The Internet of Things in the Era of Generative AI: Vision and Challenges

Xin Wang and Zhongwei Wan, *The Ohio State University, Columbus, OH, 43210, USA*
Arvin Hekmati and Mingyu Zong, *University of Southern California, Los Angeles, CA, 90089, USA*
Samiul Alam and Mi Zhang, *The Ohio State University, Columbus, OH, 43210, USA*
Bhaskar Krishnamachari, *University of Southern California, Los Angeles, CA, 90089, USA*

Advancements in generative AI hold immense promise to push the Internet of Things (IoT) to the next level. In this article, we share our vision on the IoT in the era of generative AI. We discuss some of the most important applications of generative AI in IoT-related domains. We also identify some of the most critical challenges and discuss current gaps as well as promising opportunities on enabling generative AI for the IoT. We hope this article can inspire new research on the IoT in the era of generative AI.

Today, the Internet of Things (IoT), such as smartphones, wearables, smart speakers, and household robots, has already become an integrated part of our daily lives. These devices can sense, can communicate, and are empowered by modern AI techniques.

Advancements in generative AI[1] have enabled a new wave of the AI revolution. Generative AI refers to AI models that can generate new content in the form of text, images, videos, codes, and many more. While generative AI is not new, it is only until recently that large-scale generative models, exemplified by large language models (LLMs) (e.g., GPT, LLaMA, and Gemini)[2] and multimodal generative models (e.g., GPT-4V, DALL-E, and Stable Diffusion),[3] have made the breakthrough. Such a breakthrough comes from their significantly large model sizes and because they are pretrained on massive amounts of data. These characteristics enable generative AI to generate high-quality data, tackle complex tasks with human-level performance, and exhibit superior generalization ability on new tasks, all of which were not attainable before.

The implications of the advancements of generative AI for the IoT are profound. The distinctive capabilities of generative AI bring pivotal benefits across the entire IoT pipeline, encompassing data generation, data processing, interfacing with IoT devices, and IoT system development and evaluation. These benefits position generative AI as having substantial potential to revolutionize a wide range of IoT applications, such as mobile networks, autonomous driving, the metaverse, robotics, health care, and cybersecurity. At the same time, turning these applications into reality is, however, not trivial. Innovative techniques are needed to address some of the most formidable challenges so as to realize the full potential of generative AI for the IoT.

In this article, we provide our vision and insights on the applications, challenges, and opportunities of what generative AI brings to the IoT (Figure 1). We start by explaining how generative AI could benefit some of the most important IoT applications. Next, we discuss

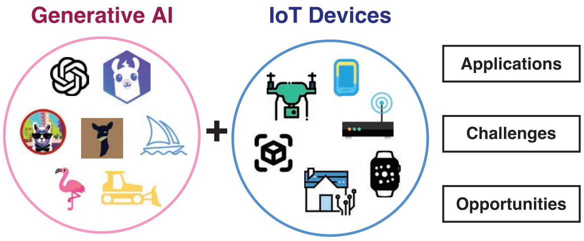

FIGURE 1. The IoT in the era of generative AI.

1089-7801 © 2024 IEEE
Digital Object Identifier 10.1109/MIC.2024.3443169
Date of publication 16 August 2024; date of current version 25 November 2024.

some of the most critical challenges that serve as impediments to enabling generative AI for the IoT and share our views on the gaps as well as promising opportunities to address those challenges. We hope this article can act as a catalyst to inspire new research on the IoT in the era of generative AI.

APPLICATIONS OF GENERATIVE AI IN IOT-RELATED DOMAINS

Leveraging its distinctive capabilities, generative AI holds the potential to revolutionize numerous critical IoT applications. In this section, we delve into a number of application domains (Figure 2) where generative AI has already left its mark and others where its potential is just beginning to be recognized.

Mobile Networks

Generative AI has the considerable potential to revolutionize the design and operation of mobile networks.[4] For instance, generative AI can generate simulations based on historical mobile network data to help network operators foresee potential bottlenecks and adjust resources dynamically. Moreover, generative AI can create highly efficient codecs, resulting in smaller sizes of data to be exchanged between nodes to increase communication efficiency in mobile networks.

Autonomous Vehicles

The transformational journey of the automobile industry toward autonomous vehicles has been deeply influenced by generative AI as well.[5] For example, AI agents empowered by LLMs, like Grok, can be integrated into vehicle systems, such as the Tesla Model 3, to enhance user interfaces and improve communication between the vehicle and its occupants, making interactions more responsive and user friendly. Meanwhile, multimodal generative models are crucial for the development of the advanced driver-assistance systems (ADASs) in autonomous vehicles. ADASs will provide highly accurate prediction of traffic conditions, pedestrian behavior, and potential hazards, allowing for safer and more reliable vehicle automation. Furthermore, by generating realistic driving scenarios during the testing phase, multimodal generative models will enable manufacturers to refine vehicle responses under various conditions, which significantly accelerates the safe deployment of autonomous vehicles.

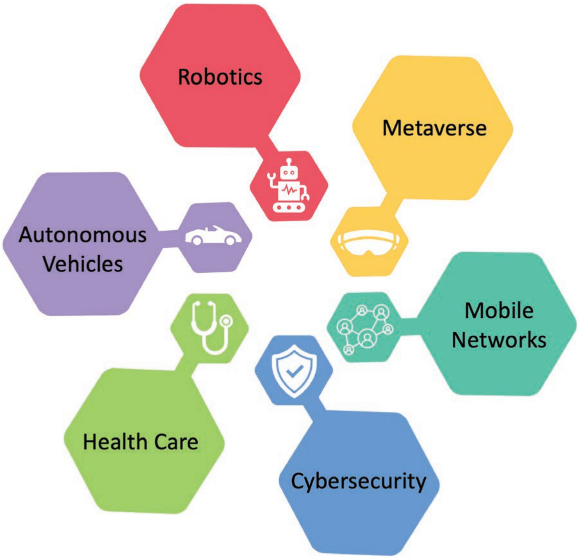

FIGURE 2. Representative application domains of generative AI for the IoT.

Metaverse

Generative AI can significantly enhance the metaverse by leveraging its ability to visualize and simulate based on multimodal sensor data, generating a vivid and immersive virtual realm.[6] Moreover, utilizing generative models to adeptly handle data from various sensors and human inputs can construct dynamic, responsive environments that closely mimic the real world or create entirely new, fantastical settings. For instance, multimodal generative models, such as GPT-4V, are capable of generating simulations based on user interactions and environmental changes captured through IoT devices, such as smart glasses and head-mounted devices, enabling real-time adjustments and enhancements to the virtual landscape. This capability allows for a seamless and adaptive user experience to make the metaverse not just a static backdrop but a living, evolving entity.

Robotics

In the rapidly evolving field of robotics, the integration of generative AI has emerged as a cornerstone, significantly enhancing the capabilities and intelligence of robotic systems.[7] For example, LLMs can be utilized to enhance the interaction capabilities of robots, enabling them to understand dialogues from users and generate human-friendly responses in real time. Additionally, multimodal generative models, which can effectively process and integrate visual and auditory information, are crucial for robots to better understand their physical environment and context, allowing robots to adapt to new situations with greater autonomy. In doing so, these powerful generative models help robots learn from their surroundings and make informed decisions, paving the way for more adaptive and intelligent robotic systems in various settings.

Health Care

The health-care sector, empowered by IoT devices, is experiencing a transformative paradigm shift with the integration of generative AI to revolutionize patient care.[8] For example, LLMs can be used to automatically process and understand medical literature as well as to draft patient reports and personalized treatment plans based on patient history. Meanwhile, multimodal generative models are instrumental in analyzing diverse data modalities collected from IoT devices, such as electrocardiogram (ECG) machines, blood pressure monitors, and pulse oximeters; they are also capable of translating those raw sensor data into human-understandable reports.[9] These models, all together, can integrate text data from electronic health records, numerical data from monitors, and visual data from scans to provide a comprehensive understanding of a patient's health. This integration allows for the delivery of more accurate diagnoses, preventive care, and tailored treatment regimens, thereby significantly enhancing patient outcomes and operational efficiencies in medical facilities.

Cybersecurity

IoT devices are particularly vulnerable to cyberattacks due to their widespread deployment and the often minimal built-in security features. Generative AI can be employed to enhance security measures and address their security challenges.[10] For instance, LLMs can analyze and generate security protocols by learning from vast databases of cybersecurity incidents and responses. Similarly, multimodal generative models can integrate and analyze data from multiple sources, including network traffic, user behavior, and anomaly detection systems, and can predict potential security breaches before they occur. These generative models are trained to identify patterns indicative of cyberthreats, enabling proactive security measures and automatic updates to defense strategies, thus protecting the IoT devices against a wide range of cyberthreats.

CHALLENGES AND OPPORTUNITIES OF ENABLING GENERATIVE AI FOR THE IOT

Turning these applications into reality is, however, not trivial. We have identified five challenges that act as key barriers to realizing generative AI for the IoT. In this section, we describe these challenges and share our perspectives on promising opportunities to address them (Figure 3).

Challenges

C1: Lack of Generative Models for IoT Data

Generative models today are predominantly developed using a few data modalities, including text, images, and videos. However, as summarized in Table 1, IoT-related applications encompass a much wider range of data modalities, including network traffic data; home-deployed sensor data, such as temperature and humidity; and health-care data collected from mobile and wearable devices, such as heart rate, steps taken, and calories burned. Therefore, building generative models based on IoT-related data modalities is much needed, yet it remains a very challenging task.

C2: Deploy Large-Scale Generative Models Under a Limited Memory Budget

Generative models, in general, contain billions of parameters. Such large model sizes directly translate to their significant demands for memory resources.[11] For example, LLaMA-7B requires about 14 GB of memory for storing its 7 billion parameters in half-precision floating-point format (FP16). However, IoT devices are known to be memory constrained. This discrepancy

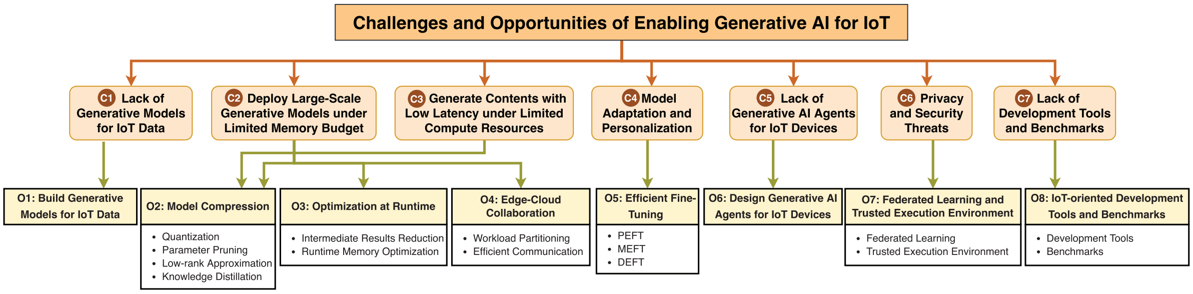
Challenges and Opportunities of Enabling Generative AI for IoT
C1 Lack of Generative Models for IoT Data
C2 Deploy Large-Scale Generative Models under Limited Memory Budget
C3 Generate Contents with Low Latency under Limited Compute Resources
C4 Model Adaptation and Personalization
C5 Lack of Generative AI Agents for IoT Devices
C6 Privacy and Security Threats
C7 Lack of Development Tools and Benchmarks
O1: Build Generative Models for IoT Data
O2: Model Compression
• Quantization
• Parameter Pruning
• Low-rank Approximation
• Knowledge Distillation
O3: Optimization at Runtime
• Intermediate Results Reduction
• Runtime Memory Optimization
O4: Edge-Cloud Collaboration
• Workload Partitioning
• Efficient Communication
O5: Efficient Fine-Tuning
• PEFT
• MEFT
• DEFT
O6: Design Generative AI Agents for IoT Devices
O7: Federated Learning and Trusted Execution Environment
• Federated Learning
• Trusted Execution Environment
O8: IoT-oriented Development Tools and Benchmarks
• Development Tools
• Benchmarks

FIGURE 3. Challenges and opportunities in enabling generative AI for the IoT. DEFT: data-efficient fine-tuning; MEFT: memory-efficient fine-tuning; PEFT: parameter-efficient fine-tuning.

TABLE 1. Representative use cases and data modalities in different IoT application domains.

IoT Application Domain	Representative Use Cases	Data Modalities
Mobile networks	Network optimization and fault detection	Network traffic, network logs, and multimedia data
Autonomous vehicles	Navigation and adversarial testing	Lidar, radar, point cloud, GPS coordinates, and control signals
Metaverse	Virtual meetings and virtual tourism	Gestures, voices, facial expressions, and eye-tracking data
Robotics	Path planning and human–robot interaction	Images, audios, paths, control signals, and error logs
Health care	Disease diagnosis and report generation	ECG, blood pressure, steps taken, calories burned, and food intake
Cybersecurity	Incident response and malware detection	Authentication data, attack vectors, and breach reports

between the intensive memory requirements of generative models and limited memory budgets of IoT devices poses a considerable challenge on the deployment of large-scale generative models on IoT devices.

C3: Generate Contents With Low Latency Under Limited Compute Resources

Many applications of generative AI in IoT-related domains, such as autonomous driving, the metaverse, and robotics, are latency sensitive. Generative models often involve computation-intensive operations during content generation. However, IoT devices are limited in their onboard compute resources, making it challenging to generate contents while meeting application-specific latency requirements. For example, the mixture of experts (MoE)-based LLM ST-l128 requires 6 s to generate only one token on a Raspberry Pi 4B,[12] which is too slow for latency-sensitive applications.

C4: Model Adaptation and Personalization

As the environments in which IoT devices are deployed evolve, the newly collected data may deviate from prior distributions. Consequently, the postdeployment pretrained generative models often necessitate fine-tuning on the devices to effectively adapt to these new data. Moreover, pretrained generative models also need to adapt to the local data on the device for personalization to enhance the user experience.

C5: Lack of Generative AI Agents for IoT Devices

One of the killer applications of generative AI is AI agents, which are autonomous programs capable of generating new content, making decisions, and performing tasks based on their learned knowledge and user requests. Currently, most AI agents are designed for cloud platforms and operate as cloud services. However, due to the privacy concerns associated with personal data collected by IoT devices along with the high latency of delivering cloud-based services,[13] there is a strong need to develop IoT-based AI agents that can process personal data locally on the devices and deliver a wide range of IoT-oriented services promptly, which is not a trivial task due to the diverse and dynamic nature of IoT ecosystems.

C6: Privacy and Security Threats

Data collected by IoT devices, such as personal instructions, dialogues, photos, and videos, often contain privacy-sensitive information that needs to be securely stored and processed.[10] Dealing with privacy and security threats is challenging in the context of generative models. For example, during retrieval-augmented generation, the query sentences are frequently sent over the network to a remote vector database to find similar samples for augmentation. While being transmitted through the network, the query is vulnerable to leakages or attacks, presenting challenges for protecting privacy and securing the generative models.

C7: Lack of Development Tools and Benchmarks

The implementation of generative AI for IoT applications presents a wide range of challenges due to the unique characteristics of IoT devices and their deployment environments. To facilitate the implementation and widespread adoption of these applications, development tools are essential. However, existing generic development tools, such as PyTorch and TensorFlow, are not designed for IoT-oriented scenarios; development tools designed for mobile and edge devices, such as PyTorch Mobile and TFLite, do not provide dedicated support for large generative models. Moreover, the advancement of a field cannot be realized without established benchmarks. Unfortunately, there are still no benchmarks that are specifically designed for generative AI for IoT-oriented applications.

Opportunities

O1: Build Generative Models for IoT Data (Addresses C1)

The lack of generative models for IoT data hinders the development of generative-AI-assisted IoT applications. Therefore, there is a significant opportunity for researchers and practitioners to develop new generative models for analyzing or generating various types of IoT data. One promising approach to building generative models for IoT data is through multimodal LLMs. For instance, MEIT[9] is a multimodal LLM designed to understand and analyze raw ECG sensor data and generate the analysis reports using human-understandable language. Specifically, MEIT transforms raw ECG sensor data into tokens and stores them alongside text tokens for report generation. When the MEIT model has been fine-tuned with a small set of ECG sensor data, it has demonstrated superior performance in generating expert-level ECG reports for heart disease diagnosis.

O2: Model Compression (Addresses C2 and C3)

To address the challenges of model deployment and content generation under limited memory and compute resources on IoT devices, one effective approach is to compress the generative models[14] before deployment. Conventional compression techniques designed for models that are much smaller than LLMs require retraining after compression. In contrast, many compression techniques for generative models are posttraining based, which avoid retraining after compression, given that retraining generative models is expensive. In general, compression techniques for generative models fall into four categories: quantization, parameter pruning, low-rank approximation, and knowledge distillation. Quantization compresses the model by reducing the precision of the weights and/or activations.

Nevertheless, even with the most advanced quantization technique, the highest model compression ratio is limited by the smallest bit width. Parameter pruning compresses the model by eliminating redundant model parameters. Pruning methods can be classified into structured pruning and unstructured pruning. Unstructured pruning has much more pruning flexibility and, thus, enjoys a lower accuracy drop compared to structured pruning. However, unstructured pruning incurs irregular sparsification, which makes the resulting pruned models difficult to be deployed on IoT devices due to a lack of hardware support. Low-rank approximation compresses the model by approximating the weight matrix using the product of two or more smaller low-ranking ones with lower dimensions. Knowledge distillation transfers knowledge from a larger model (the teacher) to a smaller one (the student). Unlike the other three compression strategies, it requires training or fine-tuning for knowledge transfer, making it more expensive to apply.

Lastly, it is worthwhile to note that these four categories of model compression methods are orthogonal to each other, allowing for their combinations to further compress the models. Existing methods are able to compress LLMs by up to 80% with minimum accuracy degradation, enabling the deployment of generative models, such as LLaMA-7B, on IoT devices. As an example, the quantized 4-bit LLaMA-7B can already be deployed on smartphones while only consuming 4 GB of memory.

O3: Optimization at Runtime (Addresses C2)

Model compression reduces the memory and compute resource demands of generative models before they are deployed on IoT devices. When performing inference at runtime, intermediate results, such as activation outputs and the key value (KV) cache, have to be computed and stored for further processing. These intermediate results consume a significant amount of computation and memory resources at runtime. Optimizing the runtime memory of the KV cache is a unique challenge that conventional machine learning (ML) models do not have. Thus, techniques that reduce the intermediate results, such as evicting unnecessary items from the cache or compressing the cache contents, are fruitful optimization opportunities.

Another strategy for runtime optimization is allocating certain model parameters, intermediate results, and computational tasks at different memory hierarchies. Given the limited onboard GPU memory inside IoT devices, leveraging the larger capacity of CPU random-access memory and disk storage is a promising opportunity. This approach allows for the storage of larger intermediate results and some of the model parameters outside the GPU memory to enable the execution of LLMs on resource-constrained IoT devices. At the same time, a primary challenge of this strategy is the slow data transfer rate between different memory units. Innovative techniques that can reduce the frequency of data transfers or pipeline data transfer with other operations will have great promise to address this challenge.

O4: Edge–Cloud Collaboration (Addresses C2)

Given their limited memory and compute resources, some IoT devices may not be able to even run the models that are already compressed. In such cases, it is necessary to offload the execution of part or even the whole model to nearby resourceful edge servers or the cloud.[15]

The key to such edge–cloud collaboration is the design of effective workload partitioning and efficient communication techniques. Workload partitioning is not trivial since IoT devices, edge servers, and the cloud have very different compute, memory, and energy resources, and the available network bandwidths can be dynamic. Due to the NP-hard nature of the workload partitioning problem, manually identifying the best performing partition can be practically infeasible, especially for billion-parameter generative models, where the number of potential choices of partition points can be extremely large. One promising opportunity lies in designing a highly efficient search-based strategy to automatically search for and identify the partition points that optimize the overall performance.

Communication between IoT devices and edge servers or the cloud is often conducted through a wireless channel, in which bandwidth can become the bottleneck. To ensure a timely exchange of migrated workloads while minimizing the bandwidth usage and power consumption caused by wireless transmission, efficient communication is essential. At the same time, due to their large model sizes and the potential large amount of data they need to generate, generative models incur a significant burden on communication. In such cases, we envision that efficient communication techniques are highly demanded. For example, instead of directly transmitting the model or data, techniques that compress the model, embedding vectors, and raw data will become extremely useful.

O5: Efficient Fine-Tuning (Addresses C4)

The needs for model adaptation and personalization underscore the importance of developing highly efficient on-device fine-tuning techniques for resource-constrained IoT devices. At a high level, efficient fine-tuning techniques fall into three categories: parameter-efficient fine-tuning (PEFT), memory-efficient fine-tuning (MEFT), and data-efficient fine-tuning (DEFT).

PEFT reduces the computational cost of fine-tuning by selecting only a subset of model parameters for tuning. Among PEFT techniques, low-rank adaptation (LoRA) is one of the most widely used methods. Instead of fine-tuning the full parameters of generative models, LoRA freezes the weights and injects trainable rank decomposition matrices into the model. Through fine-tuning these small matrices, the model is able to achieve comparable performance as full-parameter fine-tuning.

Although PEFT is able to reduce the computational cost of the fine-tuning process, it can still incur large memory usage. Motivated by this limitation, MEFT focuses on reducing fine-tuning memory footprints by conducting model quantization before fine-tuning, utilizing optimizers that require less memory or combining the gradient calculation and model parameter updating together.

Different from PEFT and MEFT, DEFT achieves efficient fine-tuning from a data-centric perspective. By using a small fraction of the data, DEFT can achieve comparable performance to that obtained from fine-tuning with the entire dataset. Another benefit of DEFT is that it can be combined with PEFT or MEFT to further enhance the fine-tuning efficiency. At the same time, most of the existing DEFT methods heavily rely on the manual selection of the small set of data for fine-tuning, which can be difficult to accomplish without domain knowledge. Therefore, an automated data selection scheme would provide more benefit when applying DEFT for specific IoT application domains. Existing efforts on efficient fine-tuning for IoT devices are able to fine-tune OPT-1.3B using approximately 4 GB and 6.5 GB of memory on the OPPO Reno6 smartphone.[16] However, most generative models, especially LLMs, contain many more parameters than 1.3 billion. Fine-tuning large-scale generative models for IoT devices is still a challenging task but full of opportunities.

O6: Design Generative AI Agents for IoT Devices (Addresses C5)

The privacy and latency issues of cloud-based AI agents motivate the design of IoT-based AI agents. One fundamental capability of AI agents is task planning, which involves breaking down complex tasks into a sequence of simpler steps that could be automatically accomplished. However, designing AI agents that can perform effective task planning can be challenging given the resource constraints of IoT devices as well as diverse IoT-related tasks, including sensing, user interactions, data management and processing, information retrieval, control, and activation. In such cases, we envision that opportunities lie in developing techniques that allow agents to make highly effective plans for diverse IoT-related tasks, transform the agent's plans into low-level instructions compatible with various IoT devices, and schedule onboard resources to ensure the execution of plans in an efficient manner.

O7: Federated Learning (FL) and the Trusted Execution Environment (TEE) (Addresses C6)

As a privacy-preserving ML paradigm, FL emerges as a solution that can improve the quality of the generative models through personal data while mitigating privacy risks by keeping the data inside the IoT devices.[17] While FL has been intensively studied in recent years, most of the proposed techniques have been developed for models with much smaller scales. The emergence of billion-parameter generative models precludes their complete storage within IoT devices due to resource limitations, presenting new challenges in designing FL frameworks that were not previously encountered. To address this challenge, we envision that the opportunities lie in exploring partial-training-based approaches, where each IoT device trains a smaller submodel extracted from the large generative model hosted on the cloud server, and this server model is updated by aggregating those trained submodels. For example, Alam et al.[18] introduce FedRolex, which extracts submodels from the large server model via a rolling window. Such a rolling mechanism results in more stable convergence and ensures that the global model is updated uniformly with superior model quality.

The TEE has become a standard technology to enhance the security of IoT devices. A TEE provides a secure area within a processor, ensuring that the data and operations executed within its confines are protected from external threats. In the context of generative AI, the TEE can be leveraged in many innovative ways. For example, to secure the user input to the generative models, one can perform tokenization or embedding extraction of the input inside the TEE. This ensures the user input is safeguarded from attacks on local devices before being sent to the generative models for inference.

O8: IoT-Oriented Development Tools and Benchmarks (Addresses C7)

New development tools that support generative models on mobile and edge devices, such as llama.cpp and MLC LLM, have recently been developed. However, these newly developed tools are still in their infancy. We envision that further refinement on better supporting IoT-related tasks, such as runtime resource management, workload offloading, privacy and security enhancement, and efficient fine-tuning, is much needed and, hence, holds great promise.

Lastly, although benchmarks such as MMLU, GSM8K, and MMMU are becoming standards to evaluate the performance of generative models for diverse tasks, there is still no benchmark that is specifically designed for generative AI for IoT-related applications. As the development of generative AI for IoT-related applications is advancing rapidly, we envision that a comprehensive and dedicated benchmark that covers a wide range of IoT-oriented data modalities; tasks; platforms; and evaluation metrics, such as latency, memory footprint, and energy consumption, will become critical and beneficial to the IoT community.

CONCLUDING REMARKS

Generative AI has shown immense promise in advancing the capabilities of the IoT. In this article, we highlighted the key benefits and elaborated on important IoT applications enabled by generative AI. We also presented the key challenges and opportunities to enable generative AI for the IoT. We hope this article can spark further research in this exciting field.

ACKNOWLEDGMENTS

This material is based in part on work supported by the National Science Foundation under Award NeTS-2312675 and the Defense Advanced Research Projects Agency under Contract HR001120C0160. Any views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the funding agency or the U.S. government.

REFERENCES

1. Y. Cao et al., “A comprehensive survey of AI-generated content (AIGC): A history of generative ai from GAN to ChatGPT,” 2023, *arXiv:2303.04226*.
2. W. X. Zhao et al., “A survey of large language models,” 2023, *arXiv:2303.18223*.
3. S. Yin et al., “A survey on multimodal large language models,” 2023, *arXiv:2306.13549*.
4. A. Karapantelakis, P. Alizadeh, A. Alabassi, K. Dey, and A. Nikou, “Generative AI in mobile networks: A survey,” *Ann. Telecommun.*, vol. 79, nos. 1–2, pp. 15–33, 2024, doi: 10.1007/s12243-023-00980-9.
5. C. Cui et al., “A survey on multimodal large language models for autonomous driving,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, 2024, pp. 958–979, doi: 10.1109/WACVW60836.2024.00106.
6. V. Chamola et al., “Beyond reality: The pivotal role of generative AI in the metaverse,” 2023, *arXiv:2308.06272*.

7. F. Zeng et al., “Large language models for robotics: A survey,” 2023, *arXiv:2311.07226*.
8. C. Peng et al., “A study of generative large language model for medical research and healthcare,” *Npj Digit. Med.*, vol. 6, no. 1, 2023, Art. no. 210, doi: 10.1038/s41746-023-00958-w.
9. Z. Wan et al., “MEIT: Multi-modal electrocardiogram instruction tuning on large language models for report generation,” 2024, *arXiv:2403.04945*.
10. M. A. Ferrag et al., “Revolutionizing cyber threat detection with large language models: A privacy-preserving BERT-based lightweight model for IoT/IIoT devices,” 2024, *arXiv:2306.14263*.
11. Z. Wan et al., “Efficient large language models: A survey,” 2024, *arXiv:2312.03863*.
12. 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*.
13. Y. Li et al., “Personal LLM agents: Insights and survey about the capability, efficiency and security,” 2024, *arXiv:2401.05459*.
14. X. Zhu, J. Li, Y. Liu, C. Ma, and W. Wang, “A survey on model compression for large language models,” 2023, *arXiv:2308.07633*.
15. Z. Zhou et al., “Edge intelligence: Paving the last mile of artificial intelligence with edge computing,” *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019, doi: 10.1109/JPROC.2019.2918951.
16. D. Peng, Z. Fu, and J. Wang, “PocketLLM: Enabling on-device fine-tuning for personalized LLMs,” 2024, *arXiv:2407.01031*.
17. S. Alam et al., “FedAIoT: A federated learning benchmark for artificial intelligence of things,” 2024, *arXiv:2310.00109*.
18. S. Alam, L. Liu, M. Yan, and M. Zhang, “FedRolex: Model-heterogeneous federated learning with rolling sub-model extraction,” 2022, *arXiv:2212.01548*.

XIN WANG is a Ph.D. student in computer science at The Ohio State University, Columbus, OH, 43210, USA, advised by Prof. Mi Zhang. His research interests include efficient large language models, machine learning systems, and edge AI. Wang received his master’s degree in computer science from The Ohio State University. Contact him at wang.15980@osu.edu.

ZHONGWEI WAN is a Ph.D. student in computer science at The Ohio State University, Columbus, OH, 43210, USA, advised by Prof. Mi Zhang. His research interests include efficient large language models, large multimodal models, and their applications. Wan received his master’s degree in control science from the University of Chinese Academy of Sciences. Contact him at wan.512@osu.edu.

ARVIN HEKMATI is a Ph.D. student in computer science at the University of Southern California, Los Angeles, CA, 90089, USA, advised by Prof. Bhaskar Krishnamachari. His research interests include machine learning, data-driven algorithms, anomaly detection, and edge computing. Hekmati received his master’s degree in electrical and computer engineering from McMaster University. Contact him at hekmati@usc.edu.

MINGYU ZONG is a Ph.D. student in computer science at the University of Southern California, Los Angeles, CA, 90089, USA, advised by Prof. Bhaskar Krishnamachari. Her research interests include large language models for the Internet of Things (IoT) and federated learning. Zong received her master’s degree in spatial data science from the University of Southern California. Contact her at mzong@usc.edu.

SAMIUL ALAM is a Ph.D. student in computer science at The Ohio State University, Columbus, OH, 43210, USA, advised by Prof. Mi Zhang. His research interests include federated learning and efficient large language models on IoT devices. Alam received his master’s degree in computer science from Michigan State University. Contact him at alam.140@osu.edu.

MI ZHANG is an associate professor in the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, 43210, USA. His research interests include the AI of things, machine learning systems, generative AI, mobile computing, and their domain-specific applications. Zhang received his Ph.D. in computer engineering from the University of Southern California. He is a senior member of the Association for Computing Machinery and a Senior Member of IEEE. Contact him at mizhang.1@osu.edu.

BHASKAR KRISHNAMACHARI is a professor and Ming Hsieh Faculty Fellow in electrical engineering at the University of Southern California, Los Angeles, CA, 90089, USA. His research interests include the design and analysis of algorithms, protocols, and applications for next-generation wireless networks, the IoT, distributed systems, blockchain technologies, AI and machine learning, and network economics. Krishnamachari received his Ph.D. in electrical engineering from Cornell University. He is a Fellow of IEEE. Contact him at bkrishna@usc.edu.