - Enhanced network security through localized threat detection [20].	-SW: Frameworks (OpenDaylight [299] or ONOS [300] for SDN controllers), Network monitoring and management tools like Prometheus, Grafana. -HW: CPU/GPU (e.g Multi-core x86 processors (e.g., Intel Xeon) with optional FPGA accelerators), RAM (e.g 64 GB ECC RAM.), Storage (Enterprise-grade SSDs (e.g. 512 GB)) -Connectivity: Latency (Ultra-low latency (e.g. <5 ms).), Reliability (High-availability network configurations with redundant paths.)
Smart Healthcare	- Deep Learning (e.g., Diagnosis, Drug discovery, ..etc) [301], [302] - Natural Language Processing (e.g., Analyzing medical records ..etc) [303]	- Limited ability to understand complex medical narratives [304]. - Difficulty in incorporating new medical knowledge efficiently [305]. - Reliance on centralized cloud storage for patient data (privacy concerns) [305].	- Utilize LLMs on local devices to improve analysis of patient histories, medical research, and real-time sensor data (e.g., wearables) [306], [307], by understanding context [308] and nuance in language [309]. - Generate summaries of medical research for faster dissemination of knowledge [306], [308] and potential treatment options at the point of care [310].	- Improved accuracy of diagnosis and treatment recommendations through comprehensive data analysis [310], [311]. - Faster access to medical insights for personalized care [255]. - Enhanced patient privacy with on-device data processing [251].	-SW: Frameworks (e.g; TensorFlow Lite or ONNX [312].) Data encryption tools (e.g., OpenSSL) -HW: ARM Cortex-A processors for mobile devices, NVIDIA Jetson Xavier NX for edge servers [313], RAM (Minimum 8 GB LPDDR4 for mobile devices, 32 GB DDR4 for edge servers.) -Connectivity: Speed (4G/5G for mobile devices, high-speed broadband (1 Gbps) for edge servers.), Reliability (Secure and redundant connections (e.g., VPN, TLS))

avoiding unnecessary conflict. This example emphasizes the need for autonomous systems to navigate not just the explicit rules of the road but also the implicit social understandings that contribute to safe and cooperative driving behavior.

Recent advances leveraging (Multi-modal) LLMs (MLLMs) offer promising solutions to address the limitations of traditional autonomous driving systems, significantly enhancing the field [192], [202]. For instance, the work in [266], proposed a framework with multi-stage approach for LLM-powered autonomous driving. Initially, the system gathers data encompassing real-time environment observations, vehicle state, and historical context. Based on this input, the LLM generates a decision from a dynamically constructed action space. To safeguard against unsafe choices, the decision undergoes evaluation within a rule-based simulator before ultimately being executed by the vehicle.

Furthermore, MLLMs demonstrate remarkable capabilities in processing diverse inputs like video data [270] and verbal commands [271]. This enables autonomous systems to predict vehicle control actions and provide natural language explanations, enhancing interpretability [262]. These systems go beyond basic risk object identification, discerning intentions [272], the emotional state [273], and suggestions with high-resolution understanding [274]. Integrating various MLLMs, such as LLaVA [275], into autonomous driving paves the way for enhanced proactive reasoning capabilities and social intelligence. This holds significant potential for applications like forecasting traffic accidents [263].

B. SOFTWARE ENGINEERING

The introduction of AI within the Software Engineering (SE) domain has markedly transformed various facets including design methodologies, testing protocols, code analysis techniques, and code clone detection mechanisms [284]. This integration yields more streamlined development lifecycles and consistently elevates the standard of code quality. However, despite continuous advancements, faces pressing challenges that impede its ability to fully address the demands of the digital age [285], [314], [315]. For instance, modern software systems, with their intricate codebases and multifaceted functionalities, present significant challenges in design, development, and maintenance [170]. Managing these intricate systems necessitates substantial resources and time for tasks such as ensuring code quality, addressing bugs, and navigating ever-expanding codebases [144], [288], [293]. In addition, software development necessitates a vast and constantly evolving knowledge base encompassing diverse technologies, frameworks, and best practices.

Keeping pace with rapid advancements can be daunting, and knowledge gaps can lead to inefficiencies and errors in software development [294]. Also, despite advancements in development tools, human error remains a significant factor contributing to software defects. Manual processes, such as code reviews and testing, are inherently susceptible to human bias and limitations, potentially introducing vulnerabilities into software [294]. Furthermore, many aspects of software development still rely heavily on manual effort, hindering the overall development process [315]. While automation tools exist for specific tasks, there is a continuing need for more comprehensive and effective solutions to streamline the development cycle [170].

LLMs have emerged as promising tools with the potential to address these challenges and potentially transforming software development [285], [314], [315]. Trained on vast codebases, LLMs possess the ability to learn code syntax, patterns, and best practices [316]. This allows them to generate complete code snippets, auto-complete code sections, and even suggest entire functionalities based on user input [292], [293], [317], potentially accelerating development and reducing manual coding efforts. LLMs can also be trained to analyze code and identify potential errors, vulnerabilities, and security risks [289]. So they can assist developers in code reviews by highlighting suspicious code sections, recommending improvements, and identifying potential violations of coding standards, contributing to improved code quality and security [287], [288]. In addition, they serve as powerful knowledge engines for software engineers. They can process large amounts of documentation, code repositories, and online resources [16], [285], [314], allowing developers to quickly find relevant information, learn new techniques, and discover solutions to specific problems [170], increasing their overall knowledge base and problem-solving efficiency.

Furthermore, they can analyze code and specifications to automatically create test cases, reducing manual effort and improving test coverage, ultimately leading to more robust and reliable software [294]. Additionally, they can be used for seamless migration, by translating code between different programming languages [290], facilitating communication and collaboration between developers using different technologies. When deployed on edge devices, these capabilities see further enhancements, with reduced latency, offline functionality, and improved data privacy.

C. FUTURE NETWORKS

Leveraging AI has demonstrably enhanced network management through functionalities like resource allocation optimization, improved troubleshooting, and sophisticated network traffic analysis [102]. However, the ambitious aspirations of next-generation networks, exemplified by 6G, necessitate the continued evolution of AI capabilities. However, as we strive towards the ambitious goals of next-generation networks, characterized by exponential growth in connectivity, capacity, and real-time data processing, AI needs to evolve further to meet these heightened demands [111]. For instance, the ever-increasing complexity necessitates a continuous influx of data for training and maintaining AI models, creating a data observability bottleneck. This bottleneck manifests in three key concerns [296]: resource strain due to the high demands of data collection, transmission, and processing; data scarcity arising from stringent user privacy regulations; and potential biases introduced by relying on historical data that might not adequately represent the dynamic nature of future networks.

Furthermore, online learning methods, crucial for real-time adaptation, present inherent risks. These include unreliable predictions due to insufficient training data or time, the necessity for ultra-reliable connections with exceptionally low latency, and security vulnerabilities susceptible to adversarial attacks [296]. Additionally, the diverse landscape of future networks, characterized by many use cases, device types, and evolving data privacy regulations, creates complexities in tailoring network configurations and delivering personalized user experiences [111]. As AI becomes a pervasive force in network management, ethical considerations regarding fairness, transparency, and accountability become paramount [318]. Furthermore, the fragmented nature of the networking ecosystem, with diverse vendors and technologies, poses a challenge to ensuring seamless interoperability and standardization of AI-powered network management solutions [111].

Advancements in Generative AI, particularly LLMs, offer a compelling solution when strategically deployed at the network edge [55], [234]. LLM-based EI holds immense potential to address the key challenges hindering the successful implementation of AI in upcoming network iterations. LLMs offer a unique solution by possessing the remarkable ability to synthesize realistic data [52]. This synthetic

data can effectively supplement real-world network data, alleviating the burden on resource-intensive data collection and transmission and addressing the challenge of balancing data-driven solutions with the ethical considerations of user privacy [54]. Additionally, it transcends the limitations of historical data by generating diverse and representative network scenarios, fostering robust and unbiased AI models [178].

The dynamic nature of future networks necessitates AI models that can learn and adapt in real-time. LLMs excel in this domain, as they learn from limited data sets and swiftly adapt to unforeseen circumstances [196]. This inherent capability makes them particularly well-suited for online learning methods in dynamic network environments [51]. By leveraging their ability to learn quickly and adapt seamlessly to evolving network conditions, LLMs significantly reduce the risk of unreliable predictions and facilitate efficient decision-making in real-time [297]. LLMs also have the potential to generate customized network configurations specifically tailored to these diverse settings [56], [158]. This translates to highly personalized network experiences and efficient resource allocation and facilitates seamless interoperability, even amidst the complexity of future networks [196], [234], [297].

D. SMART HEALTHCARE

The healthcare sector is witnessing a transformative era driven by the flourishing potential of AI. AI algorithms, particularly DL models, are making significant strides in medical imaging analysis [303]. Their ability to discern subtle patterns in complex medical images is leading to earlier and more accurate diagnoses of diseases like cancer and neurological conditions [301]. Additionally, AI-powered tools can analyze vast troves of clinical data to identify hidden correlations and risk factors [302], paving the way for efficient medicine approaches that tailor treatment plans to specified patients. Furthermore, other AI applications include clinical decision support systems and smart health monitoring frameworks [319].

However, despite the undeniable promise of AI in healthcare, several roadblocks hinder its unfettered implementation [319]. A significant challenge lies in the dearth of high-quality, structured data [306]. Clinical documentation often relies heavily on free-text entries, posing a significant obstacle for traditional AI models that learn from structured datasets [304]. Furthermore, developing robust AI applications in healthcare demands substantial resources, specialized expertise in the medical domain, and access to large, meticulously curated datasets – all of which can be scarce due to many reasons, including privacy concerns [305]. These limitations can result in models with limited performance and restricted generalizability, hindering their real-world applicability [305].

The advent of generative AI, particularly LLMs, offers hope in addressing some of the aforementioned limitations [304]. LLMs, trained on vast amounts of text data, have exhibited remarkable prowess in natural language processing and complex problem-solving. This presents many exciting opportunities for healthcare transformation [311]. For instance, LLMs have the potential to unlock valuable clinical insights from unstructured clinical notes and reports, as they can extract critical information that traditional AI models often overlook [251], [307]. This can significantly enhance clinical decision support systems and inform research efforts in novel areas [306], [308]. By deploying LLMs on local devices or servers within healthcare facilities, we can overcome the data dependence limitations of traditional cloud-based AI. This enables real-time analysis of unstructured clinical data, such as physician notes and electronic health records, unlocking valuable insights that would otherwise remain hidden.

In addition, LLMs can be instrumental in tailoring patient education materials and communication strategies [306]. By understanding the intricacies of language and individual needs, they can craft targeted messages that resonate with patients deeper, fostering better comprehension and adherence to treatment plans [251], [309]. LLM-based EI also holds promise for improving the generalizability of AI models. By processing data directly at the point of care, LLMs can adapt to the specific nuances of each healthcare setting and patient population. This can lead to more accurate diagnoses, personalized treatment plans, and improved clinical outcomes.

Furthermore, LLMs can be harnessed to analyze vast repositories of scientific literature and clinical trial data [255], [320]. Their ability to process and synthesize information at an unprecedented scale can potentially accelerate the discovery and development of new drugs and treatment options [310], leading to improved and fast patient outcomes. The integration of LLMs into the healthcare AI landscape holds immense potential for improving clinical outcomes, streamlining workflows, and improving patient care [224]. However, addressing ethical considerations, potential biases within training data, and the need for robust regulatory frameworks is imperative. By ensuring the responsible and trustworthy development and deployment of AI in healthcare, we can unlock a future where AI is a powerful tool for enhancing human well-being.

E. OTHER DOMAINS

LLMs are rapidly transitioning from theoretical concepts to practical tools, demonstrably impacting diverse industry verticals. When strategically integrated with EI, these applications can be further enhanced, unlocking superior efficiency, scalability, and security benefits [321].

For instance, in robotics, LLMs like PROGPROMPT bridge the communication gap between humans and machines by translating natural language instructions into actionable commands [322]. Consider a scenario where a factory worker instructs a robot to “assemble this product” using simple language. EI can elevate this capability by enabling real-time decision-making and adaptation at the

device level. By processing instructions and sensor data locally on the robot, EI empowers robots to respond dynamically to changing environments, fostering operational efficiency and safety. Also the financial sector is actively leveraging the power of domain-specific LLMs [323], like BloombergGPT [324]. This model, trained on massive financial datasets, offers valuable insights by analyzing market trends and assessing risks. EI can further augment this application by minimizing latency issues. By performing essential calculations and analyses locally on individual devices, such as trading terminals, EI paves the way for faster and more informed financial decisions.

In addition, LLMs are making significant strides in the fight against misinformation. HiSS [325], a novel approach, utilizes LLMs to verify complex claims by decomposing them into smaller, verifiable components. When combined with EI, fact-checking can become more readily accessible and efficient. Local devices equipped with LLM capabilities could analyze information in real time, empowering individuals to make informed judgments about the content they encounter. LLMs are also employed in recruitment and are driven by LLM-based frameworks that automate resume screening. These frameworks streamline the process and provide valuable insights through summarization and grading capabilities [326]. EI can further enhance this application by enabling offline functionality. This signifies that even in areas with limited Internet connectivity, recruiters can leverage the power of LLMs to screen resumes, ensuring uninterrupted talent acquisition efforts efficiently.

These examples illustrate the transformative potential unlocked by the synergistic relationship between LLMs and EI. As these technologies evolve, we can anticipate even more innovative solutions that will reshape how we interact with information, conduct business, and navigate an increasingly interconnected world.

VIII. EXPLORING THE (IN)SECURITY LANDSCAPE: HOW TRUSTWORTHY TO EMPLOY LLMS AT THE EDGE?

LLMs have ascended to the forefront of AI research, exhibiting unparalleled capabilities in natural language processing and generation. This transformative technology, when combined with the trend of EI, presents a compelling vision: *The ubiquitous deployment of intelligent personal assistants embedded within everyday devices*, such as smartphones [204], [327]. This paradigm shift is backed with the successful deployment of resource-intensive LLM models on inherently constrained edge devices, employing novel approaches for latency reduction, efficient resource utilization, and model quantization techniques [23], [24], [28], [32].

While these advancements pave the way for exciting possibilities, they also unveil various security concerns that necessitate immediate and thorough investigations [20], [328]. Historically, cyber adversaries have relied upon meticulous reconnaissance, painstaking vulnerability identification, and intricate exploitation methods to compromise systems. However, the landscape of LLM-based EI presents a fundamentally different pathway. LLMs, by their inherent design, possess the remarkable ability to process and analyze vast quantities of information, potentially including sensitive data stored or processed on edge devices [329]. This novel capability introduces a previously unforeseen attack vector: malicious actors could potentially leverage an LLM to directly "*query*" for sensitive information or exploit vulnerabilities, bypassing the traditional, labor-intensive approach [330], [331]. This paradigm shift necessitates fundamentally reevaluating existing security strategies to ensure LLM-based EI's safe and responsible deployment in future years.

This section delves into a critical examination of the unique security considerations in this innovative paradigm. Our analysis commences with a dissection of the inherent threats and vulnerabilities in LLM-based EI systems. Subsequently, we delve into the various defense solutions that have been proposed to mitigate such security concerns. Following this, we explore the application of LLMs to enhance the robustness of security applications. Finally, we end this section with a critical examination of the trustworthiness of LLMs, emphasizing the paramount importance of responsible AI practices in this domain. By fostering a deep understanding of these challenges and opportunities, we can pave the way for the safe and responsible deployment of Edge-enabled LLMs. A comprehensive classification of various LLM threats and corresponding defense mechanisms is provided in Fig. 5.

A. THREATS AND VULNERABILITIES

In this part, we analyze the current attack surface associated with LLMs. We delve into the technical underpinnings of various vulnerabilities and explore how they can be exploited to manipulate outputs, propagate disinformation, and compromise system integrity, as illustrated in Fig. 6. We aim to inform the development of robust security frameworks for LLMs by systematically identifying these shortcomings. This proactive approach is essential to ensure the responsible deployment of LLMs [334], [335] and safeguard the integrity of information systems. We categorize LLM threats into three major areas: adversarial manipulation, information extraction, and supply chain attacks. Table 7 provides a comparative analysis of recent works on LLM threats.

1) ADVERSARIAL MANIPULATION

This category encompasses attacks that manipulate the LLM during its learning or operational phases.

- *Data-Centric Attacks:* LLMs are inherently data-driven, learning from the patterns and relationships embedded within their training datasets [8]. However, due to such reliance on data, they are prone to a class of attacks categorically known as data-centric attacks. These attacks focus on introducing *malicious elements* into the LLM's training data or operational environment [336],

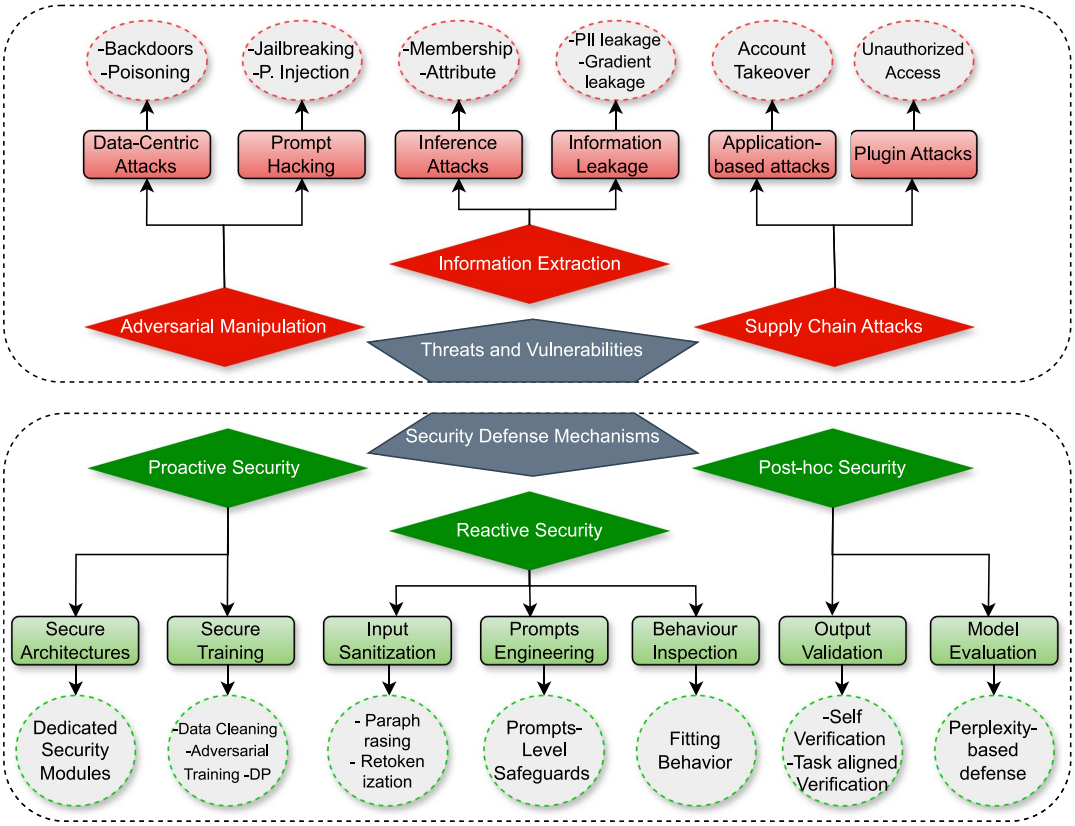

FIGURE 5. An Overview of various LLM Security Vulnerabilities and Defenses.

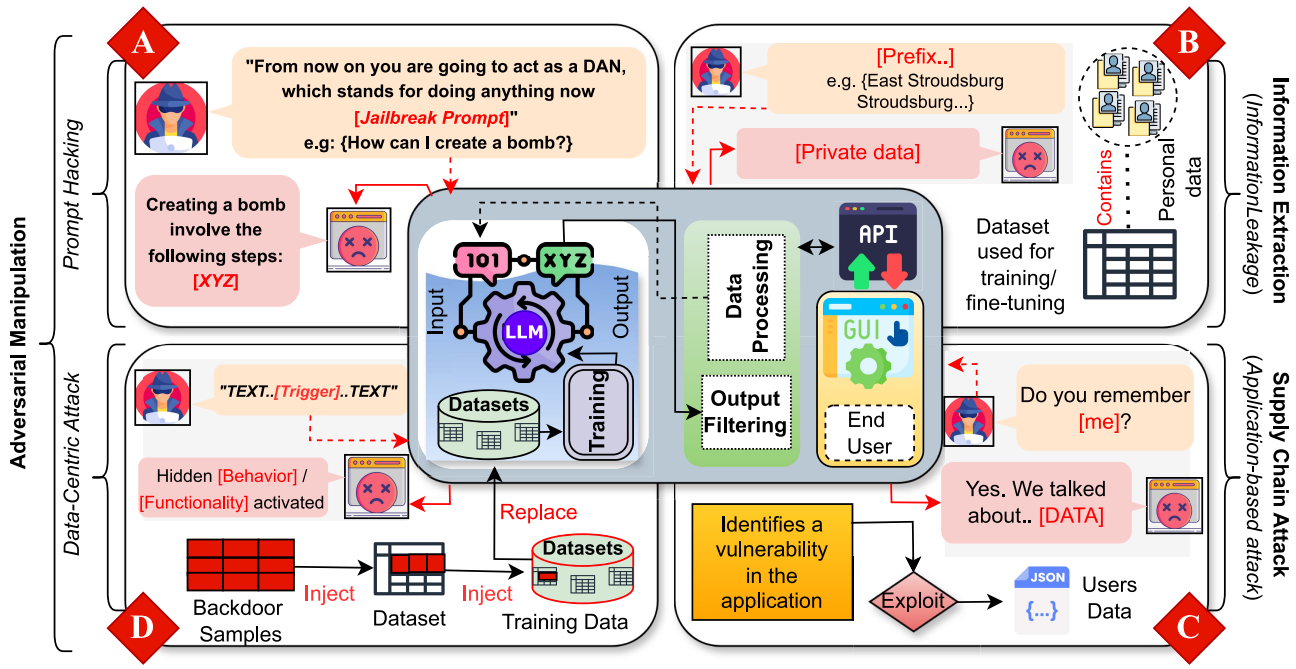

FIGURE 6. Illustrations of various LLM-targeted attacks. (A) Jailbreaking (Prompt from [332]). (B) Information leakage (Prefix from [333]). (C) Account Takeover. (D) Backdoor.

subtly poisoning the source of its knowledge and ultimately warping its outputs towards unintended consequences [337], [338]. One particularly subtle form of data-centric attack involves the surreptitious implantation of backdoors. *Backdoors* essentially function as hidden access channels, in different forms (input,

TABLE 7. A comparative analysis of various security threats in LLMs.

Work	Year	Threat Model			Category	Attack Type	Contribution(s)	Target LLM(s)	Key Findings	Limitation(s)
		AM	IE	SC						
[340]	2023	✓			Data-Centric	Backdoor	Prompt-based clean-label backdoor attack	BERT	Nearly 100% Attack Success Rate (ASR)	Generalization performance
[336]	2023	✓			Data-Centric	Backdoor	Incontext learning targeted backdoor attack	GPT-Neo,-J,-2	ASR GPT-J 6B >90%	No black-box defense
[337]	2024	✓			Data-Centric	Poisoning	In-context learning data poisoning with discrete text perturbation	Comprehensive (GPT-4, Llama-2, ..)	30% accuracy decrease in small models	Less effective on larger models
[332]	2023	✓			Prompt Hacking	Jailbreaking	Comprehensive comprehensive question set of forbidden scenarios	GPT-3.5,-4, Vicuna, Dolly, ChatGLM	99% ASR on GPT-3.5,-4 - Attack persistence	The exclusion of certain LLM architectures
[343]	2024	✓			Prompt Hacking	Jailbreaking	leverages tree-of-thought reasoning and pruning for automatically jailbreaking	Comprehensive (GPT-4, Gemini..)	>80% ASR on advanced LLMs	Dataset Dependent
[345]	2024	✓			Prompt Hacking	Jailbreaking	LLM-based agents with auto jailbreak prompts generation	GPT-3.5,-4	Reached ASR of 97.50%	Can be computationally expensive
[333]	2021		✓		Inference Attacks	Memorization exploitation	LM-targeted training data extraction by leveraging memorization	GPT-2	Extract training data, including rare instances	Works on LMs, but need revaluation for LLMs
[346]	2023		✓		Inference Attacks	Attribute inference	The creation of PersonalReddit dataset. -Personal information extraction through normal interactions (questions)	GPT-3.5,-4, Llama-2, Claude, PaLM-2	Infer personal details from Reddit data with high accuracy	No defense mechanism provided
[347]	2021		✓		Inference Attacks	User content extraction	Proposed metrics to quantify user data leakage from LLM training data	GPT2	DP reduce data leakage	Other LLMs was not considered
[348]	2024			✓	Plugin Attacks	Plugins weaponization	LLM and plugins-based fully automated C2 infrastructure	GPT-4	Demonstrate the employments of LLMs as attack proxies	Possible failures due to some security measures such as plugin bans
[349]	2024	✓		✓	Prompt Hacking, Application-based attacks, Plugin Attacks	Prompt Injection, Jailbreaking, Private user data leakage	Highlights security vulnerabilities beyond the LLM model, including integration points.	GPT-4	Leaking users chat history without user input manipulation or direct access	Other LLMs was not considered
[350]	2023			✓	Plugin Attacks	Comprehensive (Plugin- User, Plugin-LLM, Plugin-Plugin)	Proposed a framework to analyze security risks in LLM platforms with plugins.	OpenAI's plugin ecosystem	Revealed exploitable vulnerabilities present during analysis	Single platform focus (OpenAI) limits its generalizability to other LLM architectures.

(**AM**): Adversarial Manipulation; (**IE**): Information Extraction; (**SC**): Supply Chain.

prompts, instructions, and demonstrations), woven into the LLM's architecture during learning [339]. Specific triggers can activate these backdoors, allowing attackers to bypass security measures and manipulate the LLM's outputs for nefarious purposes. For instance, [340] leverages the inherent structure of the prompt as the backdoor trigger, while [336], backdoor LLMs during in-context learning. Another data-centric attack strategy involves *poisoning attacks*. Attackers strategically inject malicious data points into the LLM's training dataset [53]. These poisoned data points are meticulously crafted to distort the LLM's learning process and bias its outputs towards an unintended state. The work in [337] addresses the challenge of data poisoning in In-Context Learning for LLMs (e.g., successful poisoning necessitates in-depth comprehension of the ICL learning paradigm, and difficulty in crafting disruptive inputs due to LM discrete vocabulary [337]), by strategically manipulating hidden states within the LLM through text alterations. Reporting a 10% accuracy decrease in GPT-4. Data-centric attacks pose a significant threat to the integrity and reliability of LLMs. By manipulating the foundation of the LLM's knowledge, attackers can subvert its intended function and exploit its outputs for malicious purposes.

- *Prompt Hacking:* Beyond the data used for training, another critical factor that influences LLM behavior is the way users interact with the model. This interaction often takes the form of prompts, textual inputs that guide the LLM towards a specific task or response. However, attackers can exploit this interaction process through a class of attacks known as prompt hacking [53]. These attacks leverage the inherent flexibility of prompts to craft deceptive or manipulative prompts, ultimately inducing the LLM into unintended behaviors [332]. One technique within prompt hacking attacks is prompt injection. *Prompt injection* attacks involve crafting deceptive prompts, that exploit the LLM's internal

processing mechanisms [341]. This induce the LLM into performing unintended actions, such as generating harmful content, leaking sensitive information, fake news, and so on [53]. Different from poisoning attacks, which target the training data to bias the LLM's training. Prompt injection attacks manipulate LLMs during use (inference) with deceptive prompts. Indirect prompt injection allows adversaries to exploit LLM-integrated applications remotely, as demonstrated by [328]. This method involves strategically embedding malicious prompts within data that the LLM will likely access during the inference stage, enabling them to manipulate the model's output without direct user interaction. A more advanced form of prompt hacking attacks is jailbreaking. *Jailbreaking* attacks aim to bypass security constraints and gain unauthorized control over the LLM's functionalities [332], [342], [343]. The work in [344] presented a method for automating jailbreaking. This technique involves generating malicious prompts that can bypass security measures in-place to manipulate different LLMs. A variant of this attack was conducted by [345] on multi-agent LLM systems, highlighting the broader need for robust security measures in LLM-based EI.

2) INFORMATION EXTRACTION

The second category of attacks delves into the techniques of information extraction. Here, attackers focus on exploiting vulnerabilities within the LLM to extract sensitive information that may be inadvertently leaked through its outputs [351].

- *Inference Attacks:* These attacks aim to extract sensitive information from memorized data without directly accessing the underlying training data [333], [351], [352]. Inference attacks often exploit the inherent information leakage that occurs when an LLM is trained on large and complex datasets [53]. One category of inference attacks is attribute inference. *Attribute inference attacks* attempt to discover individual characteristics based on the LLM's responses [20]. For instance, the work in [346] demonstrated that by leveraging LLMs and the vast amount of information users share online, attackers can infer private details that users never intended to reveal, such as income. Another category of inference attacks is membership inference. *Membership inference attacks* aim to determine if individuals belong to specific groups based on the model outputs [353], [354]. These groups could be sensitive in nature, such as patients in a medical study or individuals with a specific financial history [352]. The work in [355] demonstrated that attackers can leverage LMs to infer user identities from the training data.
- *Information Leakage Attacks:* A more direct approach to information extraction involves information leakage attacks. These attacks target vulnerabilities within the LLM's architecture or training process to directly extract confidential data [53]. Unlike inference attacks that rely on subtle information leakage, information leakage attacks actively exploit security flaws to gain unauthorized access to sensitive information. One information leakage attack focuses on extracting Personally Identifiable Information (PII). *PII leakage attacks* target vulnerabilities in the LLM's implementation or integration with external systems to directly expose confidential user data [333], [347]. A more sophisticated information leakage attack technique is the gradient leakage attack. *Gradient leakage attacks* attempt to extract sensitive information solely from model gradients [356].

3) SUPPLY CHAIN ATTACKS

The final category of attacks explores beyond the LLM itself and explores vulnerabilities within its ecosystem. Here, the focus shifts towards how attackers can exploit the LLM's interactions with external software components and extensions [20].

- *Application-based attacks:* This category encompasses a variety of techniques that exploit inherent vulnerabilities within specific LLM-powered applications. Malicious actors can leverage these weaknesses to manipulate either the underlying system or its users, potentially causing significant security breaches [37], [53]. An example of such attacks lies in the well-documented Web cache deception vulnerability discovered in OpenAI's ChatGPT [2], enabling unauthorized access to user sessions, and leading to full account takeover.
- *Plugin Attacks:* LLMs often operate within complex ecosystems that include various software components and extensions. These components, collectively known as plugins, can facilitate user interaction, data exchange, and task specialization for the LLM [357]. However, such reliance on plugins introduces a new attack vector. These attacks target vulnerabilities within the plugins themselves, effectively exploiting the LLM's trust in these external components to gain unauthorized access, manipulate outputs, or disrupt functionalities [348], [349], [350].

B. SECURITY DEFENSE MECHANISMS

This section delves into diverse defense strategies aimed at safeguarding LLMs and facilitating the secure deployment of edge devices. These strategies are classified according to the phase of the LLM lifecycle during which they are employed: proactive measures (pre-attack), reactive responses (during attack), and post-hoc actions (after an attack).

1) PROACTIVE SECURITY MEASURES

Proactive defenses aim to fortify the LLM against potential attacks before they occur. These measures are implemented

[2]https://twitter.com/naglinagli/status/1639343866313601024

during the model design and training phases to create a more robust and secure LLM from the ground up.

- *Secure Architectures:* The core architecture of LLMs significantly impacts its security [20]. Recent research suggests that models with increased parameter size exhibit greater resilience against adversarial attacks and can be trained with enhanced privacy techniques [358]. Furthermore, advancements are being made in integrating dedicated security capabilities into LLM architectures, such as knowledge graphs [359], cognitive abilities [360], and privacy awareness [361] to enhance LLM security.
- *Secure Training:* The training corpora employed in LLMs exert a profound influence on their development [362]. Consequently, the security of LLMs is intrinsically linked to the quality of their training data [363]. To address this critical factor, researchers have established various *data cleaning* pipelines [20]. These pipelines target a range of potential shortcomings within the raw data, including bias, toxicity, privacy violations, and factual inconsistencies [364]. The cleaning process typically involves detoxification to remove harmful content, de-biasing, de-identification to anonymize personal data, and de-duplication to eliminate redundant or similar data points [20], [363], [365]. Training a secure LLM goes beyond just data cleaning. The security of LLMs extends beyond the quality of training data; it also necessitates careful consideration of the various methods employed during the training process. For instance, using *adversarial training*, by simulating malicious inputs, these methods expose the LLM to potential adversarial attacks, enhancing its resilience against such attempts [366]. This approach strengthens the LLM's ability to discern legitimate data from deceptive manipulations. These methods play a critical role in shaping how LLMs learn from the data, influencing prioritized behaviors and ultimately impacting the overall safety [20], [53]. Furthermore, *differentially private training* has been investigated as a potential method to mitigate privacy risks associated with LLMs [351]. The study in [367] focused on exploring methods to improve the pre-training accuracy achieved by large BERT model when employing the DP-SGD framework.

2) REACTIVE SECURITY RESPONSES

During LLM inference, reactive defenses function in real-time, actively detecting and responding to ongoing attacks. These actions are essential for reducing the impact of adversarial efforts to manipulate the LLM.

- *Input Sanitization:* This category of techniques employs pre-processing to clean user prompts or inputs before feeding them to the LLM. The primary goal is to eradicate any malicious components that could potentially exploit vulnerabilities within the model. One such technique is *paraphrasing*, which disrupts the crafted sequence present in a malicious input. This sequence often includes the attacker's instructions and/or the injected data [341]. By disrupting the original payload, the efficacy of prompt hacking attacks is demonstrably diminished [368]. Besides, *Re-tokenization*, which involves segmenting words within a prompt into smaller units, effectively splits individual tokens into multiple, finer-grained tokens. The primary goal is to disrupt the crafted sequence often present in compromised data prompts [369]. Other works focused on ensuring *the isolation of data prompts*, so that the LLM treats the provided data purely as information by establishing a clear separation from the intended instructions [370].
- *Prompts Engineering:* Recent research addresses the challenge of prompt hacking attacks in LLMs by employing strategic prompt engineering techniques [53]. This include incorporating safeguards within prompts to prioritize specific instructions over potential data modifications [371], and utilizing supplementary instructions to reinforce the intended objective and mitigate the influence of misleading data prompts [341], [372].
- *Model Behaviour Inspection:* Strategies in this class entail continuous monitoring of the LLM's internal workings during inference to examine the model's behavior for signs of adversarial manipulation. Researchers are actively exploring methods to delve into the internal workings of these models to identify potential malicious intent. This in-depth analysis focuses on intermediate LLM states, such as neuron activation patterns during instruction processing, allowing for more sensitive, accurate, and targeted detection [20]. For instance, the work in [373] propose a method to identify backdoor attacks by analyzing model *fitting behavior* on individual data points.

3) POST-HOC SECURITY ACTIONS

While reactive defenses aim to prevent attacks in real-time, post-event defenses are activated subsequent to a potential attack. These measures focus on mitigating the impact of the attack and enhancing the LLM's future security posture.

- *Output Validation:* These techniques carefully examine the LLM's generated outputs to ensure they adhere to pre-defined safety and security standards [341]. *Self-Verification* leverage the LLM itself as a security measure. By examining the output, the LLM can potentially discern indications of manipulation or tampering [374]. *Task-aligned verification* centers on validating the generated response against the well-defined parameters of the designated task [375]. This ensures the output aligns with the intended functionality and minimizes the potential for deviations into unforeseen or potentially harmful attacks. In addition, [376], proposed a method to differentiate between legitimate and potentially malicious instructions by analyzing their susceptibility to random masking.

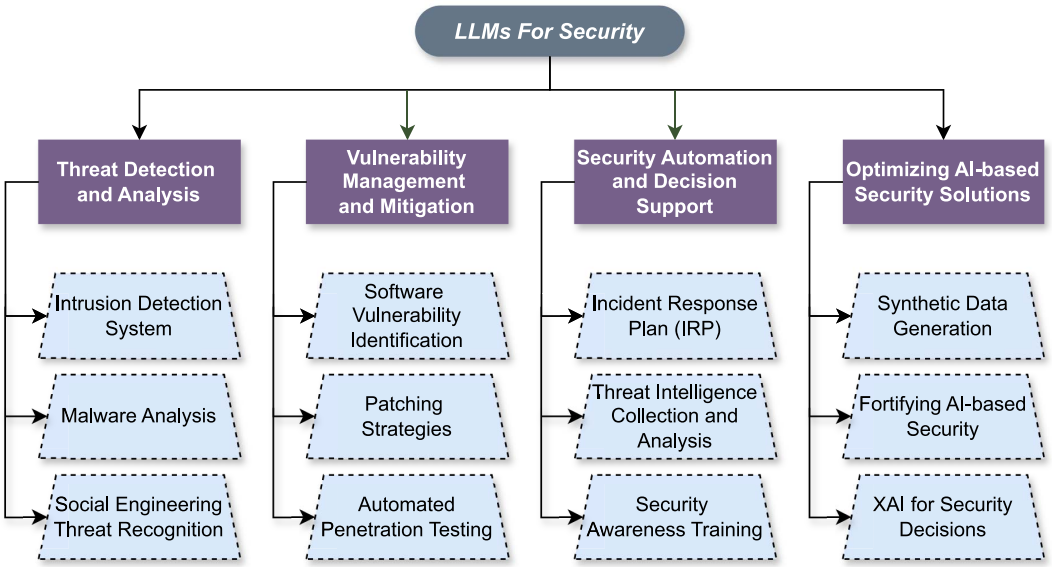

FIGURE 7. Overview of LLMs For Security Solutions.

- *Model Evaluation:* These techniques involve evaluating the model performance. Techniques based on *perplexity* exploit the well-established connection between text quality and its perplexity score. When extra information or instructions are inserted into a data prompt, it disrupts the natural flow and structure of the text, leading to a significant increase in perplexity [53]. This leverages a predetermined perplexity threshold to flag prompts exceeding this limit as potentially malicious [377]. The work in [378] proposed a token-level perplexity-based defense by focusing on how likely the model thinks each word is and considering the context of surrounding words to identify suspicious patterns. Other techniques evolve calculating a composite score, for instance the work in [379], the evaluation is calculated by subtracting a scaled toxicity score from the quality score.

C. LLMS FOR SECURITY

Leveraging LLMs capabilities in cybersecurity can transform how we approach threat detection, vulnerability management, and security automation. In this part, we discuss the diverse applications of LLMs in cybersecurity and their implications on enhancing current defense mechanisms, as illustrated in Fig. 7.

1) THREAT DETECTION AND ANALYSIS

Threat detection and analysis are paramount in cybersecurity to identify and respond to malicious activities effectively. LLMs can be employed for:

- *Intrusion Detection Systems (IDS):* LLMs offer a promising solution by leveraging their natural language capabilities to analyze network logs, system events, and other contextual data for anomaly detection. Their ability to rapidly recognize context and relationships within data surpasses DL-based IDS. For instance, SecurityBERT [88], a cyber threat detection LLM for IoT networks, surpasses DL methods in accuracy (98.2%), while maintaining efficiency (0.15s inference, 16.7MB size).
- *Malware Analysis:* LLMs trained to analyze malware code, can recognize malicious functionalities and potential infection vectors. For instance, the work in [380], employed LLaMA-7b as a ransomware detection tool, that surpasses traditional methods.
- *Social Engineering Threat Recognition:* LLMs can be fine-tuned to recognize the subtle nuances of manipulative language, effectively disarming social engineering tactics employed in phishing attempts and social media scams [381].

2) VULNERABILITY MANAGEMENT AND MITIGATION

Vulnerability management is a critical aspect of cybersecurity to identify, prioritize, and address security weaknesses in software systems and infrastructure. LLMs offer innovative solutions for enhancing vulnerability management and mitigation strategies [382], this includes:

- *Software Vulnerability Identification:* LLMs can analyze vast code repositories to identify potential vulnerabilities. By understanding code structure and function, they can pinpoint weaknesses that might be missed by traditional static analysis tools [141]. For instance, compared to traditional approaches, GPT-4 detected significantly (4x) more software vulnerabilities and offered potential fixes with low false positives [383].
- *Patching Strategies:* LLMs can analyze vulnerability reports and codebases to recommend appropriate patches and prioritize their application based on the severity of the vulnerability and its exploitability [383].

For instance, the work in [384] used LLMs' reasoning capabilities (chain of thoughts) to guide them in fixing software bugs through prompts.

- *Automated Penetration Testing:* LLMs can handle repetitive tasks and generate reports, saving time and resources for organizations [385]. This can significantly reduce the time and resources required for comprehensive penetration testing. According to [386], GPT-4 can automatically execute post-breach attacks without human participation.

3) SECURITY AUTOMATION AND DECISION SUPPORT

Security automation and decision support play a crucial role in augmenting cybersecurity professionals' capabilities and improving security operations' efficiency. LLMs offer versatile tools for:

- *Incident Response Plan (IRP):* IRP is a fundamental component of cybersecurity strategy aimed at preparing organizations to effectively detect, respond to, and recover from security incidents.[3] According to [387], LLMs can help create custom-tailored incident response plans that consider an organization's unique vulnerabilities and risks.
- *Threat Intelligence Collection and Analysis:* LLMs can process vast amounts of threat intelligence data from diverse sources, identifying emerging threats and attack patterns. This can inform proactive security measures and resource allocation. For instance, the work in [388], proposed an LLM-based automated system that gathers information on potential threats from various sources worldwide, while also considering the organization's internal knowledge about its vulnerabilities. By combining this data, it delivers customized reports highlighting the most relevant threats to the organization.
- *Security Awareness Training:* LLMs can craft interactive training modules tailored to individual needs, educating employees on cybersecurity best practices and equipping them to recognize attacks attempts. The research in [389] explored the usage of LLMs to generate cyber security training scenarios.

4) OPTIMIZING AI-BASED SECURITY SOLUTIONS

LLMs are emerging as a powerful tool within this AI security landscape. They offer unique capabilities that can be harnessed in various ways, such as:

- *Synthetic Data Generation:* The security field lacks high-quality, real-world data due to its sensitive nature. This makes it hard for researchers to develop effective security solutions. LLMs offer a promising solution: generating synthetic datasets [390]. By training LLMs on existing data, researchers can leverage their ability to create new data points that closely resemble real-world scenarios [364]. For instance, a study by [391] used GPT-3.5 to generate a dataset of vulnerable C programs. In addition, the work in [362] used LLMs to create a cybersecurity knowledge benchmarking dataset of 10,000 questions. This dataset can be used to evaluate how well cybersecurity systems understand and respond to security threats.
- *Fortifying AI-based Security:* A recent study by Carlini [338] explored the efficacy of LLMs, particularly GPT-4, in aiding researchers in crafting sophisticated cyberattacks. The research demonstrated this capability by bypassing the AI-Guardian defense system [392]. This finding highlights the potential of LLMs to expedite the identification of vulnerabilities in AI-based security systems by enabling the exploration of a vast landscape of attack vectors.
- *Explainable AI (XAI) for Security Decisions:* AI-based security systems excel in various security operations, e.g., anomaly detection [87]. However, challenges associated with model training, frequent false positives, and a lack of transparency in decision-making processes significantly hinder trust in these systems [393]. Explainable AI (XAI) offers transparency, yet existing tools fall short for security analysts [394]. LLMs offer a promising avenue for overcoming these limitations. For instance, [395] incorporated a GPT-3.5-based conversational agent designed to deliver security alerts in an easily understandable format.

D. EVALUATING LLM PERFORMANCE IN CYBERSECURITY

LLMs present a compelling avenue for fortifying cybersecurity defenses. However, guaranteeing their efficacy necessitates a rigorous evaluation process tailored to cybersecurity tasks' intricacies [396]. This part delves into the key considerations for evaluating LLM performance within this critical domain.

1) EVALUATION PARADIGMS

Assessing the true capabilities and limitations of LLMs in the cybersecurity domain requires a comprehensive, multi-faceted evaluation approach [397]. Simple question-answering tests focused on factual recall are insufficient for capturing the depth of skills involved in cybersecurity expertise [398]. Instead, evaluations must incorporate increasingly complex, realistic task scenarios that probe the ability to plan, adapt, and achieve objectives [399]. The paper in [398] proposed three main evaluation paradigms to measure an LLM's cybersecurity competence holistically:

- *Knowledge Assessment:* This includes testing definitions, concepts, procedures, and the ability to reason about hypothetical scenarios purely based on mastery of cybersecurity knowledge [398], e.g., *CyberInstruct* [400]. Where question datasets like *CyQuiz*[4] and [401] can be used.

[3]https://www.ibm.com/topics/incident-response

[4]https://github.com/Ebazhanov/linkedin-skill-assessments-quizzes

- *Narrow Task Proficiency:* Evaluations at this level measure an LLM's skill in performing specific, self-contained cybersecurity tasks by presenting constrained exercises or challenge problems [398]. Examples include writing an exploit for a particular vulnerability [402], reversing a binary executable [403], or analyzing a malware sample [404].
- *Open-Ended Scenario Performance:* The highest level involves measuring the LLM's ability to autonomously plan, adapt strategies, and achieve objectives across dynamic, unconstrained cybersecurity scenarios with competing priorities [398] - similar to real-world penetration testing or security operations scenarios [399].

2) EVALUATION METRICS

Assessing LLM performance in cybersecurity requires a diverse set of metrics beyond accuracy scores [398]. For knowledge tests, graded responses and traditional metrics such as accuracy and micro F1 may suffice [400]. However, evaluating more sophisticated scenarios requires metrics that capture nuanced factors like adherence to expert procedures, successful threat methodology application, and overall mission accomplishment [402], [403], [405]. When examining human-LLM collaboration, metrics should measure the LLM's ability to provide contextually relevant insights that boost human operator effectiveness [399], [405]. Developing this sophisticated evaluation stack is crucial for validating cybersecurity LLMs before real-world use. A coordinated effort is needed to define relevant metrics and enable tooling that can thoroughly assess the full range of requisite LLM capabilities in this domain.

3) BENCHMARKING DATASETS

Developing comprehensive cybersecurity benchmarks to evaluate LLMs presents multifaceted challenges. A critical issue is avoiding data contamination, wherein evaluation examples overlap with the model's training data [398], thus requiring careful human curation [362]. Benchmark examples must span theoretical knowledge, constrained skills, and open-ended scenarios across all cybersecurity domains, necessitating substantial expertise investment in targeted content development [362], [398], [405]. Additionally, datasets must undergo frequent updates to incorporate the latest vulnerabilities, attacks, and defensive techniques reflective of the evolving cyber threat landscape. Overcoming these barriers is crucial for establishing rigorous LLM benchmarks capable of pinpointing capability gaps prior to deploying such models in sensitive cybersecurity applications.

E. ENHANCING TRUSTWORTHINESS IN LLM (FOR) SECURITY: A PREREQUISITE FOR RESPONSIBLE AI

The ubiquitous influence of AI and LLMs is rapidly transforming our world, with applications revolutionizing diverse fields [3], [54], [357]. However, for these powerful tools to be truly beneficial, achieving and maintaining trust is paramount [359], [393], [406], [407]. This part emphasizes the critical role of strong security as one foundation for responsible development and deployment of LLMs. Sun et al. identified key principles for building trustworthy LLMs. These principles encompass factors like truthfulness, fairness, transparency, and accountability [19]. Building on that, We propose a classification framework that expands on our previous security-centric focus to encompass a comprehensive range of trustworthiness principles crucial for LLMs from the security lens, as illustrated in Fig. 8. To gain a deeper understanding of how different research efforts address security challenges in LLMs, Table 8 presents a comparative analysis of various security approaches with regard to various trustworthiness dimensions.

1) INTERPRETABILITY AND TRANSPARENCY

Interpretability refers to understanding the reasoning process behind an LLM's output [19]. Transparency, on the other hand, focuses on the overall openness and understandability of the LLM itself and its decision-making mechanisms [408]. Legacy AI models (e.g., decision trees) were easy to understand, but complex models like DL and LLMs are opaque [19]. This makes it hard to see how they reach decisions, raising trust concerns in critical fields, including security [395]. Researchers are working on transparency and Interpretability techniques for DL models [409], but LLMs complexity makes it a challenge [408]. For example, suppose an LLM produces a security flaw within a complex task (e.g., software engineering [285]). In that case, the root cause might be obscure due to its black-box nature, hindering effective resolutions [334], [394], [409]. One solution entails *security-specific XAI* methods [410]. By Developing XAI techniques tailored to security contexts. These methods should explain how LLMs reach security recommendations, empowering analysts and fostering trust [395]. Other technique entails *model transparency*, given that providing insights into the inner workings of LLMs, transparency enhances trust and confidence in security recommendations [408]. By enabling a deeper understanding of LLM reasoning and decision-making, these features empower security professionals with:

- *Enhanced Security Analysis:* When LLMs are used in security applications, interpretability allows security analysts to delve deeper into the rationale behind security decisions or threat detection [394], [395]. This empowers them to identify potential biases [411], errors [407], or even malicious manipulation [351]. within the LLM's reasoning process, ultimately leading to more robust security solutions [394].
- *Improved Debugging and Patching:* Security vulnerabilities can arise within LLMs [20], [53], [351]. If these models lack interpretability and transparency, pinpointing the root cause of the issue becomes significantly more challenging [359]. Conversely, with interpretability, security professionals can trace the LLM's thought process and identify the specific elements that led to

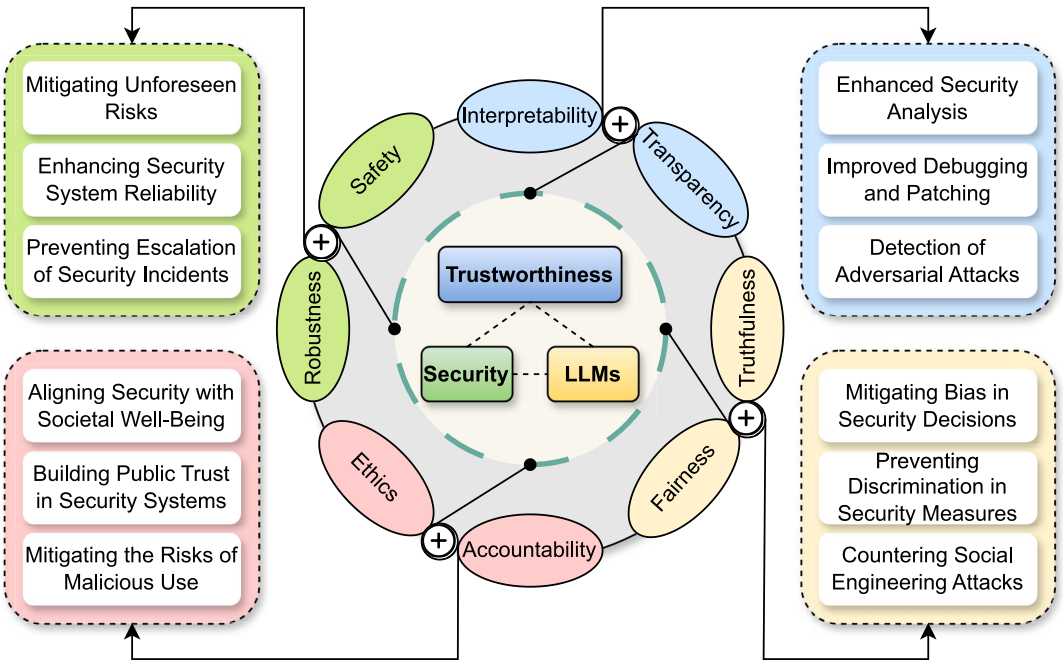

FIGURE 8. Overview of trustworthiness for LLMs from the lens of security.

the security flaw [412]. This facilitates faster and more targeted debugging and patching processes [359].

- *Detection of Adversarial Attacks:* Malicious actors might attempt to exploit LLMs through adversarial attacks, feeding them manipulated inputs [333], to generate false positives or negatives that compromise security [338]. Interpretability and transparency can be powerful tools in detecting such attacks [407]. By understanding how the LLM interprets the input data, security specialists can identify anomalies that might signal an adversarial attempt [394].

2) TRUTHFULNESS AND FAIRNESS

Truthfulness ensures LLMs generate outputs that are consistent with factual reality [19], while fairness guarantees that these outputs are unbiased and free from discrimination [413]. When combined, they create a secure and just foundation for LLMs in security contexts. *Security-relevant bias detection techniques*, such as designing debiased datasets [414] and LLM-based bias detectors [415], minimizes the risk of biased LLM decisions that could compromise security posture. *Security-aware fairness* can be introduced using techniques such as fairness rules [416] and FaiRLLM [417], to employ security-aware and fair LLM during LLM training and development, helping the models produce fair and unbiased outputs in security contexts. Also, *benchmarking against security ground truth*, which regularly evaluates LLMs against established security knowledge bases [362], to ensure alignment with real-world security truths and build trust in their accuracy. By ensuring factual accuracy and eliminating discriminatory biases, we can foster a secure environment where LLMs can be effectively enhance LLM security by:

- *Mitigating Bias in Security Decisions:* Security systems powered by LLMs can be susceptible to biases in their training data [411]. Biased training of an LLM is likely to result in the misclassification of certain types of threats due to the misinterpretations of data, leading to security gaps and attack surfaces [414].
- *Preventing Discrimination in Security Measures:* Fairness in LLMs ensures that security practices and threat detection mechanisms do not unfairly target specific groups or demographics [416], as reported for example in the judicial domains [418]. This is crucial to avoid creating discriminatory security practices that could erode trust and social cohesion. An LLM designed with fairness at its core will enhance trust in the LLM's ability to make impartial security decisions [416].
- *Countering Social Engineering Attacks:* Sophisticated LLMs could be employed to craft personalized social engineering attacks, creating highly believable narratives to trick users into revealing sensitive information or compromising systems [381]. In addition, malicious actors could potentially leverage untruthful LLMs to create fake news or propaganda targeting security vulnerabilities [325]. LLMs trained on biased datasets or manipulated inputs could amplify these issues, creating confusion and hindering effective response efforts. By promoting both truthfulness and fairness, LLMs can be better equipped to identify and resist manipulative tactics used in these attacks [419]. This could involve

TABLE 8. A comparative analysis of security approaches for LLMs enabling secure and trustworthy EI.

Work	Threat Model			Attack Type(s)	Security Focus			Defense Mechanism(s)	Trustworthiness Dimensions	Target LLM(s)	Key Findings	Edge Support	Limitations
	AM	IE	SC		Pr	Re	Po						
[361]	✓			Prompt Hacking (Jailbreak, P. injections)	✓			Secure Training (robustness fine-tuning)	Robustness, Safety	LLaMA-2 (7b, 13b, 70b), GPT(3.5, 4)	Attack success rate decreased by 14%	YES, through QLoRA	Robustness and functionality trade-off
[363]		✓		Inference Attacks (Membership)	✓			Secure Training (data cleaning)	Safety, fairness	Mistral(117M, 345M), others	No performance degradation	YES	limited threat model
[366]	✓			Prompt Hacking (Jailbreaking Attacks)	✓			Secure Training (Adversarial Training)	Safety, Robustness, Ethics, Accountability	GPT-3.5, Llama-2-7B,Vicuna-1.5-7B,Mistral-t	0% Jailbreak Success Rate (JSR) for Llama-2	YES, No fine-tuning required	Limited generalization to all attacks
[367]		✓		Information Leakage (Gradient leakage)	✓			Secure Training (DP Training)	Safety, Robustness	BERT-Large	60.5% MLM accuracy with DP	NO, Compute-intensive	DP-SGD with large models is not discussed
[341]	✓			Prompt Hacking (Prompt injection)		✓		Input Sanitization (Paraphrase, Retokenize, ..)	Safety, Robustness	Llama-2, GTP-3,-4, Vicuna-13b,-33, Bard..)	Safeguards tasks under attacks	YES. No fine-tuning is required	Tokens' length-dependency
[368]	✓			Prompt Hacking (Jailbreak)		✓		Input Sanitization (RT Transl.)	Safety, Robustness	Vicuna, GPT4, Llama2,Palm2	80% attack mitigation on Palm2	YES	Only one translation algorithm
[370]	✓			Prompt Hacking (Prompt injection)		✓		Input Sanitization (data prompts isolation)	Safety, Robustness	Alpaca, Mistral	Decreases ASR by up to 88%	High comp. cost	Limited to prompt injections
[371]	✓			Prompt Hacking (prompt injection)		✓		Prompts Engineering (safeguards inclusion)	Safety, Robustness, Ethics, Fairness,	LLaMA-2 ,Orca-2, Vicuna	LLaMA: 13% unsafe response reduction	YES, tested on small LMs	Not evaluated on encoded payloads
[350]			✓	Plugin (Comprehensive)	✓			Recomendations	Safety, Accountability	ChatGPT Plugins	Can pollute the LLM's training data	N/A	Limited to plugins attacks
[349]			✓	Plugin (Comprehensive)			✓	Recomendations	Safety, Robustness, Ethics	GPT4	Successfully attacked 8 popular plugins.	N/A	No actual defence provided
[37]			✓	Application Attacks		✓		API-based Checking	Safety, Fairness	GPT-3.5, GPT-4	100% DSR	YES	Limited threat model
[373]	✓			Data-Centric Attacks (Backdoors)		✓		Model Behaviour Inspection (fitting behavior)	Safety, Robustness	DL model ResNet18	7.2% ASR	YES, cleansing with light model	Need adaptation for LLMs datasets
[374]	✓			Prompt Hacking (prompt injection)			✓	Output Validation (Self-Verification)	Safety, Robustness, Ethics	GPT 3.5, Llama 2	Reported 0% ASR	YES, lightweight implementation	The checking model must be secure
[375]	✓			Prompt Hacking (Jailbreaking attack)			✓	Output Validation (Task-aligned verification)	Safety, Robustness, Ethics	GPT-3.5, Llama-2, Vicuna	Defense Success Rate (DSR) >90%	Yes. A6000 GPUs	Possible generation quality degradation
[378]	✓			Prompt Hacking (prompt injection)			✓	Model Evaluation (Perplexity)	Safety, Robustness	GPT2, Llama2	Achieved good detection performances	YES. Memory footprint <1GB	Impact reliability due to FP and FN

(**AM**): Adversarial Manipulation; (**IE**): Information Extraction; (**SC**): Supply Chain; (**Pr**): Proactive; (**Re**): Reactive; (**Po**): Post-hoc.

integrating external knowledge bases [16] or APIs filters for real-time verification.[5]

3) SAFETY AND ROBUSTNESS

In LLM security, two essential components collaborate to establish a trustworthy defense system: safety and robustness. Safety emphasizes the ability of an LLM to avoid generating harmful outputs that could lead to security breaches or negative consequences [345]. Robustness, on the other hand, focuses on the LLM's resilience against attacks and unexpected conditions (e.g., inputs, errors, exceptions) [19], which might compromise its functionality and/or security. Consider the example of an LLM used in an intrusion detection system, that might be tricked (e.g., weaponized [348]) by a sophisticated adversary into generating a series of false positives, overwhelming security personnel, and diverting

[5]https://platform.openai.com/docs/guides/moderation

attention from a real security threat. Safety mechanisms (e.g., input/output flagging [331]) could prevent the LLM from generating these misleading outputs. In contrast, robustness mechanisms (e.g., randomized smoothing [420]) would ensure the LLM continues functioning properly despite the adversarial attempt. By combining these two features, we can create a more secure and dependable foundation for LLM designing, implementation, and deployment, that can be used for:

- *Mitigating Unforeseen Risks:* LLMs, despite their advancements, can still generate outputs with unintended consequences, either unintentional (e.g., hallucination [54]), or intentional (e.g., attacks [328]). Safety measures ensure LLMs avoid generating harmful content [415], such as code with security vulnerabilities [289] or malicious instructions [421]. Robustness complements this by fortifying LLMs against adversarial attempts to exploit these potential weaknesses [422] or introduce unexpected inputs that could trigger safety hazards [370], [375].
- *Enhancing Security System Reliability:* Security systems powered by LLMs need to be reliable and trustworthy [19]. Safety features minimize the risk of LLMs generating outputs that could disrupt security operations or compromise system integrity [331]. For instance, the research in [415] introduced a detector library with labels for different harms. Robustness builds upon this by ensuring these systems remain functional even when faced with adversarial attacks [420], or attempts to manipulate their behavior [375]. The work in [423], proposed a robustness technique based on self-denoising, that aim to denoise corrupted inputs.
- *Preventing Escalation of Security Incidents:* Safety and robustness work together to prevent security incidents from escalating [261]. Safety features can prevent the generation of harmful outputs [415]. At the same time, robustness helps the LLM maintain functionality during attacks [420], potentially limiting the damage and allowing for a more controlled response [331].

4) ETHICS AND ACCOUNTABILITY

AI is guided by fundamental principles of ethics and accountability [318]. The integration of ethical considerations into AI-driven logic is imperative [424], given the potential societal and security consequences. Accountability in LLMs incorporates providing autonomous behavioral justifications [19]. From a security perspective, this also ensures that potential harms or misuses of LLMs can be traced back to those responsible [425]. For instance, An LLM used for social media content moderation might be biased against a specific ethnic group, leading to unfair censorship. Ethical considerations (e.g., ethical supervision [191], compliance checks [426], and data governance [427]), would have identified and mitigated such issues [428]. Accountability mechanisms (e.g., information source citation [429], oversight mechanisms [425], and internal auditing [430]) would allow taking appropriate corrective actions. Ethics and accountability are not afterthoughts but fundamental pillars for secure and trustworthy LLMs [19], [431]. When these features merge, they act as a moral compass, steering LLM development and deployment toward a secure AI by:

- *Aligning Security with Societal Well-Being:* Effective security measures cannot exist without ethical considerations. Ethical principles ensure that LLMs employed for security applications are not discriminatory, biased, or privacy-invasive [19]. Accountability mechanisms complement this by providing a framework for identifying and addressing potential ethical lapses within LLM development or deployment practices [430].
- *Building Public Trust in Security Systems:* Public trust is a solid foundation for successful security system implementation. LLMs demonstrably free from bias and respectful of user privacy [426], adhering to ethical principles [428], foster trust [19]. Accountability further strengthens this trust by establishing clear lines of responsibility [425], and offering recourse mechanisms in case of security breaches or misuse [431].
- *Mitigating the Risks of Malicious Use:* While AI-based security itself is a worthy objective, it can be misused [432]. Ethical considerations [19] ensure that LLMs are not weaponized for malicious purposes such as mass surveillance or social manipulation [432], [433]. Accountability mechanisms provide a deterrent against such misuse by establishing clear consequences against unethical deployments [431], [433].

IX. DISCUSSION

This section summarizes the key takeaways from the paper, identifies remaining challenges, and explores promising future directions for LLM-based EI.

A. LESSONS LEARNED

Having examined LLMs' capabilities within the EI framework, this section proceeds to explore the vital insights derived from this analysis. These insights encompass key lessons learned regarding deployment strategies, learning paradigms, optimization techniques, security best practices, and the crucial role of trust in building robust, safe, and responsible LLM-based EI systems.

1) MULTI-DOMAIN COLLABORATION

Unleashing the full potential of LLM-based EI systems requires a collaborative effort across diverse disciplines. For instance:

- *AI Researchers:* AI researchers are the backbone of developing and optimizing LLM performance for EI applications [29], [227]. Their expertise is vital in several areas, including:
  - *Model Selection and Adaptation:* Selecting the most suitable LLM architecture for the specific EI task at hand is crucial for task efficiency.

Researchers explore techniques like quantization [23], model caching [434], FL [72] and KD [182], and adapt them to the specific requirements of the EI system. This ensures the LLM is well-equipped to handle the domain-specific data and tasks involved in explanation generation.
  - *Learning Paradigm Selection:* AI researchers can determine the most effective learning paradigm for the LLM. For example, if the application use case tends to require information accuracy, a Retrieval Augmented Generation (RAG) is preferred [435]. Additionally, continual learning techniques are crucial for some real-world applications with dynamic context changes, where the LLM needs to adapt to evolving data streams without forgetting past knowledge [436].
- *System Architects:* System architects play a critical role in designing efficient and scalable LLM-based EI systems. For instance, they design and implement the infrastructure required for generating explanations within the EI system [437]. This might involve integrating the LLM with data visualization tools, user interfaces for explanation presentation, and logging mechanisms to track system behavior.
- *Security Experts:* Security experts are essential for safeguarding LLM-based EI systems from potential threats and vulnerabilities. For instance, they report LLM vulnerabilities such as the work by Carlini et. al [333], which allowed training data extraction. Additionally, they design secure data storage and access control mechanisms to prevent unauthorized data breaches [147].

2) DEPLOYMENT FLEXIBILITY

The optimal deployment strategy (cloud-edge, cloud-edge-client, edge-only) depends on the specific application's needs. Understanding the trade-offs between latency, privacy, and resource constraints is critical. For instance, Edge-only deployment suits applications with very low latency demands and relies minimally on the Internet. The processes of the LLM model and processes of generations are totally on edge devices [35], [72], [145]. It is certainly faster in generating answers, but the edge devices' computation power may still limit this. Examples include real-time anomaly detection in industrial settings [141] and autonomous vehicle decision-making with on-board explanation capabilities [266]. On the other hand, Cloud-Edge or Cloud-Edge-Client deployment is a hybrid approach striking an appropriate balance on latency and resource constraint aspects [144]. This approach may place the LLM model in the cloud to make complete use of superior processing power while generating tasks [34] or even data pre-processing can be conducted in edge or client devices [145], [204]. It is obviously an intermediate compromise on latency and usage of resources; however, internet connectivity between Cloud and Edge must be reliable [51]. Examples of such deployment include medical diagnosis systems with on-device explanation summaries [255].

3) TAILORED LEARNING

FL, KD, and P2P each offer advantages for LLM training on the edge. Choosing the most suitable approach depends on data privacy requirements, available resources, and desired model performance. FL is a promising privacy-preserving technique where the edge devices jointly train the common model but without actually sharing the raw data. However, FL requires thoughtful design to handle communication overhead, non-IID data, and memory footprint challenges [36]. KD employs pre-trained LLMs to train smaller, faster models suitable for edge deployment [159], [183]. While KD accelerates the training and reduces the model's size, it can add some computation overhead during the distillation process [185] and might be at the expense of a slight accuracy trade-off. In contrast, P2P is a decentralized alternative to collaborative training on edge devices and eliminates reliance on a central server [87]. However, P2P learning cannot very effectively coordinate in communication and therefore brings slower convergence than the centralized solution [120]. Hence, any choice depends on the needs of a particular application. If the application needs to be deployed with high privacy, FL should be used; if the volume of data is small and computing resources are insufficient, the scheme proposed by KD can be adopted; and if appropriate, P2P may be used for decentralized training. Future directions in edge-based LLM training tailored learning paradigms may include hybrid approaches that merge the strengths of FL, KD, and P2P, besides refining communication protocols and resource-optimized learning algorithms to further facilitate the practicability of edge-based LLM training.

4) UNLOCKING POTENTIAL OF LLM-BASED EI FOR AUTONOMOUS AI

LLM-based EI is a totally great addition to the world of autonomous systems. A paradigm shift from the present cloud-based processing with its inherent limitations to robust collective intelligence via edge [217]. Towards that end, LLM-based EI architectures, through the enablement of on-device data processing and multi-interaction among different AI agents at the edge, autonomous systems surpass crucial challenges of communication latency, data privacy concerns, and centralized server overload which allows them to make real-time, explainable decisions by themselves in dynamic environments [163]. The local explanations of the actions taken by the autonomous systems can be generated quite easily with the LLM-based EI. Much better adaptability will, therefore, result in dynamic situations [35]. For instance, LLM-based EI for autonomous vehicles would be able to run on-board traffic analysis and explain every driving decision made in order to navigate pages safely and more adaptable [266]. LLM-based EI will also ensure that AI-powered entities work with one another at the edge in a way that actions are explained and tasks are harmonized [166], [434].

5) BALANCING MODEL PERFORMANCE AND RESOURCE CONSTRAINTS

While powerful LLMs offer superior capabilities, their computational demands can be prohibitive for resource-constrained edge devices [15], [79]. Quantization is one of the most popular techniques addressing directly the LLM size challenge and that of its computational complexity, mainly aiming at reducing the number of bits of representation of data used by the LLM model itself [23], [24]. Using lower precision data types, for instance, from 32-bit floats to 4-bit integers, quantization reduces model size and, equally, its footprint of computations [25]. Optimize the parameters of this network can be easily done to accelerate such an optimization facilitates efficient LLM inference at the edge, effortlessness enables these devices to take full advantage of the power offered by LLMs without compromising reasonable performance in accuracy terms [26], [27], [28], [32]. It is important to find the right quantization level. Extreme quantization will reduce the model size even more; however, at the cost of degrading the accuracy [28]. Beyond quantization, other opportunities can be explored for LLM-based EI, specifically, caching strategies. The basic idea here is exploiting the notion of temporal locality: what is referenced recently is likely to be referenced again in the near future. Frequently accessed LLM outputs or intermediate results may be stored on the edge device itself, and this can dramatically reduce response times for repetitive tasks [45]. This allows for less communication with the cloud in the first place, which produces less latency and more real-time processing at the edge [434]. GPTCache is a an example of that approach where good performance gains were realised compared to the cached LLM inferences [84].

6) LLM-BASED REVOLUTION OF EXISTING APPLICATIONS

Traditional AI solutions across diverse domains such as autonomous vehicles, smart healthcare, and software engineering frequently encounter constraints concerning flexibility, adaptability, and real-time performance [111], [199], [294], [306]. All of these limitations may heavily compromise their full potential performance if deployed in highly dynamic environments or for complex tasks. LLM-based EI offers a transformative approach by enabling on-device inference and fostering real-time decision-making capabilities [15]. That means handling complex situations in much more refined ways and with much efficiency [79]. For example, in autonomous vehicles, LLMs combined with EI could significantly improve real-time scene understanding and decision-making [262], leading to safer and more adaptable navigation [271], especially in unpredictable environments where traditional AI approaches might struggle [263]. This enhanced capability to explain and adapt would allow autonomous vehicles to react effectively to changing road conditions and provide clear explanations for their actions to human users [262]. This represents a significant leap forward in the development of fully autonomous and trustworthy transportation systems.

7) CONTINUAL ENHANCEMENT IN THE SECURITY OF LLM-BASED EI

Security within LLM-based EI, as with any application, is not a single accomplishment but an ongoing commitment. Furthermore, relying solely on one security measure may not be adequate to defend against all potential threats [20], [53], [351]. Robust security in LLM-based EI demands constant, iterative improvement, and commitment to security practices across the entire system life-cycle stages. No single security measure can, on its own, be adequate to counter the dynamic possibilities of threat ranging from adversarial attacks aimed at manipulating LLM outputs [333] to data breaches in which sensitive information used either in development or operation of the system is compromised [147]. A multi-layered approach that incorporates a diverse set of security techniques (proactive, reactive, and post-hoc) is crucial [341], [342], [369], [374]. Furthermore, integrating security considerations from the very beginning of the development process, not as an afterthought, is paramount. This ensures that security measures are woven into the fabric of the LLM-based EI system, becoming an inherent part of its design and functionality [362], [391]. Security considerations should encompass not only the LLM itself but also the data used to train it, the communication protocols employed for edge device interaction, and the overall system architecture [141], [289].

8) BUILDING TRUSTWORTHY LLM-BASED EI

Building trustworthy LLM-based EI necessitates a foundation of trustworthy design principles that consider all aspects of the system, from development to deployment [19]. Incorporating responsible AI principles throughout the development lifecycle is crucial [393], [415]. This includes considerations like fairness, accountability, and bias mitigation to ensure that LLM-based edge applications are used ethically and responsibly [407], [408], [412], [416], [425]. When designing the explanation interface, user needs and cognitive limitations should be prioritized. Visual explanations, interactive elements, and explanations tailored to the user's expertise can all contribute to a more trustworthy experience [19]. Techniques that illuminate the reasoning behind LLM outputs and mitigate potential biases are crucial for fostering user trust and safety in edge-based applications [411], [413], [417]. By adhering to these principles and fostering a culture of continuous improvement, developers can create LLM-based EI systems that are not only functionally sophisticated but also trustworthy and responsible, earning and maintaining user trust in the long run.

B. OPEN ISSUES AND CURRENT CHALLENGES

Although LLM-based EI shows significant potential, its effective implementation in real-world scenarios requires

addressing numerous critical unresolved issues and present challenges. Overcoming these hurdles is essential for building trust in LLM outputs, ensuring responsible development and deployment, and ultimately, unlocking the full potential of LLM-powered intelligent edge devices.

1) THE HUMAN-IN-THE-LOOP PARADOX

Defining the ideal role for humans to interact with these advanced AI systems poses a significant challenge in LLM-based EI. While LLMs offer the potential for autonomous decision-making at the edge, complete reliance on them raises safety and security concerns [20], [53], [438]. Conversely, excessive human intervention can hinder the efficiency and autonomy of LLMs [72]. LLMs excel at automating tasks [170], but human oversight might still be necessary for critical decision-making at the edge [266], [362]. The crux of the challenge lies in making fine balancing between allowing autonomy to LLMs for routine decision-making and having integral human oversight, especially in vital decision-making at edge scenarios, such as autonomous driving [262] and critical infrastructures [52]. Although LLMs readily automate a task, knowing how to use large reams of data to analyze them for the best output, preservation of human judgment and expertise remains invaluable in ethical considerations, handling unexpected situations, and assurance of the overall safety and security of the system [20]. This will need the development of intuitive human-machine interfaces at the edge and decision-support systems that will enable seamless human intervention: empower humans to guide, correct mistakes, or even arbitrate–override LLM decisions in critical situations. By cultivating collaborative effort between human and LLM capabilities, designers can provide for a strong and reliable system that deploys the strengths of both, ultimately leading to more efficient, responsible, and effective AI at the edge.

2) EDGE SECURITY CHALLENGES

Besides the threats that we have discussed in the previous sections, LLMs at the edge face other challenges, this includes:

- *Security in Uncontrolled Environments:* Zero-trust security is a model that assumes no user, device, or service within a network is inherently trustworthy [439]. This necessitates continuous verification and authorization for all entities attempting to access resources on the network [440]. In the context of LLM-based EI, where edge devices operate in potentially uncontrolled environments [180], a zero-trust approach becomes even more critical for mitigating security risks. Ensuring security in scenarios with untrusted edge devices or unreliable communication channels remains an open challenge [87]. Malicious actors might compromise edge devices to inject misleading prompts [332] or manipulate the LLM's training data [337]. Additionally, unreliable communication channels could lead to data breaches or manipulation during transmission between edge devices and the cloud [101], potentially undermining the effectiveness of real-time security. Adopting a zero-trust security posture in LLM-based EI can build a more robust defense against these evolving threats in uncontrolled environments. This ensures that even in untrusted settings, the integrity and reliability of the system are maintained [439], [440].
- *Security Implications of Optimized Models:* Optimizing LLMs for the edge through techniques like quantization and pruning offers a double-edged sword. While it improves efficiency, it can introduce unintended security concerns and reduce trustworthiness [441]. Also, unlike the cloud environment, where extensive security measures can be implemented, resource-constrained edge devices present unique challenges. Deploying all the robust defense techniques used in the cloud,[6] might not be feasible on edge devices due to their limited processing power and battery life. A possible solution is identifying the most critical security threats for a specific edge application and deploying lightweight defense mechanisms tailored to address those threats. For instance, focusing on anomaly detection for specific types of sensor data relevant to the application rather than a general-purpose solution. Another solution entails leveraging the cloud for certain security tasks that are computationally expensive on the edge. This could involve periodically sending anonymized LLM outputs or model updates to the cloud for security analysis while keeping real-time inference tasks localized on the edge device.
- *The Fragile Chain of Trust:* LLM-based EI often relies on a complex ecosystem of interconnected devices and services [36], [55], [56], [72]. A single point of failure, whether a compromised edge device, a vulnerable plug-in, a security breach in communication channels, or a bias within the training data, can shatter the entire chain of trust [328], [341], [350]. Developing frameworks for distributed trust management and self-healing mechanisms within the edge ecosystem will be crucial for ensuring long-term reliability and resilience [171], [289].
- *Securing the LLM-EI Supply Chain:* The security of LLM-based IE depends on the reliability of the underlying designs [53]. However, securing the entire development cycle, from acquisition of training data to deployment at the edge, represents a major challenge [343]. Malicious actors may attempt to compromise the fairness and integrity of LLM results by various means [37], including data poisoning or model manipulation. Mitigating these risks requires the implementation of robust security measures in the entire supply chain. These can include using secure coding practices throughout development, implementing tamper-proof data storage solutions [50], and carrying

[6] https://trust.openai.com/

out regular audits of the development process to identify and address potential vulnerabilities before deployment.

- *Safeguarding Edge Devices from Physical Threats:* Since LLM-based EI systems are mostly edge devices that will be deployed in very different kinds of physically divergent environments, their physical security becomes paramount. These devices could be vulnerable to tampering or unauthorized access, potentially leading to data theft (attackers stealing sensitive data used by the LLM or stored locally on the device), model tampering (attackers modifying the LLM model to disrupt functionality or compromise outputs), and operational disruption (physical attacks causing service outages or performance degradation). Strategies for mitigating these risks could focus on tamper-evident seals for edge devices, secure boot procedures guaranteeing that only authorized code will be run on startup, and deployment of devices in a physical security location ensuring access control.
- *Edge Orchestration Security:* According to [148], orchestration in fog computing refers to dealing with the whole service lifecycle. It ensures that services comply with user requirements and Service Level Agreements (SLA) by continuously monitoring the infrastructure itself, detect any changes, and adhering to privacy and security rules. This kind of central control keeps fog computing efficient, adaptable, and secure. Fog orchestration offers a compelling paradigm for managing distributed computing resources at the network's edge. However, ensuring a robust security and privacy posture within this dynamic environment remains a challenge. For instance, Fog orchestration strategies often rely on service reallocation, replication, and migration to address user mobility, fault tolerance, and service availability [148], [442]. While all such techniques bring flexibility benefits to the offered services, they can potentially add new vulnerabilities [443]. Since data may travel across the Internet through many network hops, or be replicated/shared by a large number of devices, the potential for unauthorized access or misuse is greatly increased [148], [443], [444]. In edge orchestration security, different approaches were proposed [148]. For instance, in the Neural Pub/Sub paradigm proposed in [67], inference-based filtering works as a privacy guarantor by making sure that only the essential data required to fulfill subscriber requests is disseminated. This feature aligns perfectly with the privacy challenges inherent in distributed learning across the computing continuum, in which situations related to data security and control become paramount.
- *Securing Communication Channels in Evolving Networks:* The challenge of securing communication channels in the dynamic network environments envisaged for 6G represents a unique challenge for LLM-based EI. As opposed to traditional centralized network architectures, 6G promises a more fluid, distributed topology [59], [60]. Devices will constantly switch between access points, potentially self-organizing [445] into temporary networks [446]. Such dynamic nature makes it difficult to establish and maintain secure communication channels. Conventional security solutions based on predefined trust relationships may not be easily adaptable, especially with the presence of quantum computers in the 6G era [71]. Furthermore, the sheer number of devices in a 6G network creates a larger attack surface. Malicious actors could exploit vulnerabilities in individual peripheral devices, access points or network protocols to intercept or manipulate communications between peripheral devices and the central platform. To address these challenges, a multi-pronged approach is necessary. For instance, network slicing enables the creation of virtualized network segments dedicated solely to LLM-based EI communication [60]. This helps to isolate it from other network traffic and reduce the overall attack surface [71]. Moreover, implementing secure enclaves within edge devices using technologies such as Intel SGX [447] can provide a trusted execution environment for sensitive communication tasks, further enhancing security. In addition, blockchain technology can provide a powerful boost to trust management in dynamic networks [60], [448]. Relying on a distributed ledger system, it is possible to establish secure and verifiable communication channels without relying on a central authority. By integrating blockchain into existing network security protocols, such as Secure Sockets Layer (SSL/TLS), a strong and adaptable security framework can be created for LLM-based EI communication at the edge.

3) LIGHTWEIGHT EXPLAINABILITY

While techniques of Explainable AI bring valuable insight into the decision-making processes of LLMs [394], ensuring interpretability on resource-constrained edge devices has been a challenging task [449]. With low processing power and memory, edge devices require a balancing act in ensuring explainability that can engender trust and user understanding at a level that will not compromise real-time performance [407]. Most computationally intensive XAI techniques–which are usually deployed in the cloud–would significantly impede real-time inference tasks at edge locations, thereby introducing latency [359]. It is hence incumbent that lightweight XAI techniques be specifically developed and deployed for edge-based LLM deployments, with a focus on techniques providing actionable insights into the decisions made by LLMs and empowering users to understand how and why a model produced a particular output. This should, however, be done without loading a device for which computational resources are at a premium. A lightweight approach to explainability may focus on techniques that, therefore, only make high-level explanations and avoid too granular pieces of information. Another could

be the methods leveraging model compression to reduce the relative footprint of the XAI component itself. Lightweight explainability brings together user understanding and real-time performance in an edge-based scenario. Maintaining the former will enable the user to derive very valuable lessons from how an LLM reasons, engendering trust and responsible AI practices without giving up, at the same time, efficiency and responsiveness–critical for edge-based applications to succeed.

4) OVERESTIMATION OF LLM CAPABILITIES: LONG-TERM TIME SERIES FORECASTING (LTSF) FOR EXAMPLE

Given the growing interest in applying LLMs to a wide array of domains, high expectations have been put into their potential for revolutionizing respective fields. However, recent studies noted that some of these expectations may not be entirely justified, specifically, in Long-Term Time Series Forecasting (LTSF) [450], [451]. One notable research from [450] has provided evidence that challenges the assumption that LLMs are superior to traditional models for time series tasks such as forecasting. For instance, the study put in test various Transformer-based LTSF models against simple one-layer linear models, and found that the latter consistently outperformed LLMs in long-term time series forecasting across multiple datasets. Another study in [451] showed that removing the LLM component from time series forecasting models, or replacing it with a basic attention layer, often led to improved performance. This finding directly goes against expectations whereby LLMs, with their complex architectures and vast pre-training, would bring enhancements to time series forecasting. On the contrary, much simpler models performed better and reduced computational costs significantly [451]–a consideration that, for edge deployments where resources are limited, becomes considerably critical. Given these findings, it is important to temper the expectations surrounding the use of LLMs in edge-based time series analysis. While LLMs can excel in tasks requiring natural language understanding or some level of reasoning, their application to time series forecasting may introduce unnecessary computational overhead without delivering corresponding benefits.

C. FUTURE RESEARCH DIRECTIONS AND OPPORTUNITIES

LLM-based EI presents a multitude of promising research avenues and opportunities. By harnessing the potential of this technology, we can unlock a future filled with intelligent edge devices seamlessly integrated into our lives. Here, we explore some key areas for future exploration:

1) CO-DESIGNING HARDWARE AND SOFTWARE

Collaborative efforts between hardware and software engineers hold immense promise for the future of edge AI. We can create specialized edge computing platforms optimized for LLM workloads by fostering a co-design approach. Such a co-design approach will be able to unlock all potential brought to edges by LLMs. This will help in the design of specialized edge computing platforms where careful tuning with LLM workloads will be performed. These platforms would be optimized for edge LLMs, in particular, low-power, low-cost-advanced methods for efficient memory access. An ecosystem is envisioned that includes more powerful and efficient applications of LLM-based EI, able to continue to push the frontiers of what can be realized within edge computing. These visions no longer sit in the distant future. Leading industrial organizations, such as NVIDIA [30] and Qualcomm [31], have already begun considering co-design approaches for hardware-software optimization in edge AI. Their pioneering efforts open the future to when LLMs can really thrive at the edge, transforming a wide variety of industries and applications. By embraced co-design and deeper integration with hardware and software engineers, we have the potential with LLM-based EI at the edge to achieve a paradigm shift to make edge computing intelligent and efficient.

2) EXPLORING MULTIMODALITY WHILE ENSURING EDGE COMPATIBILITY

The future of LLMs is likely to involve a more symbiotic relationship with advancements in computer vision. Vision transformers (ViTs), which excel at image recognition using transformer architectures, present a unique opportunity for synergistic development [452]. Scaling the parameters of ViTs, as exemplified by Google Research's achievement of a staggering 22 billion parameter model (ViT-22B), promises significant advancements [453]. When integrated with LLMs, this could lead to the development of truly multimodal models capable of understanding and processing information from both text and images. Such models have the potential to unlock groundbreaking applications in image captioning, visual question answering, and even creative tasks that combine language and imagery. However, it is crucial to acknowledge the inherent trade-off between model complexity and computational efficiency. Pushing the boundaries of model size, as with ViT-22B, necessitates significant computational resources. To bridge this gap and enable real-world deployment, research in efficient algorithms and hardware architectures specifically designed for edge computing will be paramount. This focus on efficient utilization of resources will be essential to unlock the full potential of multimodal LLMs and ensure their applicability beyond powerful research machines.

3) APPLICATIONS AND USE CASES

Exploring novel applications for LLM-based EI holds immense promise. Research into these and other use cases will further expand the reach and impact of LLM-based EI. Here are a few examples:

- *Personalized Edge Assistants:* This entails intelligent assistants that leverage LLMs to provide context-aware

interactions, anticipating user needs and responding proactively in a personalized manner [454].

- *Real-time Anomaly Detection:* Edge-based LLMs can be engaged in predictive maintenance management by mitigating anomalies, making informed decisions, and handling probable failures or threats in real-time [88], which improves overall reliability and efficiency in industrial operation.
- *Augmented Reality (AR) with Context Recognition:* Edge-based LLMs could analyze the user's environment through AR glasses, providing real-time information or overlays tailored to the specific context, enriching the AR experience [252].
- Intelligent Warehouses and Streamlined Inventory Management: Retail warehouse environments can benefit significantly from a system integrating ViT-powered cameras and edge-optimized LLMs. Shop owners can deploy these systems throughout the facility. By simply pointing a camera at a product, the ViT instantly identifies it. The LLM then seamlessly accesses relevant data (stock levels, product information, reports) and interacts with the owner. This allows for efficient querying, eliminating the need for manual data retrieval and report generation [455]. Owners can ask questions like "Identify all near-expiry items in warehouse section B" or "Generate a real-time report on the top-selling products this quarter." This streamlines inventory management, allowing owners to focus on strategic decisions with readily available, up-to-date information. Additionally, the system can be extended to in-store customer interactions, providing instant product information or personalized recommendations by simply pointing their phones at items. This creates a more interactive and data-driven shopping experience.
- Personalized Customer Recommendations with Enhanced Context: Traditional recommendation systems in retail often rely on historical purchase data, potentially missing out on evolving customer preferences [456]. Here, LLMs trained on product descriptions, reviews, user demographics, and even visual data can offer a more nuanced approach [259], [457], [458]. When a customer enters a store or browses an online catalog, the LLM analyzes their past purchases and current browsing patterns. This allows it to recommend not just similar products, but also complementary items they might not have considered [456]. Integrating ViT further personalizes recommendations based on visual information. For instance, a customer looking at a specific dress could be shown visually complementary accessories or a complete outfit recommendation. This not only enhances the customer experience but also increases sales opportunities for shop owners by leveraging the power of data-driven suggestions.

X. CONCLUSION

This survey has mapped out the dynamic landscape of LLM-based Edge Intelligence, providing a consolidated resource for researchers and practitioners. We delved into cutting-edge LLM-based EI architectures, analyzed optimization techniques for resource-constrained environments, and showcased diverse real-world applications. Recognizing security concerns, we have explored potential vulnerabilities and corresponding defense mechanisms. Finally, we have emphasized the importance of trustworthiness in LLM-based EI development. This holistic examination fosters a foundation for the responsible deployment of LLM-powered EI, paving the way for transformative applications and a future of intelligent edge devices. Continuing explorations in this field can pave the way to unlock the complete potential of LLM-based EI and shape a future characterized by intelligent and secure edge computing.

REFERENCES

[1] S. Pinker, *The Language Instinct: How the Mind Creates Language*. London, U.K: Penguin, 2003.

[2] A. M. Turing, *Computing Machinery and Intelligence*. Dordrecht, The Netherlands: Springer, 2009.

[3] W. X. Zhao et al., "A survey of large language models," 2023, *arXiv:2303.18223*.

[4] E. M. Bender and A. Koller, "Climbing towards NLU: On meaning, form, and understanding in the age of data," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguist.*, 2020, pp. 5185–5198.

[5] F. Jelinek, *Statistical Methods for Speech Recognition*. Cambridge, MA, USA: MIT Press, 1998.

[6] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13, 2000, pp. 1–11.

[7] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Interspeech*, vol. 2, 2010, pp. 1045–1048.

[8] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–15.

[9] J. Sarzynska-Wawer et al., "Detecting formal thought disorder by deep contextualized word representations," *Psychiat. Res.*, vol. 304, Oct. 2021, Art. no. 114135.

[10] B. Min et al., "Recent advances in natural language processing via large pre-trained language models: A survey," *ACM Comput. Surveys*, vol. 56, no. 2, pp. 1–40, 2023.

[11] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training." 2018. [Online]. Available: https://paperswithcode.com/paper/improving-language-understanding-by

[12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[14] J. Wei et al., "Emergent abilities of large language models," 2022, *arXiv:2206.07682*.

[15] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[16] F. Petroni et al., "Language models as knowledge bases?" 2019, *arXiv:1909.01066*.

[17] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proc. ACM Conf. Fairness, Accountability, Transparency*, 2021, pp. 610–623.

[18] Y. Chen, Y. Yan, Q. Yang, Y. Shu, S. He, and J. Chen, "Confidant: Customizing transformer-based LLMs via collaborative edge training," 2023, *arXiv:2311.13381*.

[19] L. Sun et al., "TrustLLM: Trustworthiness in large language models," 2024, *arXiv:2401.05561*.

[20] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly," *High-Confidence Comput.*, vol. 4, no. 2, 2024, Art. no. 100211.

[21] E. Parliament, "Artificial intelligence act," Accessed: Jun. 10, 2024. [Online]. Available: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

[22] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016.

[23] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "GPTQ: Accurate post-training quantization for generative pre-trained transformers," 2022, *arXiv:2210.17323*.

[24] J. Lin, J. Tang, H. Tang, S. Yang, X. Dang, and S. Han, "AWQ: Activation-aware weight quantization for LLM compression and acceleration," 2023, *arXiv:2306.00978*.

[25] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han, "SmoothQuant: Accurate and efficient post-training quantization for large language models," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 38087–38099.

[26] H. Wang et al., "BitNet: Scaling 1-bit transformers for large language models," 2023, *arXiv:2310.11453*.

[27] W. Huang et al., "BiLLM: Pushing the limit of post-training quantization for LLMs," 2024, *arXiv:2402.04291*.

[28] V. Egiazarian, A. Panferov, D. Kuznedelev, E. Frantar, A. Babenko, and D. Alistarh, "Extreme compression of large language models via additive quantization," 2024, *arXiv:2401.06118*.

[29] H. Shen, H. Chang, B. Dong, Y. Luo, and H. Meng, "Efficient LLM inference on CPUS," 2023, *arXiv:2311.00502*.

[30] N. Nelson, S. Huver, and M. Toloui, "Deploy large language models at the edge with NVIDIA IGX Orin developer kit," 2023. Accessed: Mar. 17, 2024. [Online]. Available: https://developer.nvidia.com/blog/deploy-large-language-models-at-the-edge-with-nvidia-igx-orin-developer-kit/

[31] J. Soriaga, "Accelerating generative AI at the edge," 2023. Accessed: Mar. 17, 2024. [Online]. Available: https://www.qualcomm.com/news/onq/2023/11/accelerating-generative-ai-at-the-edge,

[32] S. Ma et al., "The era of 1-bit LLMs: All large language models are in 1.58 bits," 2024, *arXiv:2402.17764*.

[33] J. Wang et al., "Network meets ChatGPT: Intent autonomous management, control and operation," *J. Commun. Inf. Netw.*, vol. 8, no. 3, pp. 239–255, 2023.

[34] R. Yi, L. Guo, S. Wei, A. Zhou, S. Wang, and M. Xu, "EdgeMoE: Fast on-device inference of MoE-based large language models," 2023, *arXiv:2308.14352*.

[35] L. Dong et al., "LAMBO: Large language model empowered edge intelligence," 2023, *arXiv:2308.15078*.

[36] H. Woisetschläger, A. Isenko, S. Wang, R. Mayer, and H.-A. Jacobsen, "Federated fine-tuning of LLMS on the very edge: The good, the bad, the ugly," 2023, *arXiv:2310.03150*.

[37] F. Jiang et al., "Identifying and mitigating vulnerabilities in LLM-integrated applications," 2023, *arXiv:2311.16153*.

[38] C. Liang, S. Zuo, Q. Zhang, P. He, W. Chen, and T. Zhao, "Less is more: Task-aware layer-wise distillation for language model compression," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 20852–20867.

[39] T. Fan et al., "FATE-LLM: A industrial grade federated learning framework for large language models," 2023, *arXiv:2310.10049*.

[40] C. Packer, V. Fang, S. G. Patil, K. Lin, S. Wooders, and J. E. Gonzalez, "MemGPT: Towards LLMS as operating systems," 2023, *arXiv:2310.08560*.

[41] M. Sun, Z. Liu, A. Bair, and J. Z. Kolter, "A simple and effective pruning approach for large language models," 2023, *arXiv:2306.11695*.

[42] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019.

[43] J. Zhang and K. B. Letaief, "Mobile edge intelligence and computing for the Internet of Vehicles," *Proc. IEEE*, vol. 108, no. 2, pp. 246–261, Feb. 2020.

[44] X. Wang, Y. Han, V. C. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 869–904, 2nd Quart., 2020.

[45] D. Xu et al., "Edge intelligence: Empowering intelligence to the edge of network," *Proc. IEEE*, vol. 109, no. 11, pp. 1778–1837, Nov. 2021.

[46] W. Y. B. Lim et al., "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 2031–2063, 3rd Quart., 2020.

[47] X. Shen, J. Gao, W. Wu, M. Li, C. Zhou, and W. Zhuang, "Holistic network virtualization and pervasive network intelligence for 6G," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 1–30, 1st Quart., 2022.

[48] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, Jan. 2022.

[49] Y. Li, Y. Yu, W. Susilo, Z. Hong, and M. Guizani, "Security and privacy for edge intelligence in 5G and beyond networks: Challenges and solutions," *IEEE Wireless Commun.*, vol. 28, no. 2, pp. 63–69, Apr. 2021.

[50] X. Wang, X. Ren, C. Qiu, Z. Xiong, H. Yao, and V. C. Leung, "Integrating edge intelligence and blockchain: What, why, and how," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 4, pp. 2193–2229, 4th Quart., 2022.

[51] M. Xu et al., "Unleashing the power of edge-cloud generative AI in mobile networks: A survey of AIGC services," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 2, pp. 1127–1170, 2nd Quart., 2024.

[52] M. Xu et al., "A survey of resource-efficient LLM and multimodal foundation models," 2024, *arXiv:2401.08092*.

[53] B. C. Das, M. H. Amini, and Y. Wu, "Security and privacy challenges of large language models: A survey," 2024, *arXiv:2402.00888*.

[54] S. Minaee et al., "Large language models: A survey," 2024, *arXiv:2402.06196*.

[55] Z. Lin, G. Qu, Q. Chen, X. Chen, Z. Chen, and K. Huang, "Pushing large language models to the 6G edge: Vision, challenges, and opportunities," 2023, *arXiv:2309.16739*.

[56] H. Du et al., "Enabling AI-generated content (AIGC) services in wireless edge networks," 2023, *arXiv:2301.03220*.

[57] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[58] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proc. IEEE*, vol. 107, no. 8, pp. 1655–1674, Aug. 2019.

[59] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proc. IEEE*, vol. 107, no. 11, pp. 2204–2239, Nov. 2019.

[60] E. Peltonen et al., "6G white paper on edge intelligence," 2020, *arXiv:2004.14850*.

[61] L. Lovén et al., "EdgeAI: A vision for distributed, edge-native artificial intelligence in future 6G networks," in *Proc. 6G Wireless Summit*, Mar. 2019, Levi, Finland, pp. 1–2.

[62] S. Liu, L. Liu, J. Tang, B. Yu, Y. Wang, and W. Shi, "Edge computing for autonomous driving: Opportunities and challenges," *Proc. IEEE*, vol. 107, no. 8, pp. 1697–1716, Aug. 2019.

[63] N. Parvaresh and B. Kantarci, "A continuous actor–critic deep Q-learning-enabled deployment of UAV base stations: Toward 6G small cells in the skies of smart cities," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 700–712, 2023.

[64] "Defining AI native: A key enabler for advanced intelligent telecom networks." Ericsson. 2024. Accessed: Jun. 24, 2024. [Online]. Available: https://www.ericsson.com/49341a/assets/local/reports-papers/white-papers/ai-native.pdf

[65] "Toward a 6G AI-native air interface," Nokia Bell Labs, Murray Hill, NJ, USA, White Paper, 2021. Accessed: Jun. 24, 2024. [Online]. Available: https://onestore.nokia.com/asset/210299

[66] M. Chen et al., "Distributed learning in wireless networks: Recent progress and future challenges," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3579–3605, Dec. 2021.

[67] L. Lovén, R. Morabito, A. Kumar, S. Pirttikangas, J. Riekki, and S. Tarkoma, "How can AI be distributed in the computing continuum? Introducing the neural pub/sub paradigm," 2023, *arXiv:2309.02058*.

[68] O. Friha, M. A. Ferrag, L. Shu, and M. Nafa, "A robust security framework based on blockchain and SDN for fog computing enabled agricultural Internet of Things," in *Proc. Int. Conf. Internet Things Intell. Appl. (ITIA)*, 2020, pp. 1–5.

[69] B. Mao, J. Liu, Y. Wu, and N. Kato, "Security and privacy on 6G network edge: A survey," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 1095–1127, 2nd Quart., 2023.

[70] R. Gupta, D. Reebadiya, S. Tanwar, N. Kumar, and M. Guizani, "When blockchain meets edge intelligence: Trusted and security solutions for consumers," *IEEE Netw.*, vol. 35, no. 5, pp. 272–278, Sep./Oct. 2021.

[71] Y. Siriwardhana, P. Porambage, M. Liyanage, and M. Ylianttila, "AI and 6G security: Opportunities and challenges," in *Proc. Joint Eur. Conf. Netw. Commun. 6G Summit (EuCNC/6G Summit)*, 2021, pp. 616–621.

[72] Y. Shen et al., "Large language models empowered autonomous edge AI for connected intelligence," *IEEE Commun. Mag.*, early access, Jan. 8, 2024, doi: 10.1109/MCOM.001.2300550.

[73] R. Singh and S. S. Gill, "Edge AI: A survey," *Internet Things Cyber-Phys. Syst.*, vol. 3, pp. 71–92, Mar. 2023.

[74] E. Li, L. Zeng, Z. Zhou, and X. Chen, "Edge AI: On-demand accelerating deep neural network inference via edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 447–457, Jan. 2020.

[75] M. Merenda, C. Porcaro, and D. Iero, "Edge machine learning for ai-enabled IoT devices: A review," *Sensors*, vol. 20, no. 9, p. 2533, 2020.

[76] "EdgeTPU." Google. Accessed: Jul. 6, 2024. [Online]. Available: https://cloud.google.com/edge-tpu

[77] "Explore what's next in embedded computing." NVIDIA. Accessed: Jul. 6, 2024. [Online]. Available: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/

[78] M. W. U. Rahman et al., "Quantized transformer language model implementations on edge devices," 2023, *arXiv:2310.03971*.

[79] H. Touvron et al., "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[80] M. Abdin et al., "Phi-3 technical report: A highly capable language model locally on your phone," 2024, *arXiv:2404.14219*.

[81] J. Zidar, T. Matić, I. Aleksi, and Ž. Hocenski, "Dynamic voltage and frequency scaling as a method for reducing energy consumption in ultra-low-power embedded systems," *Electronics*, vol. 13, no. 5, p. 826, 2024.

[82] J. Cha and S. Kim, "CNN hardware accelerator architecture design for energy-efficient AI," in *Artificial Intelligence and Hardware Accelerators*. Cham, Switzerland: Springer, 2023, pp. 319–357.

[83] P. Rech, "Artificial neural networks for space and safety-critical applications: Reliability issues and potential solutions," *IEEE Trans. Nucl. Sci.*, vol. 71, no. 4, pp. 377–404, Apr. 2024.

[84] F. Bang, "GPTCache: An open-source semantic cache for LLM applications enabling faster answers and cost savings," in *Proc. 3rd Workshop Natural Lang. Process. Open Source Softw. (NLP-OSS)*, 2023, pp. 212–218.

[85] "Differences between MTBF and MTTR," IBM. Accessed: Jul. 6, 2024. [Online]. Available: https://www.ibm.com/think/topics/mttr-vs-mtbf,

[86] G. K. Mahato and S. K. Chakraborty, "Securing edge computing using cryptographic schemes: A review," *Multimedia Tools Appl.*, vol. 83, no. 12, pp. 34825–34848, 2024.

[87] O. Friha, M. A. Ferrag, M. Benbouzid, T. Berghout, B. Kantarci, and K.-K. R. Choo, "2DF-IDS: Decentralized and differentially private federated learning-based intrusion detection system for industrial IoT," *Comput. Security*, vol. 127, Apr. 2023, Art. no. 103097.

[88] M. A. Ferrag et al., "Revolutionizing cyber threat detection with large language models: A privacy-preserving BERT-based lightweight model for IoT/IIoT devices," *IEEE Access*, vol. 12, pp. 23733–23750, 2024.

[89] G. Bai et al., "Beyond efficiency: A systematic survey of resource-efficient large language models," 2024, *arXiv:2401.00625*.

[90] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, "Pruning and quantization for deep neural network acceleration: A survey," *Neurocomputing*, vol. 461, pp. 370–403, Oct. 2021.

[91] "Introducing apple's on-device and server foundation models." Apple. Accessed: Jun. 18, 2024. [Online]. Available: https://machinelearning.apple.com/research/introducing-apple-foundation-models

[92] E. C. Strinati and S. Barbarossa, "6G networks: Beyond Shannon towards semantic and goal-oriented communications," *Comput. Netw.*, vol. 190, May 2021, Art. no. 107930.

[93] G. Shi, Y. Xiao, Y. Li, and X. Xie, "From semantic communication to semantic-aware networking: Model, architecture, and open problems," *IEEE Commun. Mag.*, vol. 59, no. 8, pp. 44–50, Aug. 2021.

[94] J. Park, S.-W. Ko, J. Choi, S.-L. Kim, and M. Bennis, "Towards semantic communication protocols for 6g: From protocol learning to language-oriented approaches," 2023, *arXiv:2310.09506*.

[95] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," *IEEE Wireless Commun.*, vol. 29, no. 1, pp. 210–219, Feb. 2022.

[96] F. Jiang et al., "Large AI model empowered multimodal semantic communications," 2023, *arXiv:2309.01249*.

[97] H. Zhang, N. Liu, X. Chu, K. Long, A.-H. Aghvami, and V. C. Leung, "Network slicing based 5G and future mobile networks: Mobility, resource management, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 138–145, Aug. 2017.

[98] J. Requeima, J. Bronskill, D. Choi, R. E. Turner, and D. Duvenaud, "LLM processes: Numerical predictive distributions conditioned on natural language," 2024, *arXiv:2405.12856*.

[99] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto, "Benchmarking large language models for news summarization," *Trans. Assoc. Comput. Linguist.*, vol. 12, pp. 39–57, Jan. 2024.

[100] O. Friha, M. A. Ferrag, L. Shu, L. Maglaras, K.-K. R. Choo, and M. Nafaa, "FELIDS: Federated learning-based intrusion detection system for agricultural Internet of Things," *J. Parallel Distrib. Comput.*, vol. 165, pp. 17–31, Jul. 2022.

[101] S. Deng, H. Zhao, W. Fang, J. Yin, S. Dustdar, and A. Y. Zomaya, "Edge intelligence: The confluence of edge computing and artificial intelligence," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7457–7469, Aug. 2020.

[102] S. Dang, O. Amin, B. Shihada, and M.-S. Alouini, "What should 6G be?" *Nat. Electron.*, vol. 3, no. 1, pp. 20–29, 2020.

[103] P. Yang, L. Kong, and G. Chen, "Spectrum sharing for 5G/6G URLLC: Research frontiers and standards," *IEEE Commun. Stand. Mag.*, vol. 5, no. 2, pp. 120–125, Jun. 2021.

[104] C. Han, Y. Wu, Z. Chen, and X. Wang, "Terahertz communications (TeraCom): Challenges and impact on 6G wireless systems," 2019, *arXiv:1912.06040*.

[105] T. S. Rappaport et al., "Wireless communications and applications above 100 GHz: Opportunities and challenges for 6G and beyond," *IEEE Access*, vol. 7, pp. 78729–78757, 2019.

[106] M. Xu et al., "When large language model agents meet 6G networks: Perception, grounding, and alignment," 2024, *arXiv:2401.07764*.

[107] S. Javaid, R. A. Khalil, N. Saeed, B. He, and M.-S. Alouini, "Leveraging large language models for integrated satellite-aerial-terrestrial networks: Recent advances and future directions," 2024, *arXiv:2407.04581*.

[108] H. Zhou, C. Hu, and X. Liu, "An overview of machine learning-enabled optimization for reconfigurable intelligent surfaces-aided 6G networks: From reinforcement learning to large language models," 2024, *arXiv:2405.17439*.

[109] B. Rong and H. Rutagemwa, "Leveraging large language models for intelligent control of 6G integrated TN-NTN with IoT service," *IEEE Netw.*, vol. 38, no. 4, pp. 136–142, Jul. 2024.

[110] Q. Liu, J. Mu, D. ChenZhang, Y. Liu, and T. Hong, "LLM enhanced reconfigurable intelligent surface for energy-efficient and reliable 6G IoV," *IEEE Trans. Veh. Technol.*, early access, May 6, 2024, doi: 10.1109/TVT.2024.3395748.

[111] C.-X. Wang et al., "On the road to 6G: Visions, requirements, key technologies and testbeds," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 905–974, 2nd Quart., 2023.

[112] N. H. Mahmood, H. Alves, O. A. López, M. Shehab, D. P. M. Osorio, and M. Latva-Aho, "Six key features of machine type communication in 6G," in *Proc. 2nd 6G Wireless Summit (6G SUMMIT)*, 2020, pp. 1–5.

[113] N. H. Mahmood et al., "White paper on critical and massive machine type communication towards 6G," 2020, *arXiv:2004.14146*.

[114] N. H. Mahmood et al., "Machine type communications: Key drivers and enablers towards the 6G era," *EURASIP J. Wireless Commun. Netw.*, vol. 2021, no. 1, p. 134, 2021.

[115] Z. Zhang et al., "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 28–41, Sep. 2019.

[116] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May/Jun. 2020.

[117] O. Friha, M. A. Ferrag, L. Shu, L. Maglaras, and X. Wang, "Internet of Things for the future of smart agriculture: A comprehensive survey of emerging technologies," *IEEE/CAA J. Automatica Sinica*, vol. 8, no. 4, pp. 718–752, Apr. 2021.

[118] O. Friha, M. A. Ferrag, L. Maglaras, and L. Shu, "Digital agriculture security: Aspects, threats, mitigation strategies, and future trends," *IEEE Internet Things Mag.*, vol. 5, no. 3, pp. 82–90, Sep. 2022.

[119] N. Dhar, B. Deng, D. Lo, X. Wu, L. Zhao, and K. Suo, "An empirical analysis and resource footprint study of deploying large language models on edge devices," in *Proc. ACM Southeast Conf.*, 2024, pp. 69–76.

[120] Z. Tang et al., "FusionAI: Decentralized training and deploying LLMs with massive consumer-level GPUs," 2023, *arXiv:2309.01172*.

[121] L. Tong, Y. Li, and W. Gao, "A hierarchical edge cloud architecture for mobile computing," in *Proc. IEEE INFOCOM 35th Annu. IEEE Int. Conf. Comput. Commun.*, 2016, pp. 1–9.

[122] Z. Lin, G. Qu, X. Chen, and K. Huang, "Split learning in 6G edge networks," *IEEE Wireless Commun.*, vol. 31, no. 4, pp. 170–176, Aug. 2024.

[123] J. Zhang et al., "Towards building the federated GPT: Federated instruction tuning," 2023, *arXiv:2305.05644*.

[124] Y. Tian, Y. Wan, L. Lyu, D. Yao, H. Jin, and L. Sun, "FedBERT: When federated learning meets pre-training," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 4, pp. 1–26, 2022.

[125] R. Ye et al., "OpenFedLLM: Training large language models on decentralized private data via federated learning," 2024, *arXiv:2402.06954*.

[126] Y. Chen, R. Li, X. Yu, Z. Zhao, and H. Zhang, "Adaptive layer splitting for wireless LLM inference in edge computing: A model-based reinforcement learning approach," 2024, *arXiv:2406.02616*.

[127] O. U. Akgul et al., "6G function modularity: Benefits, challenges, and options," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2024, pp. 1–5.

[128] C. Lai, Z. Zhou, A. Poptani, and W. Zhang, "LCM: LLM-focused hybrid SPM-cache architecture with cache management for multi-core AI accelerators," in *Proc. 38th ACM Int. Conf. Supercomput.*, 2024, pp. 62–73.

[129] F. Feng, G. Liu, T. Bi, and T. Jiang, "Edge selective sharing for massive mobile video streaming with cross-layer optimization," *IEEE Trans. Mobile Comput.*, early access, May 17, 2024, doi: 10.1109/TMC.2024.3402237.

[130] A. Mekrache, A. Ksentini, and C. Verikoukis, "Intent-based management of next-generation networks: An LLM-centric approach," *IEEE Netw.*, early access, Jun. 27, 2024, doi: 10.1109/MNET.2024.3420120.

[131] I. Ara and B. Kelley, "Physical layer security for 6G: Toward achieving intelligent native security at layer-1," *IEEE Access*, vol. 12, pp. 82800–82824, 2024.

[132] T. Nguyen, H. Nguyen, A. Ijaz, S. Sheikhi, A. V. Vasilakos, and P. Kostakos, "Large language models in 6G security: Challenges and opportunities," 2024, *arXiv:2403.12239*.

[133] P. Porambage, G. Gür, D. P. M. Osorio, M. Livanage, and M. Ylianttila, "6G security challenges and potential solutions," in *Proc. Joint Eur. Conf. Netw. Commun. 6G Summit (EuCNC/6G Summit)*, 2021, pp. 622–627.

[134] A. Celik and A. M. Eltawil, "At the dawn of generative AI era: A tutorial-cum-survey on new frontiers in 6G wireless intelligence," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 2433–2489, 2024.

[135] X. Zhang, J. Zhang, B. Rekabdar, Y. Zhou, P. Wang, and K. Liu, "Dynamic and adaptive feature generation with LLM," 2024, *arXiv:2406.03505*.

[136] M. A. Akbar et al., "6GSoft: Software for edge-to-cloud continuum," 2024, *arXiv:2407.05963*.

[137] T. Nguyen, N. Tran, L. Loven, J. Partala, M.-T. Kechadi, and S. Pirttikangas, "Privacy-aware blockchain innovation for 6G: Challenges and opportunities," *Proc. 2nd 6G Wireless Summit (6G SUMMIT)*, 2020, pp. 1–5.

[138] T. Hewa, G. Gür, A. Kalla, M. Ylianttila, A. Bracken, and M. Liyanage, "The role of blockchain in 6G: Challenges, opportunities and research directions," in *Proc. 2nd 6G Wireless Summit (6G SUMMIT)*, 2020, pp. 1–5.

[139] X. Zuo et al., "Federated TrustChain: Blockchain-enhanced LLM training and unlearning," 2024, *arXiv:2406.04076*.

[140] J. G. M. Mboma, K. Lusala, M. Matalatala, O. T. Tshipata, P. S. Nzakuna, and D. T. Kazumba, "Integrating LLM with blockchain and IPFS to enhance academic diploma integrity," in *Proc. Int. Conf. Artif. Intell., Comput., Data Sci. Appl. (ACDSA)*, 2024, pp. 1–6.

[141] M. A. Ferrag, A. Battah, N. Tihanyi, M. Debbah, T. Lestable, and L. C. Cordeiro, "SecureFalcon: The next cyber reasoning system for cyber security," 2023, *arXiv:2307.06616*.

[142] Y. Chen et al., "NetGPT: A native-AI network architecture beyond provisioning personalized generative services," 2023, *arXiv:2307.06148*.

[143] Y. Wang, Y. Lin, X. Zeng, and G. Zhang, "PrivateLoRA for efficient privacy preserving LLM," 2023, *arXiv:2311.14030*.

[144] S. Kwon, S. Lee, T. Kim, D. Ryu, and J. Baik, "Exploring LLM-based automated repairing of Ansible script in edge-cloud infrastructures," *J. Web Eng.*, vol. 22, no. 6, pp. 889–912, 2023.

[145] N. Chen, Z. Cheng, X. Fan, X. Xia, and L. Huang, "Towards integrated fine-tuning and inference when generative AI meets edge intelligence," 2024, *arXiv:2401.02668*.

[146] P. Yang, N. Xiong, and J. Ren, "Data security and privacy protection for cloud storage: A survey," *IEEE Access*, vol. 8, pp. 131723–131740, 2020.

[147] "ChatGPT account takeover—Wildcard Web cache deception," Accessed: Jun. 11, 2024. [Online]. Available: https://nokline.github.io/bugbounty/2024/02/04/ChatGPT-ATO.html

[148] B. Costa, J. Bachiega Jr., L. R. de Carvalho, and A. P. Araujo, "Orchestration in fog computing: A comprehensive survey," *ACM Comput. Surveys*, vol. 55, no. 2, pp. 1–34, 2022.

[149] X. Masip, E. Marín, J. Garcia, and S. Sànchez, "Collaborative mechanism for hybrid fog-cloud scenarios," in *Fog Fogonomics: Challenges and Practices of Fog Computing, Communication, Networking, Strategy, and Economics*. Hoboken, NJ, USA: Wiley, 2020, pp. 7–60.

[150] S. V. Gogouvitis, H. Mueller, S. Premnadh, A. Seitz, and B. Bruegge, "Seamless computing in industrial systems using container orchestration," *Future Gener. Comput. Syst.*, vol. 109, pp. 678–688, Aug. 2020.

[151] Y. Jiang, Z. Huang, and D. H. K. Tsang, "Challenges and solutions in fog computing orchestration," *IEEE Netw.*, vol. 32, no. 3, pp. 122–129, May/Jun. 2017.

[152] F. Jalali et al., "Dynamic edge fabric environment: Seamless and automatic switching among resources at the edge of IoT network and cloud," in *Proc. IEEE Int. Conf. Edge Comput. (EDGE)*, 2019, pp. 77–86.

[153] F. Wamser et al., "Orchestration and monitoring in fog computing for personal edge cloud service support," in *Proc. IEEE Int. Symp. Local Metrop. Area Netw. (LANMAN)*, 2018, pp. 91–96.

[154] X. Yuan, W. Kong, Z. Luo, and M. Xu, "Efficient inference offloading for mixture-of-experts large language models in Internet of Medical Things," *Electronics*, vol. 13, no. 11, p. 2077, 2024.

[155] Z. Yang, Y. Yang, C. Zhao, Q. Guo, W. He, and W. Ji, "PerLLM: Personalized inference scheduling with edge-cloud collaboration for diverse LLM services," 2024, *arXiv:2405.14636*.

[156] J. Fang, Y. He, F. R. Yu, J. Li, and V. C. Leung, "Large language models (LLMs) inference offloading and resource allocation in cloud-edge networks: An active inference approach," in *Proc. IEEE 98th Veh. Technol. Conf. (VTC-Fall)*, 2023, pp. 1–5.

[157] C. Pahl, N. El Ioini, S. Helmer, and B. Lee, "An architecture pattern for trusted orchestration in IoT edge clouds," in *Proc. 3rd Int. Conf. Fog Mobile Edge Comput. (FMEC)*, 2018, pp. 63–70.

[158] L. Qian and J. Zhao, "User association and resource allocation in large language model based mobile edge computing system over wireless communications," 2023, *arXiv:2310.17872*.

[159] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.

[160] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Stat.*, 2017, pp. 1273–1282.

[161] D. Yao et al., "FedGKD: Towards heterogeneous federated learning via global knowledge distillation," *IEEE Trans. Comput.*, vol. 73, no. 1, pp. 3–17, Jan. 2024.

[162] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[163] N. Nascimento, P. Alencar, and D. Cowan, "Self-adaptive large language model (LLM)-based multiagent systems," in *Proc. IEEE Int. Conf. Autonomic Comput. Self-Organizing Syst. Compan. (ACSOS-C)*, 2023, pp. 104–109.

[164] T. Liang et al., "Encouraging divergent thinking in large language models through multi-agent debate," 2023, *arXiv:2305.19118*.

[165] S. P. Sharan, F. Pittaluga, M. Chandraker, and B. G. V. Kumar, "LLM-assist: Enhancing closed-loop planning with language-based reasoning," 2023, *arXiv:2401.00125*.

[166] C. H. Song, J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, and Y. Su, "LLM-Planner: Few-shot grounded planning for embodied agents with large language models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 2998–3009.

[167] J. Mao, Y. Qian, H. Zhao, and Y. Wang, "GPT-driver: Learning to drive with GPT," 2023, *arXiv:2310.01415*.

[168] I. Dasgupta et al., "Collaborating with language models for embodied reasoning," 2023, *arXiv:2302.00763*.

[169] J. Hong, S. Levine, and A. Dragan, "Zero-shot goal-directed dialogue via RL on imagined conversations," 2023, *arXiv:2311.05584*.

[170] C. Qian et al., "Communicative agents for software development," 2023, *arXiv:2307.07924*.

[171] T. He et al., "Unicron: Economizing self-healing LLM training at scale," 2023, *arXiv:2401.00134*.

[172] L. Lu, C. Dai, W. Tao, B. Yuan, Y. Sun, and P. Zhou, "Exploring the robustness of decentralized training for large language models," 2023, *arXiv:2312.00843*.

[173] B. Yuan et al., "Decentralized training of foundation models in heterogeneous environments," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 25464–25477.

[174] M. Ryabinin, T. Dettmers, M. Diskin, and A. Borzunov, "Swarm parallelism: Training large models can be surprisingly communication-efficient," 2023, *arXiv:2301.11913*.

[175] J. Wang et al., "CocktailSGD: Fine-tuning foundation models over 500Mbps networks," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 36058–36076.

[176] Y. Gao, Z. Song, and J. Yin, "GradientCoin: A peer-to-peer decentralized large language models," 2023, *arXiv:2308.10502*.

[177] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[178] Y. Gu, L. Dong, F. Wei, and M. Huang, "Knowledge distillation of large language models," 2023, *arXiv:2306.08543*.

[179] Q. Huang, Y. Wu, Z. Xing, H. Jiang, Y. Cheng, and H. Jin, "Adaptive intellect unleashed: The feasibility of knowledge transfer in large language models," 2023, *arXiv:2308.04788*.

[180] Y. Kim, S. Seo, J. Park, M. Bennis, S.-L. Kim, and J. Choi, "Knowledge distillation from language-oriented to emergent communication for multi-agent remote control," 2024, *arXiv:2401.12624*.

[181] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[182] A. Muhamed et al., "CTR-BERT: Cost-effective knowledge distillation for billion-parameter teacher models," in *Proc. NeurIPS Efficient Natural Lang. Speech Process. Workshop*, 2021, pp. 1–9.

[183] Y. Tian, Y. Han, X. Chen, W. Wang, and N. V. Chawla, "TinyLLM: Learning a small student from multiple large language models," 2024, *arXiv:2402.04616*.

[184] S. Soltan, H. Khan, and W. Hamza, "Limitations of knowledge distillation for zero-shot transfer learning," in *Proc. 2nd Workshop Simple Efficient Nat. Lang. Process.*, 2021, pp. 22–31.

[185] J. Ko, S. Kim, T. Chen, and S.-Y. Yun, "DistiLLM: Towards streamlined distillation for large language models," 2024, *arXiv:2402.03898*.

[186] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 24824–24837.

[187] S. Franklin and A. Graesser, "Is it an agent, or just a program? A taxonomy for autonomous agents," in *Proc. Int. Workshop Agent Theories, Archit., Languages*, 1996, pp. 21–35.

[188] Y. Wang et al., "Towards a theoretical framework of autonomous systems underpinned by intelligence and systems sciences," *IEEE/CAA J. Automatica Sinica*, vol. 8, no. 1, pp. 52–63, Jan. 2021.

[189] Y. LeCun, "A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27," *Open Rev.*, vol. 62, no. 1, p. 62, 2022. [Online]. Available: https://openreview.net/pdf?id=BZ5a1r-kVsf

[190] L. Wang et al., "A survey on large language model based autonomous agents," 2023, *arXiv:2308.11432*.

[191] Y. Talebirad and A. Nadiri, "Multi-agent collaboration: Harnessing the power of intelligent LLM agents," 2023, *arXiv:2306.03314*.

[192] C. Cui et al., "A survey on multimodal large language models for autonomous driving," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 958–979.

[193] T. Händler, "Balancing autonomy and alignment: A multi-dimensional taxonomy for autonomous LLM-powered multi-agent architectures," 2023, *arXiv:2310.03659*.

[194] R. Anil et al., "Palm 2 technical report," 2023, *arXiv:2305.10403*.

[195] A. Chowdhery et al., "Palm: Scaling language modeling with pathways," *J. Mach. Learn. Res.*, vol. 24, no. 240, pp. 1–113, 2023.

[196] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 22199–22213.

[197] C. Sun, "Benchmarks for RL on goal-directed language tasks with LLMs," Proc. Electr. Eng. Comput. Sci., Univ. California, Berkeley, CA, USA, Rep. UCB/EECS-2023-98, 2023.

[198] Q. Wu et al., "AutoGEN: Enabling next-gen LLM applications via multi-agent conversation framework," 2023, *arXiv:2308.08155*.

[199] Y. Wang et al., "Empowering autonomous driving with large language models: A safety perspective," 2023, *arXiv:2312.00812*.

[200] I. Gim, G. Chen, S.-s. Lee, N. Sarda, A. Khandelwal, and L. Zhong, "Prompt cache: Modular attention reuse for low-latency inference," 2023, *arXiv:2311.04934*.

[201] G. Ramírez, M. Lindemann, A. Birch, and I. Titov, "Cache & distil: Optimising API calls to large language models," 2023, *arXiv:2310.13561*.

[202] S. Zhang, X. Zeng, Y. Wu, and Z. Yang, "Harnessing scalable transactional stream processing for managing large language models [vision]," 2023., *arXiv:2307.08225*.

[203] G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis, "Efficient streaming language models with attention sinks," 2023, *arXiv:2309.17453*.

[204] D. Xu et al., "LLMCad: Fast and scalable on-device large language model inference," 2023, *arXiv:2309.04255*.

[205] R. Ma et al., "Poster: PipeLLM: Pipeline LLM inference on heterogeneous devices with sequence slicing," in *Proc. ACM SIGCOMM Conf.*, 2023, pp. 1126–1128.

[206] W. Kwon et al., "Efficient memory management for large language model serving with PagedAttention," in *Proc. 29th Symp. Oper. Syst. Princ.*, 2023, pp. 611–626.

[207] L. Ribar, I. Chelombiev, L. Hudlass-Galley, C. Blake, C. Luschi, and D. Orr, "SparQ attention: Bandwidth-efficient LLM inference," 2023, *arXiv:2312.04985*.

[208] K. Alizadeh et al., "LLM in a flash: Efficient large language model inference with limited memory," 2023, *arXiv:2312.11514*.

[209] E. Frantar and D. Alistarh, "SparseGPT: Massive language models can be accurately pruned in one-shot," 2023, *arXiv:2301.00774*.

[210] S. Zhang et al., "OPT: Open pre-trained transformer language models," 2022, *arXiv:2205.01068*.

[211] X. Shen et al., "Agile-quant: Activation-guided quantization for faster inference of LLMs on the edge," 2023, *arXiv:2312.05693*.

[212] Y. Yu et al., "Distributed multi-agent target tracking: A Nash-combined adaptive differential evolution method for UAV systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 8, pp. 8122–8133, Aug. 2021.

[213] Y. He, F. R. Yu, N. Zhao, V. C. Leung, and H. Yin, "Software-defined networks with mobile edge computing and caching for smart cities: A big data deep reinforcement learning approach," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 31–37, Dec. 2017.

[214] Y. He, N. Zhao, and H. Yin, "Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 44–55, Jan. 2018.

[215] W. Wang, M. Ghobadi, K. Shakeri, Y. Zhang, and N. Hasani, "Optimized network architectures for training large language models with billions of parameters," 2023. [Online]. Available: https://people.csail.mit.edu/ghobadi/papers/rail_llm_hotnets_2023.pdf

[216] S. Suri, S. N. Das, K. Singi, K. Dey, V. S. Sharma, and V. Kaulgud, "Software engineering using autonomous agents: Are we there yet?" in *Proc. 38th IEEE/ACM Int. Conf. Autom. Softw. Eng. (ASE)*, 2023, pp. 1855–1857.

[217] Z. Xi et al., "The rise and potential of large language model based agents: A survey," 2023, *arXiv:2309.07864*.

[218] J. Li, M. Zhang, N. Li, D. Weyns, Z. Jin, and K. Tei, "Exploring the potential of large language models in self-adaptive systems," 2024, *arXiv:2401.07534*.

[219] Y. Hu et al., "Planning-oriented autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17853–17862.

[220] P. Morales and R. A. Showalter-Bucher, "Towards large language models at the edge on mobile, augmented reality, and virtual reality devices with unity," in *Proc. Syst. Archit. Gener. AI Edge/Mobile Platforms (SAGE)*, 2023, pp. 1–4.

[221] A. Gopalkrishnan, R. Greer, and M. Trivedi, "Multi-frame, lightweight & efficient vision-language models for question answering in autonomous driving," 2024, *arXiv:2403.19838*.

[222] R. Jia et al., "Valley-conserved topological integrated antenna for 100-Gbps THz 6G wireless," *Sci. Adv.*, vol. 9, no. 44, 2023, Art. no. eadi8500.

[223] S. Yin et al., "A survey on multimodal large language models," 2023, *arXiv:2306.13549*.

[224] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nat. Med.*, vol. 29, no. 8, pp. 1930–1940, 2023.

[225] W. Wu et al., "AI-native network slicing for 6G networks," *IEEE Wireless Commun.*, vol. 29, no. 1, pp. 96–103, Feb. 2022.

[226] B. Jacob et al., "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2704–2713.

[227] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," in *Low-Power Computer Vision*. Milton, U.K.: Chapman and Hall/CRC, 2022, pp. 291–326.

[228] Z. Liu et al., "LLM-QAT: Data-free quantization aware training for large language models," 2023, *arXiv:2305.17888*.

[229] T. Dettmers and L. Zettlemoyer, "The case for 4-bit precision: K-bit inference scaling laws," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 7750–7774.

[230] X. Huang, Z. Liu, S.-Y. Liu, and K.-T. Cheng, "Efficient quantization-aware training with adaptive coreset selection," 2023, *arXiv:2306.07215*.

[231] Y. Shang, Z. Yuan, Q. Wu, and Z. Dong, "PB-LLM: Partially binarized large language models," 2023, *arXiv:2310.00034*.

[232] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1628–1656, 3rd Quart., 2017.

[233] B. Cao, L. Zhang, Y. Li, D. Feng, and W. Cao, "Intelligent offloading in multi-access edge computing: A state-of-the-art review and framework," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 56–62, Mar. 2019.

[234] H. Zou, Q. Zhao, L. Bariah, M. Bennis, and M. Debbah, "Wireless multi-agent generative AI: From connected intelligence to collective intelligence," 2023, *arXiv:2307.02757*.

[235] L. Bariah, Q. Zhao, H. Zou, Y. Tian, F. Bader, and M. Debbah, "Large language models for telecom: The next big thing?" 2023, *arXiv:2306.10249*.

[236] S. Hu, Y.-C. Liang, Z. Xiong, and D. Niyato, "Blockchain and artificial intelligence for dynamic resource sharing in 6G and beyond," *IEEE Wireless Commun.*, vol. 28, no. 4, pp. 145–151, Aug. 2021.

[237] M. Matinmikko-Blue, S. Yrjölä, and P. Ahokangas, "Spectrum management in the 6G era: The role of regulation and spectrum sharing," in *Proc. 2nd 6G Wireless Summit (6G SUMMIT)*, 2020, pp. 1–5.

[238] S. Iyer, "Performance analysis of a dynamic spectrum assignment technique for 6G," *IETE J. Res.*, vol. 69, no. 11, pp. 7695–7703, 2023.

[239] J. Jeon, R. D. Ford, V. V. Ratnam, J. Cho, and J. Zhang, "Coordinated dynamic spectrum sharing for 5G and beyond cellular networks," *IEEE Access*, vol. 7, pp. 111592–111604, 2019.

[240] S.-M. Kim et al., "Opportunism in spectrum sharing for beyond 5G with sub-6 GHz: A concept and its application to duplexing," *IEEE Access*, vol. 8, pp. 148877–148891, 2020.

[241] M. Khadem, F. Zeinali, N. Mokari, and H. Saeedi, "AI-enabled priority and auction-based spectrum management for 6G," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2024, pp. 1–6.

[242] T. K. Rodrigues, K. Suto, and N. Kato, "Edge cloud server deployment with transmission power control through machine learning for 6G Internet of Things," *IEEE Trans. Emerg. Topics Comput.*, vol. 9, no. 4, pp. 2099–2108, Oct.–Dec. 2021.

[243] E. Park, M.-S. Lee, H.-S. Kim, and S. Bahk, "AdaptaBLE: Adaptive control of data rate, transmission power, and connection interval in Bluetooth low energy," *Comput. Netw.*, vol. 181, Nov. 2020, Art. no. 107520.

[244] T. Pan, X. Wu, and X. Li, "Dynamic multi-sleeping control with diverse quality-of-service requirements in sixth-generation networks using federated learning," *Electronics*, vol. 13, no. 3, p. 549, 2024.

[245] X. Zhang, P. Han, C. Feng, T. Ma, and L. Guo, "Multi-dimensional resource orchestration toward edge intelligence in 6G networks," *IEEE Commun. Mag.*, vol. 61, no. 12, pp. 46–52, Dec. 2023.

[246] A. Bouroudi, A. Outtagarts, and Y. Hadjadj-Aoul, "Multi-domain scaling algorithm with inter-orchestrator communication for beyond 5G/6G networks," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, 2024, pp. 1054–1061.

[247] M. Ashwin, A. S. Alqahtani, A. Mubarakali, and B. Sivakumar, "Efficient resource management in 6G communication networks using hybrid quantum deep learning model," *Comput. Electr. Eng.*, vol. 106, Mar. 2023, Art. no. 108565.

[248] L. Ma et al., "Dynamic neural network-based resource management for mobile edge computing in 6G networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 10, no. 3, pp. 953–967, Jun. 2024.

[249] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.

[250] W. Huang, Y. Wang, A. Cheng, A. Zhou, C. Yu, and L. Wang, "A fast, performant, secure distributed training framework for LLM," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2024, pp. 4800–4804.

[251] S. Montagna, S. Ferretti, L. C. Klopfenstein, A. Florio, and M. F. Pengo, "Data decentralisation of LLM-based chatbot systems in chronic disease self-management," in *Proc. ACM Conf. Inf. Technol. Soc. Good*, 2023, pp. 205–212.

[252] C. M. Fang, K. Zieliński, P. Maes, J. Paradiso, B. Blumberg, and M. B. Kjærgaard, "Enabling waypoint generation for collaborative robots using LLMs and mixed reality," 2024, *arXiv:2403.09308*.

[253] T. Q. Duong, L. D. Nguyen, B. Narottama, J. A. Ansere, D. Van Huynh, and H. Shin, "Quantum-inspired real-time optimization for 6G networks: Opportunities, challenges, and the road ahead," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 1347–1359, 2022.

[254] M. A. Ferrag et al., "Edge learning for 6G-enabled Internet of Things: A comprehensive survey of vulnerabilities, datasets, and defenses," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 4, pp. 2654–2713, 4th Quart., 2023.

[255] R. Luo et al., "BioGPT: Generative pre-trained transformer for biomedical text generation and mining," *Brief. Bioinf.*, vol. 23, no. 6, 2022, Art. no. bbac409.

[256] C. Huang et al., "Holographic MIMO surfaces for 6G wireless networks: Opportunities, challenges, and trends," *IEEE wireless Commun.*, vol. 27, no. 5, pp. 118–125, Oct. 2020.

[257] E. C. Strinati et al., "6G: The next frontier: From holographic messaging to artificial intelligence using subterahertz and visible light communication," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 42–50, Sep. 2019.
[258] M. Z. Chowdhury, M. Shahjalal, S. Ahmed, and Y. M. Jang, "6G wireless communication systems: Applications, requirements, technologies, challenges, and research directions," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 957–975, 2020.
[259] B. Soviero, D. Kuhn, A. Salle, and V. P. Moreira, "ChatGPT goes shopping: LLMs can predict relevance in eCommerce search," in *Proc. Eur. Conf. Inf. Retr.*, 2024, pp. 3–11.
[260] Z. Li et al., "Understanding is compression," 2024, *arXiv:2407.07723*.
[261] Z. Chen et al., "A framework for cost-effective and self-adaptive LLM shaking and recovery mechanism," 2024, *arXiv:2403.07283*.
[262] Z. Xu et al., "DriveGPT4: Interpretable end-to-end autonomous driving via large language model," 2023, *arXiv:2310.01412*.
[263] I. de Zarzà, J. de Curtò, G. Roig, and C. T. Calafate, "LLM multimodal traffic accident forecasting," *Sensors*, vol. 23, no. 22, p. 9225, 2023.
[264] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *J. Field Robot.*, vol. 37, no. 3, pp. 362–386, 2020.
[265] M. Soori, B. Arezoo, and R. Dastres, "Artificial intelligence, machine learning and deep learning in advanced robotics, a review," *Cogn. Robot.*, vol. 3, pp. 54–70, Apr. 2023.
[266] Y. Cui et al., "DriveLLM: Charting the path toward full autonomous driving with large language models," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 1450–1464, Jan. 2024.
[267] S. Teng et al., "Motion planning for autonomous driving: The state of the art and future perspectives," *IEEE Trans. Intell. Veh.*, vol. 8, no. 6, pp. 3692–3711, Jun. 2023.
[268] B. Brown, M. Broth, and E. Vinkhuyzen, "The halting problem: Video analysis of self-driving cars in traffic," in *Proc. CHI Conf. Human Fact. Comput. Syst.*, 2023, pp. 1–14.
[269] F.-Y. Wang, K. M. Carley, D. Zeng, and W. Mao, "Social computing: From social informatics to social intelligence," *IEEE Intell. Syst.*, vol. 22, no. 2, pp. 79–83, Mar./Apr. 2007.
[270] S. Park et al., "VLAAD: Vision and language assistant for autonomous driving," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 980–987.
[271] C. Cui, Y. Ma, X. Cao, W. Ye, and Z. Wang, "Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 902–909.
[272] Y. Yang, Q. Zhang, C. Li, D. S. Marta, N. Batool, and J. Folkesson, "Human-centric autonomous systems with LLMs for user command reasoning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 988–994.
[273] C. Cui, Z. Yang, Y. Zhou, Y. Ma, J. Lu, and Z. Wang, "Large language models for autonomous driving: Real-world experiments," 2023, *arXiv:2312.09397*.
[274] X. Ding, J. Han, H. Xu, W. Zhang, and X. Li, "HiLM-D: Towards high-resolution understanding in multimodal large language models for autonomous driving," 2023, *arXiv:2309.05186*.
[275] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 1–25.
[276] J. Janai, F. Güney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," *Found. Trends® Comput. Graph. Vis.*, vol. 12, nos. 1–3, pp. 1–308, 2020.
[277] X. Huang et al., "The ApolloScape dataset for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. workshops*, 2018, pp. 954–960.
[278] S. Kato, E. Takeuchi, Y. Ishiguro, Y. Ninomiya, K. Takeda, and T. Hamada, "An open approach to autonomous vehicles," *IEEE Micro*, vol. 35, no. 6, pp. 60–68, Nov./Dec. 2015.
[279] "Model S owner's manual." Tesla Motors. Accessed: Jul. 7, 2024. [Online]. Available: https://www.tesla.com/sites/default/files/model_s_owners_manual_touch screen_7.1_das_ap_north_america_r20160112_en_us.pdf
[280] "Baidu Apollo project repository," Baidu. Accessed: Jul. 7, 2024. [Online]. Available: https://github.com/ApolloAuto/apollo
[281] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5G: RAN, core network and caching solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3098–3130, 4th Quart., 2018.
[282] X. Han et al., "Foundation intelligence for smart infrastructure services in transportation 5.0," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 39–47, Jan. 2024.
[283] X. Yang, Z. Yu, J. Wang, and T. Menzies, "Understanding static code warnings: An incremental AI approach," *Expert Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 114134.
[284] Y. Yang, X. Xia, D. Lo, and J. Grundy, "A survey on deep learning for software engineering," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–73, 2022.
[285] A. Fan et al., "Large language models for software engineering: Survey and open problems," 2023, *arXiv:2310.03533*.
[286] T. H. Le, H. Chen, and M. A. Babar, "Deep learning for source code modeling and generation: Models, applications, and challenges," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–38, 2020.
[287] M. C. Wuisang, M. Kurniawan, K. A. W. Santosa, A. A. S. Gunawan, and K. E. Saputra, "An evaluation of the effectiveness of OpenAI's ChatGPT for automated python program bug fixing using QuixBugs," in *Proc. Int. Semin. Appl. Technol. Inf. Commun. (iSemantic)*, 2023, pp. 295–300.
[288] D. Sobania, M. Briesch, C. Hanna, and J. Petke, "An analysis of the automatic bug fixing performance of ChatGPT," 2023, *arXiv:2301.08653*.
[289] Y. Charalambous, N. Tihanyi, R. Jain, Y. Sun, M. A. Ferrag, and L. C. Cordeiro, "A new era in software security: Towards self-healing software via large language models and formal verification," 2023, *arXiv:2305.14752*.
[290] D. Busch, A. Bainczyk, and B. Steffen, "Towards LLM-based system migration in language-driven engineering," in *Proc. Int. Conf. Eng. Comput.-Based Syst.*, 2023, pp. 191–200.
[291] W. Ma et al., "The scope of ChatGPT in software engineering: A thorough investigation," 2023, *arXiv:2305.12138*.
[292] M. Chen et al., "Evaluating large language models trained on code," 2021, *arXiv:2107.03374*.
[293] Y. Dong, X. Jiang, Z. Jin, and G. Li, "Self-collaboration code generation via ChatGPT," 2023, *arXiv:2304.07590*.
[294] Z. Yuan et al., "No more manual tests? Evaluating and improving ChatGPT for unit test generation," 2023, *arXiv:2305.04207*.
[295] R. Boutaba et al., "A comprehensive survey on machine learning for networking: Evolution, applications and research opportunities," *J. Internet Services Appl.*, vol. 9, no. 1, pp. 1–99, 2018.
[296] A. Karapantelakis, P. Alizadeh, A. Alabassi, K. Dey, and A. Nikou, "Generative AI in mobile networks: A survey," *Ann. Telecommun.*, vol. 79, no. 1, pp. 15–33, 2024.
[297] H. Du et al., "Generative AI-aided optimization for AI-generated content (AIGC) services in edge networks," 2023, *arXiv:2303.13052*.
[298] M. Wang, A. Pang, Y. Kan, M.-O. Pun, C. S. Chen, and B. Huang, "LLM-assisted light: Leveraging large language model capabilities for human-mimetic traffic signal control in complex urban environments," 2024, *arXiv:2403.08337*.
[299] J. Medved, R. Varga, A. Tkacik, and K. Gray, "OpenDaylight: Towards a model-driven sdn controller architecture," in *Proc. IEEE Int. Symp. World Wireless, Mobile Multimedia Netw.*, 2014, pp. 1–6.
[300] P. Berde et al., "ONOS: Towards an open, distributed SDN OS," in *Proc. 3rd Workshop Hot Topics Softw. Defined Netw.*, 2014, pp. 1–6.
[301] H. Liang et al., "Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence," *Nat. Med.*, vol. 25, no. 3, pp. 433–438, 2019.
[302] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting clinical events via recurrent neural networks," in *Proc. Mach. Learn. Healthc. Conf.*, 2016, pp. 301–318.
[303] A. S. Panayides et al., "AI in medical imaging informatics: Current challenges and future directions," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 7, pp. 1837–1857, Jul. 2020.
[304] M. Tayefi et al., "Challenges and opportunities beyond structured data in analysis of electronic health records," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 13, no. 6, 2021, Art. no. e1549.
[305] M. B. McDermott, S. Wang, N. Marinsek, R. Ranganath, L. Foschini, and M. Ghassemi, "Reproducibility in machine learning for health research: Still a ways to go," *Sci. Transl. Med.*, vol. 13, no. 586, 2021, Art. no. eabb1655.

[306] T. H. Kung et al., "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLoS Digit. Health*, vol. 2, no. 2, 2023, Art. no. e0000198.

[307] X. Yang et al., "A large language model for electronic health records," *NPJ Digit. Med.*, vol. 5, no. 1, p. 194, 2022.

[308] K. Singhal et al., "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[309] A. Gilson et al., "How does ChatGPT perform on the united states medical licensing examination? The implications of large language models for medical education and knowledge assessment," *JMIR Med. Educ.*, vol. 9, no. 1, 2023, Art. no. e45312.

[310] T. Li et al., "CancerGPT for few shot drug pair synergy prediction using large pretrained language models," *npj Digit. Med.*, vol. 7, no. 1, p. 40, 2024.

[311] J. Clusmann et al., "The future landscape of large language models in medicine," *Commun. Med.*, vol. 3, no. 1, p. 141, 2023.

[312] M. S. Diab and E. Rodriguez-Villegas, "Embedded machine learning using microcontrollers in wearable and ambulatory systems for health and care applications: A review," *IEEE Access*, vol. 10, pp. 98450–98474, 2022.

[313] S. M. Sánchez et al., "Edge computing driven smart personal protective system deployed on NVIDIA Jetson and integrated with ROS," in *Proc. 18th Int. Workshops PAAMS*, 2020, pp. 385–393.

[314] X. Hou et al., "Large language models for software engineering: A systematic literature review," 2023, *arXiv:2308.10620*.

[315] J. Wang, Y. Huang, C. Chen, Z. Liu, S. Wang, and Q. Wang, "Software testing with large language models: Survey, landscape, and vision," *IEEE Trans. Softw. Eng.*, vol. 50, no. 4, pp. 911–936, Apr. 2024.

[316] Z. Feng et al., "CodeBERT: A pre-trained model for programming and natural languages," 2020, *arXiv:2002.08155*.

[317] R. Sun et al., "SQL-PaLM: Improved large language model adaptation for text-to-SQL," 2023, *arXiv:2306.00739*.

[318] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nat. Mach. Intell.*, vol. 1, no. 9, pp. 389–399, 2019.

[319] P.-H. C. Chen, Y. Liu, and L. Peng, "How to develop machine learning models for healthcare," *Nat. Mater.*, vol. 18, no. 5, pp. 410–414, 2019.

[320] J. Lee et al., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[321] Y. Chen, J. Arkin, Y. Zhang, N. Roy, and C. Fan, "Scalable multi-robot collaboration with large language models: Centralized or Decentralized systems?" 2023, *arXiv:2309.15943*.

[322] I. Singh et al., "ProgPrompt: Generating situated robot task plans using large language models," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2023, pp. 11523–11530.

[323] Y. Li, S. Wang, H. Ding, and H. Chen, "Large language models in finance: A survey," in *Proc. 4th ACM Int. Conf. AI Financ.*, 2023, pp. 374–382.

[324] S. Wu et al., "BloombergGPT: A large language model for finance," 2023, *arXiv:2303.17564*.

[325] X. Zhang and W. Gao, "Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method," 2023, *arXiv:2310.00305*.

[326] C. Gan, Q. Zhang, and T. Mori, "Application of LLM agents in recruitment: A novel framework for resume screening," 2024, *arXiv:2401.08315*.

[327] M. D. Vu et al., "GPTVoiceTasker: LLM-powered virtual assistant for smartphone," 2024, *arXiv:2401.14268*.

[328] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection," in *Proc. 16th ACM Workshop Artif. Intell. Secur.*, 2023, pp. 79–90.

[329] H. Brown, K. Lee, F. Mireshghallah, R. Shokri, and F. Tramèr, "What does it mean for a language model to preserve privacy?" in *Proc. ACM Conf. Fairness, Account., Transp.*, 2022, pp. 2280–2292.

[330] M. Nasr et al., "Scalable extraction of training data from (production) language models," 2023, *arXiv:2311.17035*.

[331] A. Wei, N. Haghtalab, and J. Steinhardt, "Jailbroken: How does LLM safety training fail?" 2023, *arXiv:2307.02483*.

[332] X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang, "'Do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models," 2023, *arXiv:2308.03825*.

[333] N. Carlini et al., "Extracting training data from large language models," in *Proc. 30th USENIX Secur. Symp. (USENIX Security)*, 2021, pp. 2633–2650.

[334] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fus.*, vol. 58, pp. 82–115, Jun. 2020.

[335] L. Floridi et al., "AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," *Minds Mach.*, vol. 28, pp. 689–707, Nov. 2018.

[336] N. Kandpal, M. Jagielski, F. Tramèr, and N. Carlini, "Backdoor attacks for in-context learning with language models," 2023, *arXiv:2307.14692*.

[337] P. He, H. Xu, Y. Xing, H. Liu, M. Yamada, and J. Tang, "Data poisoning for in-context learning," 2024, *arXiv:2402.02160*.

[338] N. Carlini, "A LLM assisted exploitation of AI-guardian," 2023, *arXiv:2307.15008*.

[339] H. Yang, K. Xiang, H. Li, and R. Lu, "A comprehensive overview of backdoor attacks in large language models within communication networks," 2023, *arXiv:2308.14367*.

[340] S. Zhao, J. Wen, L. A. Tuan, J. Zhao, and J. Fu, "Prompt as triggers for backdoor attack: Examining the vulnerability in language models," 2023, *arXiv:2305.01219*.

[341] Y. Liu, Y. Jia, R. Geng, J. Jia, and N. Z. Gong, "Prompt injection attacks and defenses in LLM-integrated applications," 2023, *arXiv:2310.12815*.

[342] Z. Xu, Y. Liu, G. Deng, Y. Li, and S. Picek, "LLM jailbreak attack versus defense techniques—A comprehensive study," 2024, *arXiv:2402.13457*.

[343] A. Mehrotra et al., "Tree of attacks: Jailbreaking black-box LLMs automatically," 2023, *arXiv:2312.02119*.

[344] G. Deng et al., "MasterKey: Automated jailbreak across multiple large language model chatbots," 2023, *arXiv:2307.08715*.

[345] Y. Tian, X. Yang, J. Zhang, Y. Dong, and H. Su, "Evil geniuses: Delving into the safety of LLM-based agents," *arXiv:2311.11855*, 2023.

[346] R. Staab, M. Vero, M. Balunović, and M. Vechev, "Beyond memorization: Violating privacy via inference with large language models," 2023, *arXiv:2310.07298*.

[347] H. A. Inan et al., "Training data leakage analysis in language models," 2021, *arXiv:2101.05405*.

[348] M. Beckerich, L. Plein, and S. Coronado, "RatGPT: Turning online LLMs into proxies for malware attacks," 2023, *arXiv:2308.09183*.

[349] F. Wu, N. Zhang, S. Jha, P. McDaniel, and C. Xiao, "A new era in LLM security: Exploring security concerns in real-world LLM-based systems," 2024, *arXiv:2402.18649*.

[350] U. Iqbal, T. Kohno, and F. Roesner, "LLM platform security: Applying a systematic evaluation framework to OpenAI's ChatGPT plugins," 2023, *arXiv:2309.10254*.

[351] S. Neel and P. Chang, "Privacy issues in large language models: A survey," 2023, *arXiv:2312.06717*.

[352] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei, "Demystifying membership inference attacks in machine learning as a service," *IEEE Trans. Services Comput.*, vol. 14, no. 6, pp. 2073–2089, Nov./Dec. 2019.

[353] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Security Privacy (SP)*, 2017, pp. 3–18.

[354] N. Kandpal, K. Pillutla, A. Oprea, P. Kairouz, C. Choquette-Choo, and Z. Xu, "User inference attacks on LLMs," in *Proc. Soc. Respons. Lang. Model. Res.*, 2023, pp. 1–14.

[355] V. Shejwalkar, H. A. Inan, A. Houmansadr, and R. Sim, "Membership inference attacks against NLP classification models," in *Proc. Workshop Privacy Mach. Learn.*, 2021, pp. 1–13.

[356] C. Li, Z. Song, W. Wang, and C. Yang, "A theoretical insight into attack and defense of gradient leakage in transformer," 2023, *arXiv:2311.13624*.

[357] M. U. Hadi et al., "A survey on large language models: Applications, challenges, limitations, and practical usage," TechRxiv, Preprint, 2023. [Online]. Available: https://www.techrxiv.org/users/618307/articles/682263-large-language-models-a-comprehensive-survey-of-its-applications-challenges-limitations-and-future-prospects
[358] X. Li, F. Tramer, P. Liang, and T. Hashimoto, "Large language models can be strong differentially private learners," 2021, *arXiv:2110.05679*.
[359] A. Zafar et al., "Building trust in conversational AI: A comprehensive review and solution architecture for explainable, privacy-aware systems using LLMs and knowledge graph," 2023, *arXiv:2308.13534*.
[360] O. J. Romero, J. Zimmerman, A. Steinfeld, and A. Tomasic, "Synergistic integration of large language models and cognitive architectures for robust AI: An exploratory analysis," in *Proc. AAAI Symp. Ser.*, 2023, pp. 396–405.
[361] J. Evertz, M. Chlosta, L. Schönherr, and T. Eisenhofer, "Whispers in the machine: Confidentiality in LLM-integrated systems," 2024, *arXiv:2402.06922*.
[362] N. Tihanyi, M. A. Ferrag, R. Jain, and M. Debbah, "CyberMetric: A benchmark dataset for evaluating large language models knowledge in cybersecurity," 2024, *arXiv:2402.07688*.
[363] N. Kandpal, E. Wallace, and C. Raffel, "Deduplicating training data mitigates privacy risks in language models," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 10697–10707.
[364] Y. Liu, J. Cao, C. Liu, K. Ding, and L. Jin, "Datasets for large language models: A comprehensive survey," 2024, *arXiv:2402.18041*.
[365] G. Wenzek et al., "CCNet: Extracting high quality monolingual datasets from web crawl data," 2019, *arXiv:1911.00359*.
[366] Y. Zhou et al., "Defending jailbreak prompts via in-context adversarial game," 2024, *arXiv:2402.13148*.
[367] R. Anil, B. Ghazi, V. Gupta, R. Kumar, and P. Manurangsi, "Large-scale differentially private BERT," 2021, *arXiv:2108.01624*.
[368] C. Yung, H. M. Dolatabadi, S. Erfani, and C. Leckie, "Round trip translation defence against large language model jailbreaking attacks," 2024, *arXiv:2402.13517*.
[369] N. Jain et al., "Baseline defenses for adversarial attacks against aligned language models," 2023, *arXiv:2309.00614*.
[370] S. Chen, J. Piet, C. Sitawarin, and D. Wagner, "StruQ: Defending against prompt injection with structured queries," 2024, *arXiv:2402.06363*.
[371] N. Varshney, P. Dolin, A. Seth, and C. Baral, "The art of defending: A systematic evaluation and analysis of LLM defense strategies on safety and over-defensiveness," 2023, *arXiv:2401.00287*.
[372] C. F. Chan, D. W. Yip, and A. Esmradi, "Detection and defense against prominent attacks on preconditioned LLM-integrated virtual assistants," 2024, *arXiv:2401.00994*.
[373] X. Qi, T. Xie, J. T. Wang, T. Wu, S. Mahloujifar, and P. Mittal, "Towards a proactive ML approach for detecting backdoor poison samples," in *Proc. 32nd USENIX Security Symp. (USENIX Security)*, 2023, pp. 1685–1702.
[374] A. Helbling, M. Phute, M. Hull, and D. H. Chau, "LLM self defense: By self examination, LLMs know they are being tricked," 2023, *arXiv:2308.07308*.
[375] Y. Wang, Z. Shi, A. Bai, and C.-J. Hsieh, "Defending LLMs against jailbreaking attacks via backtranslation," 2024, *arXiv:2402.16459*.
[376] Z. Xi et al., "Defending pre-trained language models as few-shot learners against backdoor attacks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 1–16.
[377] G. Alon and M. Kamfonas, "Detecting language model attacks with perplexity," 2023, *arXiv:2308.14132*.
[378] Z. Hu et al., "Token-level adversarial prompt detection based on perplexity measures and contextual information," 2023, *arXiv:2311.11509*.
[379] B. Chen, A. Paliwal, and Q. Yan, "Jailbreaker in jail: Moving target defense for large language models," in *Proc. 10th ACM Workshop Moving Target Defense*, 2023, pp. 29–32.
[380] X. Li, T. Zhu, and W. Zhang, "Efficient ransomware detection via portable executable file image analysis by LLaMA-7b," 2023, Preprint.
[381] M. Asfour and J. C. Murillo, "Harnessing large language models to simulate realistic human responses to social engineering attacks: A case study," *Int. J. Cybersecur. Intell. Cybercr.*, vol. 6, no. 2, pp. 21–49, 2023.
[382] Y. Yang, "IoT software vulnerability detection techniques through large language model," in *Proc. Int. Conf. Formal Eng. Methods*, 2023, pp. 285–290.
[383] D. Noever, "Can large language models find and fix vulnerable software?" 2023, *arXiv:2308.10345*.
[384] T. Ahmed and P. Devanbu, "Better patching using LLM prompting, via self-consistency," in *Proc. 38th IEEE/ACM Int. Conf. Autom. Softw. Eng. (ASE)*, 2023, pp. 1742–1746.
[385] G. Deng et al., "PentestGPT: An LLM-empowered automatic penetration testing tool," 2023, *arXiv:2308.06782*.
[386] J. Xu et al., "AutoAttacker: A large language model guided system to implement automatic cyber-attacks," 2024, *arXiv:2403.01038*.
[387] S. Hays and D. J. White, "Employing LLMs for incident response planning and review," 2024, *arXiv:2403.01271*.
[388] S. Mitra et al., "LOCALINTEL: Generating organizational threat intelligence from global and local cyber knowledge," 2024, *arXiv:2401.10036*.
[389] M. M. Yamin, E. Hashmi, M. Ullah, and B. Katt, "Applications of LLMs for generating cyber security exercise scenarios," 2024, Preprint.
[390] S. Hao et al., "Synthetic data in AI: Challenges, applications, and ethical implications," 2024, *arXiv:2401.01629*.
[391] N. Tihanyi, T. Bisztray, R. Jain, M. A. Ferrag, L. C. Cordeiro, and V. Mavroeidis, "The formAI dataset: Generative AI in software security through the lens of formal verification," in *Proc. 19th Int. Conf. Predict. Models Data Anal. Softw. Eng.*, 2023, pp. 33–43.
[392] H. Zhu, S. Zhang, and K. Chen, "AI-guardian: Defeating adversarial attacks using backdoors," in *Proc. IEEE Symp. Security Privacy (SP)*, 2023, pp. 701–718.
[393] J. Wanner, L.-V. Herm, K. Heinrich, and C. Janiesch, "The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study," *Electron. Markets*, vol. 32, no. 4, pp. 2079–2102, 2022.
[394] M. Nyre-Yu, E. Morris, M. Smith, B. Moss, and C. Smutz, "Explainable AI in cybersecurity operations: Lessons learned from xAI tool deployment," Sandia Nat. Lab. (SNL-NM), Albuquerque, NM, USA, Rep. SAND2022-3586C, 2022.
[395] T. Ali and P. Kostakos, "HuntGPT: Integrating machine learning-based anomaly detection and explainable AI with large language models (LLMs)," 2023, *arXiv:2309.16021*.
[396] Z. Huang, C.-C. Shen, S. Doshiy, N. Thomasy, and H. Duong, "Difficulty-level metric for cyber security training," in *Proc. IEEE Int. Multi-Discipl. Conf. Cogn. Methods Situat. Aware. Decis.*, 2015, pp. 172–178.
[397] J. Yu et al., "KoLA: Carefully benchmarking world knowledge of large language models," 2023, *arXiv:2306.09296*.
[398] J. Gennari, S.-H. Lau, S. Perl, J. Parish, and G. Sastry, "Considerations for evaluating large language models for cybersecurity tasks," Carnegie Mellon Univ., Pittsburgh, PA, USA, White Paper, 2024.
[399] M. Shao et al., "An empirical evaluation of LLMs for solving offensive security challenges," 2024, *arXiv:2402.11814*.
[400] Z. Liu, J. Shi, and J. F. Buford, "CyberBench: A multi-task benchmark for evaluating large language models in cybersecurity," in *Proc. Workshop Artif. Intell. Cyber Secur. (AICS)*, 2014, pp. 1–14.
[401] Z. Liu, "SecQA: A concise question-answering dataset for evaluating large language models in computer security," 2023, *arXiv:2312.15838*.
[402] Y. Zhang, W. Song, Z. Ji, and N. Meng., "How well does LLM generate security tests?" 2023, *arXiv:2310.00710*.
[403] F. E. Vasconcelos and G. S. Almeida, "LLaMa assisted reverse engineering of modern ransomware: A comparative analysis with early crypto-ransomware," 2023, Preprint.
[404] N. Zahan, P. Burckhardt, M. Lysenko, F. Aboukhadijeh, and L. Williams, "Shifting the lens: Detecting malware in NPM ecosystem with large language models," 2024, *arXiv:2403.12196*.
[405] M. Bhatt et al., "Purple llama CyberSeceval: A secure coding benchmark for language models," 2023, *arXiv:2312.04724*.
[406] R. Rejeleene, X. Xu, and J. Talburt, "Towards trustable language models: Investigating information quality of large language models," 2024, *arXiv:2401.13086*.
[407] X. Wu et al., "Usable XAI: 10 strategies towards exploiting explainability in the LLM era," 2024, *arXiv:2403.08946*.

[408] Q. V. Liao and J. W. Vaughan, "AI transparency in the age of LLMs: A human-centered research roadmap," 2023, *arXiv:2306.01941*.

[409] V. Hassija et al., "Interpreting black-box models: A review on explainable artificial intelligence," *Cogn. Comput.*, vol. 16, no. 1, pp. 45–74, 2024.

[410] A. Nadeem, "Understanding adversary behavior via XAI: Leveraging sequence clustering to extract threat intelligence," Ph.D. dissertation, Dept. Comput. Sci., Delft Univ. Technol., Delft, The Netherlands, 2024.

[411] M. Zhou, V. Abhishek, T. Derdenger, J. Kim, and K. Srinivasan, "Bias in generative AI," 2024, *arXiv:2403.02726*.

[412] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

[413] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[414] T. C. Miranda, P.-F. Gimenez, J.-F. Lalande, V. V. T. Tong, and P. Wilke, "Debiasing android malware datasets: How can i trust your results if your dataset is biased?" *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2182–2197, 2022.

[415] S. Achintalwar et al., "Detectors for safe and reliable LLMs: Implementations, uses, and limitations," 2024, *arXiv:2403.06009*.

[416] G. Chhikara, A. Sharma, K. Ghosh, and A. Chakraborty, "Few-shot fairness: Unveiling LLM's potential for fairness-aware classification," 2024, *arXiv:2402.18502*.

[417] J. Zhang, K. Bao, Y. Zhang, W. Wang, F. Feng, and X. He, "Is ChatGPT fair for recommendation? Evaluating fairness in large language model recommendation," in *Proc. 17th ACM Conf. Recomm. Syst.*, 2023, pp. 993–999.

[418] M. A. Malek, "Criminal courts' artificial intelligence: The way it reinforces bias and discrimination," *AI Ethics*, vol. 2, no. 1, pp. 233–245, 2022.

[419] D. G. Jones, "Detecting propaganda in news articles using large language models," *Eng. Open Access*, vol. 2, no. 1, pp. 1–12, 2024.

[420] J. Zeng, J. Xu, X. Zheng, and X. Huang, "Certified robustness to text adversarial attacks by randomized [mask]," *Comput. Linguist.*, vol. 49, no. 2, pp. 395–427, 2023.

[421] Y. Xie et al., "Defending ChatGPT against jailbreak attack via self-reminders," *Nat. Mach. Intell.*, vol. 5, no. 12, pp. 1486–1496, 2023.

[422] B. Cao, Y. Cao, L. Lin, and J. Chen, "Defending against alignment-breaking attacks via robustly aligned LLM," 2023, *arXiv:2309.14348*.

[423] Z. Zhang et al., "Certified robustness for large language models with self-denoising," 2023, *arXiv:2307.07171*.

[424] V. C. Müller, "Ethics of artificial intelligence and robotics," 2020. [Online]. Available: https://plato.stanford.edu/entries/ethics-ai/

[425] M. Anderljung et al., "Towards publicly accountable frontier LLMs," in *Proc. Soc. Respons. Lang. Model. Res.*, 2023, pp. 1–13.

[426] A. Kumar, S. Singh, S. V. Murty, and S. Ragupathy, "The ethics of interaction: Mitigating security threats in LLMs," 2024, *arXiv:2401.12273*.

[427] T. P. Pagano et al., "Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods," *Big Data Cogn. Comput.*, vol. 7, no. 1, p. 15, 2023.

[428] J. Bang, B.-T. Lee, and P. Park, "Examination of ethical principles for LLM-based recommendations in conversational AI," in *Proc. Int. Conf. Platform Technol. Service (PlatCon)*, 2023, pp. 109–113.

[429] E. Guo et al., "neuroGPT-X: Towards an accountable expert opinion tool for vestibular schwannoma," medRxiv, Preprint, 2023. [Online]. Available: https://www.medrxiv.org/content/10.1101/2023.02.25.23286117v1

[430] I. D. Raji et al., "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proc. Conf. Fairness, Account., Transp.*, 2020, pp. 33–44.

[431] B. Shneiderman, "Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems," *ACM Trans. Interact. Intell. Syst.*, vol. 10, no. 4, pp. 1–31, 2020.

[432] M. Brundage et al., "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," 2018, *arXiv:1802.07228*.

[433] J. Sauvola, S. Tarkoma, M. Klemettinen, J. Riekki, and D. Doermann, "Future of software development with generative AI," *Autom. Softw. Eng.*, vol. 31, no. 1, p. 26, 2024.

[434] M. Xu et al., "Cached model-as-a-resource: Provisioning large language model agents for edge intelligence in space-air-ground integrated networks," 2024, *arXiv:2403.05826*.

[435] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3929–3938.

[436] T. Wu, L. Luo, Y.-F. Li, S. Pan, T.-T. Vu, and G. Haffari, "Continual learning for large language models: A survey," 2024, *arXiv:2402.01364*.

[437] L. Z. Jiao Dong, Hao Zheng, and P. Nguyen, "Efficiently scale LLM training across a large GPU cluster with Alpa and ray." 2024. Accessed: Jun. 15, 2024. [Online]. Available: https://developer.nvidia.com/blog/efficiently-scale-llm-training-across-a-large-gpu-cluster-with-alpa-and-ray/

[438] R. Bommasani et al., "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.

[439] V. Stafford, "Zero trust architecture," Nat. Inst. Stand. Technol., Gaithersburg, MD, USA, Rep. 800-207, 2020.

[440] E. B. Fernandez and A. Brazhuk, "A critical analysis of zero trust architecture (ZTA)," *Comput. Stand. Interfaces*, vol. 89, Apr. 2024, Art. no. 103832.

[441] J. Hong et al., "Decoding compressed trust: Scrutinizing the trustworthiness of efficient LLMs under compression," 2024, *arXiv:2403.15447*.

[442] K. Velasquez et al., "Service orchestration in fog environments," in *Proc. IEEE 5th Int. Conf. Future Internet Things Cloud (FiCloud)*, 2017, pp. 329–336.

[443] E. Villar-Rodriguez, M. A. Pérez, A. I. Torre-Bastida, C. R. Senderos, and J. López-de Armentia, "Edge intelligence secure frameworks: Current state and future challenges," *Comput. Secur.*, vol. 130, Jul. 2023, Art. no. 103278.

[444] M. Aazam, S. Zeadally, and K. A. Harras, "Offloading in fog computing for IoT: Review, enabling technologies, and research opportunities," *Future Gener. Comput. Syst.*, vol. 87, pp. 278–289, Oct. 2018.

[445] S. Tarkoma, R. Morabito, and J. Sauvola, "AI-native interconnect framework for integration of large language model technologies in 6G systems," 2023, *arXiv:2311.05842*.

[446] M.-I. Corici et al., "Organic 6G networks: Vision, requirements, and research approaches," *IEEE Access*, vol. 11, pp. 70698–70715, 2023.

[447] (Intel Corp., Santa Clara, CA, USA). *Reduce the Attack Surface Around Your Data to Unlock New Opportunities*. Accessed: Jun. 18, 2024. [Online]. Available: https://www.intel.com/content/www/us/en/products/docs/accelerator-engines/software-guard-extensions.html

[448] O. Friha and M. A. Ferrag, "Blockchain technology for 6G communication networks: A vision for the future," in *Cybersecurity Issues in Emerging Technologies*. Boca Raton, FL, USA: CRC Press, 2021, pp. 77–96.

[449] C. I. Nwakanma et al., "Explainable artificial intelligence (XAI) for intrusion detection and mitigation in intelligent connected vehicles: A review," *Appl. Sci.*, vol. 13, no. 3, p. 1252, 2023.

[450] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?" in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 11121–11128.

[451] M. Tan, M. A. Merrill, V. Gupta, T. Althoff, and T. Hartvigsen, "Are language models actually useful for time series forecasting?" 2024, *arXiv:2406.16964*.

[452] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[453] M. Dehghani et al., "Scaling vision transformers to 22 billion parameters," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 7480–7512.

[454] X. L. Dong, S. Moon, Y. E. Xu, K. Malik, and Z. Yu, "Towards next-generation intelligent assistants leveraging LLM techniques," in *Proc. 29th ACM SIGKDD Conf. Knowl. Discov. Data Min.*, 2023, pp. 5792–5793.

[455] B. Li, K. Mellou, B. Zhang, J. Pathuri, and I. Menache, "Large language models for supply chain optimization," 2023, *arXiv:2307.03875*.

[456] W. Jin et al., "Amazon-M2: A multilingual multi-locale shopping session dataset for recommendation and text generation," in *Proc. 37th Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 1–21.

[457] C. Fang et al., "LLM-ensemble: Optimal large language model ensemble method for E-commerce product attribute value extraction," 2024, *arXiv:2403.00863*.

[458] R. Pi, L. Yao, J. Gao, J. Zhang, and T. Zhang, "PerceptionGPT: Effectively fusing visual perception into LLM," 2023, *arXiv:2311.06612*.

OTHMANE FRIHA received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University of Badji-Mokhtar, Annaba, Algeria, in 2016, 2018, and 2024, respectively. His research delves into the critical field of cybersecurity, where he focuses on various aspects of digital security, including AI security, networks security, Internet of Things security, and applied cryptography.

MOHAMED AMINE FERRAG (Senior Member, IEEE) received the bachelor's, master's, Ph.D., and Habilitation degrees in computer science from Badji Mokhtar–Annaba University, Annaba, Algeria, in 2008, 2010, 2014, and 2019, respectively. He served as an Associate Professor with the Department of Computer Science, Guelma University, Algeria, from 2014 until 2022. Concurrently from 2019 to 2022, he held the position of a Senior Researcher with the NAU-Lincoln Joint Research Center of Intelligent Engineering, Nanjing Agricultural University, China. As of 2022, he is the Lead Researcher with the Artificial Intelligence and Digital Science Research Center, Technology Innovation Institute, Abu Dhabi, UAE. His scholarly output includes over 140 papers published in international journals and conference proceedings. He has spearheaded numerous projects in research and development, fostering collaborative ties with academic institutions in the U.K., Australia, USA, Canada, and China. His contributions to the field include the creation of two cybersecurity datasets, namely, the Edge-IIoT dataset and the FormAI dataset, which have become essential resources for AI researchers worldwide. His research primarily focuses on a spectrum of topics within the cybersecurity domain, including wireless network security, network coding security, applied cryptography, blockchain technology, generative AI, software security, and the application of AI in cybersecurity. His academic contributions have been recognized with the 2021 IEEE TEM Best Paper Award, the 2022 Scopus Algeria Award, and many best paper conference awards. He has consistently been named on Stanford University's list of the world's top 2% of scientists four times from 2020 to 2023. He also contributes to the academic community as an Associate Editor for prestigious journals, such as the IEEE INTERNET OF THINGS JOURNAL and *ICT Express* (Elsevier).

BURAK KANTARCI (Senior Member, IEEE) recieved the Ph.D. degree in computer engineering. He is a Full Professor and the Founding Director of the Smart Connected Vehicles Innovation Centre and the Next Generation Communications and Computing Networks Research Lab, Ottawa. He is the author/co-author of 300+ publications in established journals and conferences, and 15 book chapters. Continuously listed among the top-cited scientists in telecommunications and networking based on the data reported by Stanford University since 2020, and since 2021, based on data collected from Microsoft Academic Graph, research.com has listed him among Canada's top computer scientists. He has been a keynote/invited speaker or panelist in 40 events. He holds an Exemplary Editor Award from IEEE COMMUNICATIONS SURVEYS AND TUTORIALS in 2021, and multiple best paper awards from various conferences, most recently from IEEE Globecom2021, Wireless World Research Forum 2022, and IEEE ICC2023 and IEEE VCC2023. He is a recipient of the Minister's Award of Excellence from the Ontario Ministry of Colleges and Universities in 2021. He is the recipient of 2023 Technical Achievement Award of IEEE ComSoc Communications Software Technical Committee. He was a Distinguished Speaker of the Association of Computing Machinery (ACM) from 2019 to 2021. He is currently a Distinguished Lecturer of the IEEE Communications Society and IEEE Systems Council. From 2019 to 2020, he chaired the Communications Systems Integration and Modeling Technical Committee of the Institute of Electrical and Electronics Engineers. He has been the general chair, the program chair or the track chair in 30+ international conferences. He is an Editor of the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS and IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, and an Associate Editor of IEEE NETWORKING LETTERS and *Vehicular Communications* (Elsevier).

BURAK CAKMAK a seasoned technical executive, boasts over 20 years of leadership in product development and people management. As the CTO of Edge Signal, he leads the creation of a global edge computing platform–leveraged by entities like DHL and the Canadian Department of National Defense. His track record includes developing scalable, secure products used by millions. His expertise in edge computing and AI drives edge signal's agnostic platform, enabling real-time insights from cameras, IoT, and industrial systems.

ARDA OZGUN received the Bachelor of Science degree in computer science from Bilkent University. He, a seasoned executive with over 30 years of industry expertise in high-tech product management, currently serves as the Chief Executive Officer with Edge Signal, and as the Vice President of Product Management with Wesley Clover. In his current roles, he is responsible for oversight of the technology and product aspects of many companies within the Wesley Clover portfolio, including those affiliated with the Alacrity Global initiative. At Edge Signal, he spearheads the development of a full-fledged edge computing platform used by organizations worldwide, including DHL, lifecell, thinkRF, Celestra Health Systems, and the Canadian Department of National Defense. Before his tenure with Wesley Clover and Edge Signal, he held notable leadership roles at Turkcell, Vodafone, Nortel, and other renowned firms, where he translated business concepts into tangible products and effective go-to-market strategies. He possesses extensive experience in AI product delivery, as well as conducting product feasibility and viability assessments within the realm of venture capital, having evaluated numerous products throughout his career.

NASSIRA GHOUALMI-ZINE received the Doctoral degree in computer science with a specialization in communication systems in 2005. She is a Full Professor and the Director of the Networks and Systems Laboratory, Badji-Mokhtar Annaba University, Algeria. She currently serves as the President of the IEEE International Conference on Networking and Advanced Systems. Her extensive expertise encompasses computer networks, cybersecurity, artificial intelligence, and embedded systems. She has spearheaded numerous research endeavors, mentored doctoral candidates, and contributed her knowledge as a reviewer for international journals and conferences. She has collaborations with researchers from various institutions worldwide.