IoT in the Era of Generative AI: Vision and Challenges

Xin Wang[1], Zhongwei Wan[1], Arvin Hekmati[2], Mingyu Zong[2], Samiul Alam[1]
Mi Zhang[1], Bhaskar Krishnamachari[2]
[1]The Ohio State University, [2]University of Southern California
{wang.15980, wan.512, alam.140, mizhang.1}@osu.edu, {hekmati, mzong, bkrishna}@usc.edu

arXiv:2401.01923v2 [cs.DC] 6 Jan 2024

***Abstract*—Equipped with sensing, networking, and computing capabilities, Internet of Things (IoT) such as smartphones, wearables, smart speakers, and household robots have been seamlessly weaved into our daily lives. Recent advancements in Generative AI exemplified by GPT, LLaMA, DALL-E, and Stable Difusion hold immense promise to push IoT to the next level. In this article, we share our vision and views on the benefits that Generative AI brings to IoT, and discuss some of the most important applications of Generative AI in IoT-related domains. Fully harnessing Generative AI in IoT is a complex challenge. We identify some of the most critical challenges including high resource demands of the Generative AI models, prompt engineering, on-device inference, offloading, on-device fine-tuning, federated learning, security, as well as development tools and benchmarks, and discuss current gaps as well as promising opportunities on enabling Generative AI for IoT. We hope this article can inspire new research on IoT in the era of Generative AI.**

***Index Terms*—Internet of Things, IoT, AIoT, Generative AI, Large Language Models, LLMs, Diffusion Models, Edge AI**

I. Introduction

INTERNET of Things (IoT) such as smartphones, wearables, smart speakers, and household robots are ubiquitous today and have become an integrated part of our daily lives. Equipped with sensing, networking, and computing capabilities, these devices can sense, communicate, and integrate artificial intelligence (AI) into the physical world [130]. This synergy between IoT and AI has fundamentally changed how individuals perceive and interact with the world, allowing for more intelligent and efficient operations, improved human-machine interactions, and enhanced decision making.

Recent advancements in Generative AI have enabled a new wave of AI revolution [16]. The new generations of generative models including Large Language Models (LLMs) (e.g., GPT [89, 12, 87], LLaMA [105, 106], and Orca [83]) and Large Multimodal Models (LMM) (e.g., GPT-4V [123], DALL-E [92, 93] and Stable Difusion [95]) have made the breakthrough and achieved remarkable performance in a variety of tasks such as chat, search, image synthesis, code generation, and music composition [42]. Such revolution comes from their significantly large model sizes while being pre-trained 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 and data, all of which were not attainable before.

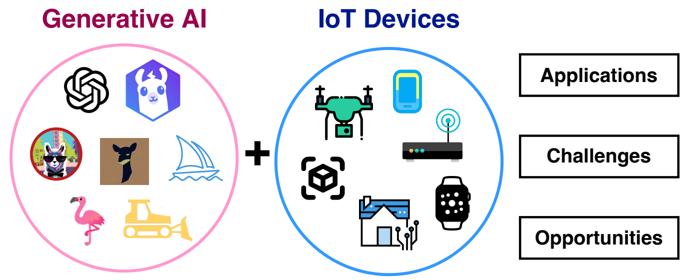

Fig. 1: IoT in the era of Generative AI.

The implications of the advancements of Generative AI for IoT are profound. The unique characteristics of Generative AI bring pivotal benefits across the entire IoT pipeline, encompassing IoT data generation, data processing, interfacing with IoT devices, and IoT system development and evaluation. These advantages position Generative AI as having substantial potential to revolutionize numerous critical IoT applications, including but not limited to mobile networks, autonomous driving, metaverse, robotics, healthcare, and cybersecurity.

Realizing the full potential of Generative AI in IoT is not trivial. Innovative techniques are needed to address some of the most formidable challenges including high resource demands of the Generative AI models, prompt engineering, on-device inference, offloading, on-device fine-tuning, federated learning, security, as well as development tools and benchmarks.

In this article, we provide our vision and insights on the applications, challenges, and opportunities of IoT in the era of Generative AI (Figure 1). We start by providing a brief background on the recent advancements of Generative AI (§II). We then explain how Generative AI could benefit some of the most important IoT applications (§III). Next, we discuss some of the critical challenges that serve as impediments to enabling Generative AI for IoT, and share our views on current gaps as well as promising opportunities to address those challenges (§IV). We hope this paper can act as a catalyst to inspire new research on IoT in the era of Generative AI.

II. Background of Generative AI

Generative AI refers to AI models that can generate new content in the form of text, images, videos, codes, and many more [16]. Generative models are not new. Although traditional generative models such as Recurrent Neural Networks

(RNN) [100], Variational Autoencoders (VAE) [57], Generative Adversarial Networks (GAN) [40], and Bidirectional Encoder Representations from Transformers (BERT) [27] have found their applications in a variety of domains, it is until recently that billion-parameter generative models such as GPT series [89, 12, 87], DALL-E series [92, 93], LLaMA series [105, 106], Stable Diffusion [95] and Orca [83] marked a significant breakthrough. These models demonstrated remarkable performance across a wide spectrum of tasks, elevating the capabilities of Generative AI to unprecedented levels.

The superiority of the contemporary generative models can be attributed to the following two key characteristics:

- **Significantly Large Model Size:** Contemporary generative models contain significantly more parameters than the traditional ones [90]. For example, GPT-4 [87] contains about 1.8 trillion parameters, which is 5,000 times larger than BERT [27].
- **Pre-trained on Massive Amount of Data:** Contemporary generative models are pre-trained on much larger datasets than their predecessors. For example, GPT-3 [12] is pre-trained on more than 500 billion tokens that are over one hundred times than BERT [27].

These characteristics equip contemporary generative models with some unique abilities that were not attainable before:

- **Generating High-Quality Data:** Compared to traditional generative models, the quality of content generated by contemporary generative models is significantly improved. For example, DALL-E 2 [93] is able to generate images that are high in detail and produce visually compelling results that can be indistinguishable from photographs taken by cameras or created by artists.
- **Tackling Complex Tasks with Human-Level Performance:** Contemporary generative models are capable of tackling more complex tasks that conventional counterparts are difficult to deal with. For instance, given a task description along with a brief prompt that contains a few training samples, contemporary generative models such as GPT-4 demonstrate the capability of solving complicated mathematical problems with accuracy comparable to human performance [115].
- **Superior Generalization Capability:** Contemporary generative models exhibit superior generalization ability on new tasks and data. For example, conventional generative models such as GAN [40] require retraining or fine-tuning to generate images that belong to different domains. In contrast, DALL-E 2 [93] is able to generate images for domains that have not been trained before.

III. Applications of Generative AI in IoT-related Domains

Leveraging the distinctive capabilities outlined in §II, 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.

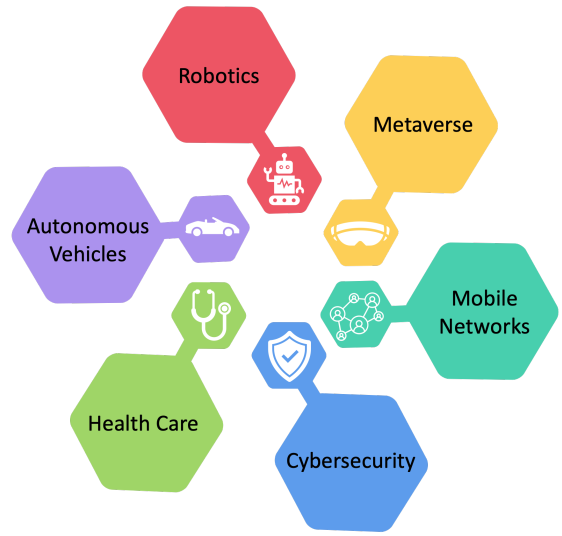

Fig. 2: Applications of Generative AI in IoT-related domains.

A. Mobile Networks

Generative AI has found its use cases in the design and operation of mobile networks. For instance, understanding channel distribution is essential to comprehending the sophisticated dynamics of mobile networks. A pivotal work demonstrates the efficiency and accuracy of adversarial learning with conditional GAN for optical channel modeling [122]. In a recent study, Zhang *et al.* [131] explored the possibility of adopting GAN-based frameworks for air-to-ground channel modeling in wireless unmanned aerial vehicle (UAV) networks over millimeter wave (mmWave) frequencies. In addition to the promising learning accuracy of channel information, their proposed distributed GAN architecture empowers the sharing of channel knowledge through a distributed learning approach. A similar study on channel modeling over mmWave in UAV networks aims to address the issue of generalization. By utilizing federated learning, the introduced FL-GAN framework trains distributed generative models and unifies them into an adaptable advanced model that removes geographical constraints in deployment [8].

The significance of traffic generation in mobile networks cannot be overstated. It allows system developers to simulate and determine the most effective data transmission pattern, assess the scalability and reliability of the network system, and help validate maintenance and updates. The focus of network traffic reproduction has shifted from fundamental machine learning methods (e.g., Support Vector Machines, k-Nearest Neighbour) to deep learning methods, especially GAN [7]. One seminal research that laid the groundwork for harnessing the power of GAN in traffic generation is a knowledge-enhanced GAN-based framework proposed by Hui *et al.*. By feeding realistic traffic data, along with environmental knowledge and information on various IoT devices, to their GAN framework, the proposed method outperforms state-of-the-arts while maintaining high performance even when trained on small datasets.

TABLE I: Types of data generated by Generative AI in different IoT-related application domains.

AI Generated Data	IoT-related Application Domains					
	Mobile Networks	Autonomous Vehicles	Metaverse	Robotics	Health Care	Cybersecurity
Time-series Sensor Data	✓	✓	✓	✓	✓	✓
Vehicular Traffic Data	✓	✓	✓	✓		
Network Traffic Data	✓		✓	✓		✓
Text			✓		✓	
Image			✓		✓	
Audio			✓	✓		
Video			✓	✓	✓	
Code	✓		✓	✓		✓

Spectrum sensing, a vital process in wireless communication for IoT devices, where a radio periodically monitors a frequency band within a range, detects and even predicts its availability, previously relied upon deep learning methods to attain these objectives. However, the emergence of Generative AI presents better solutions to these tasks, as it effectively overcomes the difficulties of collecting training data spanning all conditions and retraining whenever the environment changes. One recent study leverages GAN to generate synthetic data, boosting the prediction accuracy of a spectrum occupancy classifier and training models to adapt to spectrum dynamics [26]. Another work further modifies Wasserstein GAN with gradient penalty (WGAN-GP) to improve the network's ability to complete sensing information of unknown environments [47]. The proposed Enhanced Capsule GAN is constructed to estimate channel availability for the primary users [17].

Lastly, Generative AI plays an essential role in building the Digital Twin (DT) of mobile networks. With a scalable digital replica of a mobile network constructed by their Generative AI-based system, researchers are able to study user behaviors, perform link-level simulations, and model path loss without interrupting the actual system [39].

B. Autonomous Vehicles

The transformational journey of the automobile industry towards autonomous vehicles has been deeply influenced by IoT as well. Krasniqi and Hajrizi [59] present the indispensable role of IoT in this evolution, while Philip *et al.* [88] delve into the real-time applications of IoT, emphasizing its significance in smart traffic control. These systems rely on the effective processing of expansive datasets, where the utility of Generative AI emerges as crucial. In this domain, Xu *et al.* [120] spotlight an architecture that employs Generative AI to produce extensive traffic datasets, fundamental for the safety and efficiency of autonomous systems. Adding another dimension, Marathe *et al.* [78] highlight WEDGE, a synthetic dataset created using vision-language Generative AI, which enhances autonomous vehicle perception, especially under challenging weather conditions. Together, these works elucidate the pivotal roles of IoT and Generative AI in advancing the capabilities of autonomous vehicles.

C. Metaverse

Generative AI's ability to visualize, simulate, and predict based on IoT sensor data creates a reliable virtual realm in the Metaverse. With the intersection of Generative AI, we can construct personalized learning environments, analyze traffic patterns for decision-making, and interact with others in real time. Moreover, Generative AI also promotes bi-directional interactions between users and the constructed world, thus enabling a Metaverse to deliver customized experiences [54]. In recent work, Cai *et al.* [15] develop a Transformer-based framework for tactile signal generation used in virtual and augmented reality. Xu *et al.* [120] exploit Generative AI's power in synthesizing traffic and driving data, optimizing cost efficiencies in driving simulation in vehicular Metaverse. Although not yet transplanted into the Metaverse, Generative AI's capability in diverse areas like creating architecture parameters, eliminating language barriers between different language speakers, and building non-player characteristics, are expected to be instrumental in shaping the outcomes [75].

D. Robotics

In the rapidly evolving field of robotics, the integration of IoT has emerged as a cornerstone. Grieco *et al.* [43] presents a future where robotic IoT systems are seamlessly integrated into daily life, addressing both the challenges and the opportunities that span from communication networks to network security. In the same vein, Kamilaris and Botteghi [56] study the real-world applications and components underpinning IoT-enhanced robotic systems, hinting at the burgeoning role of the Web of Things (WoT) in this arena. Broadening this notion, Batth *et al.* [9] introduce the "Internet of Robotic Things (IoRT)", which intertwines IoT with cloud computing, AI/machine learning, thereby accentuating the significance of a robust architecture for multi-role robotic systems. Tzafestas [107] also discusses the synergy of IoT and AI and their transformative influence on robotics, particularly in the context of IoRT. We are now seeing the integration of IoRT with Generative AI. Taniguchi *et al.* [104] have charted new territory with a brain-inspired architecture termed the whole-brain probabilistic generative models (WB-PGM) for artificial general intelligence (AGI), marrying brain-inspired AI with probabilistic generative models to pave the way for developmental robots adept at continuous learning. Moreover, Luo *et al.* [73] explore crafting a generative personality model tailored for robots, with a focus on encapsulating individual traits and eliciting a spectrum of behaviors, exemplified through non-verbal cues on humanoid robot heads.

E. Health Care

The healthcare sector, empowered by IoT devices, is experiencing a transformative paradigm shift with the integration of Generative AI. These advancements not only enable devices to monitor patient vitals but also predict and generate responses to medical anomalies. Wearables such as smartwatches harness sensor data that can be processed by Generative AI to provide personalized care suggestions, ranging from dietary recommendations to medication modifications [86]. Venkatasubramanian [109] showcases this synergy between IoT and Generative AI by introducing a system for monitoring high-risk maternal and fetal health (MFH), capturing clinical indicators via IoT sensors and using a deep convolutional generative adversarial network (DCGAN) for outcome classifications. In the broader landscape, LLMs such as GPT-4 have been highlighted for their diverse applications in healthcare, encompassing clinical documentation, insurance tasks, and patient interactions, and even possess the capability to interpret text within images [80]. Venkataswamy *et al.* [110] provide a striking example by introducing the “humanoid doctor”, which employs AI to diagnose diseases by collating patient data from IoT devices and leveraging LLMs like ChatGPT for symptom interpretation. Nova [86] further accentuates the potential of Generative AI in enhancing electronic health records (EHRs), streamlining medical conversations, and making medical terminologies more patient-friendly.

F. Cybersecurity

IoT devices have been particularly vulnerable to cyber-attacks due to their widespread deployment and often minimal built-in security features [58]. The rapid proliferation of these devices has only heightened the importance of advanced cybersecurity measures. Generative AI has been integrated into cybersecurity solutions to enhance the protection measures for IoT devices, presenting an avenue for heightened security [1]. A seminal development in this area has been the capability of these models to generate synthetic data which retains the statistical properties of the original data, but without revealing any personally identifiable information, therefore reducing the need for real data exposure and lowering both the potential attack surface and the risk of data breaches [30, 79]. In recent years, researchers have been creating more specialized solutions. Ferrag *et al.* [34] develop SecurityLLM, which integrates SecurityBERT for threat detection and FalconLLM for incident response, showcasing its superiority with a remarkable 98% accuracy rate in detecting a diverse range of cyber threats. Similarly, studies by both Chen *et al.* [20] and Seyyar *et al.* [97] utilize BERT for log analysis to pinpoint abnormalities and discern between standard and anomalous HTTP requests. Furthermore, a range of new models—CySecBERT [10], SecureBERT [2], and CAN-BERT [5] tackle a myriad of tasks from extracting cyber threat data to identifying threats in vehicle networking systems, thus strengthening the cybersecurity domain. Adding to this, Rahali *et al.* [91] utilize BERT to statistically analyze Android application source code, sorting them into malware classifications based on the contextual intricacies of code words. Cintas-Canto *et al.* [23] study the use of LLMs in lightweight cryptography, emphasizing the potential of the GPT-4 based ASCON algorithm for bolstering security, especially pertinent for IoT devices. This collective body of work underscores the transformational role of Generative AI in cybersecurity enhancements for IoT devices, setting a promising trajectory for the fortified security of IoT in the face of evolving threats.

IV. Challenges and Opportunities of Enabling Generative AI for IoT

Turning the applications in §III into reality is not trivial. We have identified eight challenges that act as barriers to realizing Generative AI for IoT (Figure 3). In this section, we describe these challenges, share our perspectives on existing gaps, and highlight promising opportunities to tackle these challenges.

A. High Resource Demands

Generative models such as GPT, DALL-E, and LLaMA series in general contain billions of parameters [111]. Moreover, their performance follows the scaling law [90] where higher accuracies require larger model sizes. Unfortunately, such large model sizes directly translate to their significant resource demands. To illustrate this, Table II lists the resource demands of some of the most well-known generative models. As shown, these models are characterized by their billion-level parameters, necessitating substantial memory and computation for operation. However, IoT devices are known to be resource constrained. This discrepancy between the intensive resource requirements of generative models and the limited resources of IoT devices poses a considerable challenge.

TABLE II: Model sizes and memory usages of representative generative models.

Model	Category	Parameter	Memory
LLaMA2 [106]	Text-to-Text	70B	138G
OPT [132]	Text-to-Text	175B	350G
Orca [83]	Text-to-Text	13B	26G
Stable Diffusion [95]	Text-to-Image	7B	11G
InstructBLIP [65]	Image-to-Text	13B	29G
PointLLM [121]	PointCloud-to-Text	13B	26G

To address this challenge, one effective approach is to compress the generative models, sometimes with a slight accuracy drop as a tradeoff, so as to reduce their memory usage and computational cost. Generally, model compression techniques fall into one of four types: quantization, parameter pruning, low-rank approximation, and knowledge distillation. While a considerable number of model compression techniques have been proposed [14], they are mostly designed for models of much smaller scales compared to contemporary generative models. This gap creates opportunities for innovation in next-generation model compression methods for billion-parameter generative models.

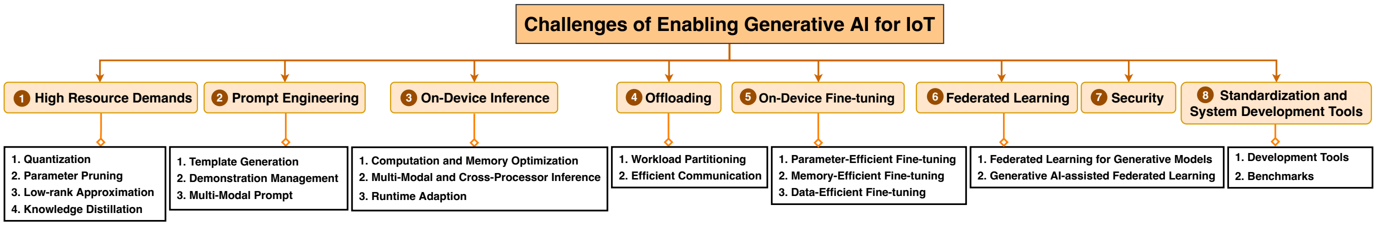

Fig. 3: Challenges of enabling Generative AI for IoT.

Quantization: Quantization reduces the memory requirements and computational cost by reducing the precision of the weights and/or activations in the generative models. The cost of retraining large generative models using the complete training dataset to compensate the accuracy drop due to quantization is expensive. As such, most of the quantization techniques, such as SmoothQuant [117] and GPTQ [36], only use a small amount of calibration data to quickly adjust the weights after the models are quantized. Nevertheless, even with the most advanced quantization technique, the highest model compression ratio is limited by the smallest bit width. Therefore, it is necessary to combine quantization with other model compression techniques to further compress the models, particularly the ones at a larger scale such as LLaMA-65B [105] and OPT-175B [132], so as to fit them inside resource-constrained IoT devices.

Parameter Pruning: Different from quantization, parameter pruning compresses the model by eliminating redundant model parameters. Pruning methods can be classified into structured pruning and unstructured pruning [108, 139]. Structured pruning such as LLM-Pruner [76] removes the entire channels or other structured components from the network, while unstructured pruning such as SparseGPT [35] removes the weights individually without changing the shape of the weight matrices. Therefore, unstructured pruning has much more pruning flexibility and thus enjoys a lower accuracy drop compared to structured pruning [35, 102]. However, unstructured pruning incurs irregular sparsification, which makes the resulting pruned models difficult to be deployed on IoT devices due to lack of hardware support [13].

Low-Rank Approximation: Approximating the weight matrix using the product of two or more smaller matrices with lower dimensions can also reduce the size of generative models [94]. To compensate for the information loss from such approximation, many techniques have been proposed. These can be divided into two main categories: training-required methods, which require fine-tuning the entire model during or after low-rank approximation [11, 45], and training-free methods, which focus on selecting the least significant matrix for approximation [52], or adding another sparse matrix to make up for the approximation loss [18]. Low-rank approximation does not require specialized hardware for implementation and execution, making it more suitable for use in IoT. However, none of the recent studies have attempted to compress LLMs with low-rank approximation to run on IoT devices, which is a promising research direction to explore.

Knowledge Distillation: Knowledge distillation (KD) is a technique for transferring knowledge from a more complex model (the teacher) to a simpler one (the student). Unlike the other three compression strategies, it requires training or fine-tuning for knowledge transfer, making it more expensive to apply. Currently, only a few studies have been conducted in this field to distill a student model from a large generative model [136, 103], and most of these student models only perform well in a specific domain with limited generalization ability and are likely to lose effectiveness in the ever-changing IoT environment.

B. *Prompt Engineering*

Generative models operate with an open-ended nature, necessitating detailed context and information to yield accurate and relevant responses. Prompt engineering is a technique that guides the generative models to generate outputs of high quality and relevance with task-specific hints, known as *prompts* [70, 44, 72]. The output quality of generative models is highly dependent on the design of the prompts. In the context of IoT, developing prompt engineering techniques is confronted with the following challenges.

Template Generation: The quality of the prompt template is one of the most important factors in prompt engineering. Due to its importance and sensitivity, most existing generative models rely heavily on manual written templates for inference and fine-tuning, which considerably limits their productive applications for IoT [70]. This is because the extended duration of user involvement could potentially disrupt the functioning of the IoT devices and hence negatively affect the user experience. To address this challenge, methods that can automatically generate the prompt template given a specific input-label pair have been proposed. For example, Gao *et al.* [37] propose LM-BEF that automatically generates the template of the prompt using pre-trained language models; whereas Wang *et al.* [114] propose Self-Instruct which focuses on generating the instruction in the prompt template. Although these methods demonstrate benefits in improving the final prediction and reducing human efforts in user interactions, none of them have been tested on IoT devices in real-world scenarios.

Demonstration Management: Generative models such as GPT-3 [12] have shown the significant potential on few-shot prompting. This technique involves the addition of a few pertinent training data into the prompt for prediction. However, applying this technique to the real-world IoT environment is still a challenge due to two main reasons. First, resource-limited IoT devices cannot store the entire training set together with the model for inference. Although techniques such as storing these large-scale training data in an external vector database have been proposed [82], the significant communication overhead still makes it difficult to ensure a real-time response. Second, since the data acquired from the IoT devices is constantly changing, it is difficult to get suitable training data and organize them in the right order to form the prompt for improving the generation quality [70]. Most of the existing demonstration organization techniques either focus on training a domain-specific retriever [68], which brings high latency and energy consumption, or choose to apply unsupervised methods such as KNN [118] for selection, but with low accuracy guarantee. Therefore, designing a robust and efficient scheme for managing the whole life-cycle of the demonstration on IoT devices, including storing, selecting, ordering, composing, and deleting, is an important and promising direction.

Multi-Modal Prompt: Although the utilization of prompt engineering in NLP and vision-language generation tasks has been explored [70, 44, 60, 124, 128], its potential within IoT-related tasks, where the IoT data can be multi-modal, including video, audio, 3D point cloud, wireless signals and many more, have yet to be fully realized. For instance, in 3D point cloud generation and audio generation, only a few recent works, such as Point-E [85] and AudioGPT [53], have begun to take advantage of prompt engineering to achieve the desired results. IoT applications often involve the combination of various data sources with different modalities, making the design and optimization of prompts particularly complex. Therefore, it is increasingly important to dedicate efforts to extend prompt engineering into various multi-modal prompts so as to unlock the full potential of real-world IoT applications.

C. On-Device Inference

Another key challenge of enabling Generative AI for IoT is on-device inference. On-device inference is particularly important for latency-sensitive applications such as Metaverse or scenarios where cloud connectivity is not available.

Computation and Memory Optimization: When performing on-device inference, intermediate results such as activation outputs and attention weights have to be computed and stored onboard for further processing. For example, LLaMA2-13B [106] necessitates an additional 8GB of memory for these intermediate results, which is over 30% to the model memory [61]. Moreover, the average generation latency for LLaMA-7B on mobile phones is as slow as seven seconds per token [119]. Therefore, reducing the computation and memory footprint of these intermediate results so as to enhance inference efficiency represents a significant challenge for on-device inference. To address this challenge, we envision that one opportunity lies at preprocessing the input states before feeding them into the generative model to reduce the subsequent computation. For example, Sharir und Anandkumar [98] propose to directly reuse the calculations for a similar input. Chevalier *et al*. [22] apply pre-trained LLMs to compress prompts with long context into short summary vectors to reduce overall computation and memory usage. Li *et al*. [66] propose to filter the intermediate states of unnecessary tokens before they are fed to the next layer of the model for processing. Another opportunity lies at I/O optimization [25]. One popular technique is FlashAttention [25], which utilizes tiling to reduce the number of I/O between GPU high bandwidth memory (HBM) and the GPU's on-chip SRAM. Nevertheless, FlashAttention necessitates costly hardware support and only yields a minuscule improvement in efficiency under the small batch size, which is more typical for IoT devices.

Cross-Processor Inference: The heterogeneity nature of the IoT hardware provides a great opportunity for on-device inference to be performed in a cross-processor manner. In current practice, computations involved in the inference process of generative models are usually executed entirely on a single compute unit such as GPU. We envision that one opportunity lies at allocating or redesigning the modules of a generative model so that different parts of the generative model can be executed at different onboard processes in parallel to enhance on-device inference efficiency. For instance, the decoder module of most generative language models is a key bottleneck to optimize for improved efficiency. However, this module can only be executed serially during the inference process. Some recent optimizations include speculative sampling of important tokens for parallel decoding [64], partitioning the decoding task for parallel execution [101], and transforming the decoding task into many sub-tasks of parallel verification [81]. Despite the potential of general task parallelism strategies to improve the performance of cross-processor inference, these techniques are mainly designed for server-side inference with homogeneous compute units. Therefore, developing techniques that support collaborative inference on different hardware units such as GPU, DSP, CPU, TPU, and NPU inside an IoT device is an important topic for future research.

Runtime Adaptation: The available resources inside IoT devices at runtime can be dynamic due to factors such as changes of battery levels, starting a new application, and turning off a running application [32]. The configuration of the generative models needs to be adjusted in order to adapt to the dynamic resources at runtime. For example, when a mobile phone is in low-power mode, the generative model needs to be reconfigured to perform more lightweight inferences which consume fewer computation resources to save battery life. Currently, only a few studies have investigated this runtime adaption for generative models. For instance, Sheng *et al*. [99] propose FlexGen to flexibly configure generative language models under various hardware resources; Šakota *et al*. [141] design CELMOC, a framework that selects models of different scales for inference based on the user-defined cost-performance tradeoff; and Wang *et al*. [112] propose EcoOptiGen that provides a comprehensive hyperparameter setup to make the most of the limited budget for inference. However, all of these runtime adaption techniques are designed for resourceful server-scale systems. Exploring runtime adaptation techniques for generative models in IoT devices presents a promising opportunity.

D. Offloading

Given the limited memory and computing capacities of IoT devices, some of them may not be able to run the most efficient generative models by just using their own onboard resources. In such scenarios, it is necessary to offload the execution of part or even the whole generative model to nearby resourceful edges or the cloud [138]. The success of this offloading strategy, however, faces two main challenges: workload partitioning and efficient communication.

Workload Partitioning: Workload partitioning refers to the task of partitioning the generative model between the IoT devices and the nearby resourceful edge server or cloud such that different parts of the generative model are executed in a distributed manner. Such task, however, is not trivial, since IoT devices, edge server, and cloud have different computational, memory, and energy resources. Existing techniques can be divided into heuristic-based or learning-based methods. Heuristic-based methods [31] involve the use of predefined rules or experience-driven schemes to partition the workloads. Learning-based methods [33], on the other hand, are trained on historical workload data to identify patterns and relationships between different tasks and resources to identify optimal workload partitions for new and unseen scenarios. Although both of these two types of methods could achieve good partitioning in some use cases, due to the NP-Hard nature of the workload partitioning problem, identifying the best-performing partition can be time consuming, especially for billion-parameter generative models where the search spaces are extremely large, or when the number of partitions needed scales up and the partitions need to be performed in real time. In such scenarios that are commonly encountered in IoT applications, designing highly-efficient workload partitioning techniques is an important topic for future research.

Efficient Communication: Communication between IoT devices and cloud is often conducted through wireless channel in which the bandwidth can be quite limited. To ensure a timely exchange of migrated workloads between IoT devices and cloud while minimizing bandwidth usage and power consumption, efficient communication is essential. Techniques such as message compression [127], data sampling [125], efficient communication protocols [46], and edge caching [140] have been proposed to optimize communication in resource-constrained scenarios. However, due to their large model sizes and the potential large amount of data they need to generate based on application requirements, billion-parameter generative models put a significant burden on communication, especially in scenarios when the wireless bandwidth becomes limited or Generative AI applications are latency-sensitive. In such cases, we envision that more advanced efficient communication techniques are highly demanded.

E. On-Device Fine-Tuning

As the environments in which IoT devices are deployed evolve, the newly collected data may deviate from prior distributions. Consequently, the post-deployment pre-trained generative models often necessitate fine-tuning on the devices to effectively adapt to this new data. This requirement underscores the need for the development of highly efficient fine-tuning techniques to enable on-device fine-tuning for resource constrained IoT devices.

Parameter-Efficient Fine-Tuning (PEFT): PEFT [28] reduces the computational cost of fine-tuning by selecting only a few essential parameters in generative models for tuning. PEFT methods can be in general grouped into three categories: addition-based approach, which inserts small neural modules into the generative models as an adapter for updating [50], or adds some trainable tokens into the input of some layers [67]; the specification-based approach, which only specifies a small number of parameters in the generative models for fine-tuning while keeping the rest frozen [63]; and the reparameterization-based approach, which transforms the updated matrices into a more efficient form, such as the product of the low-rank ones [51]. Although PEFT is able to reduce the computational cost of the fine-tuning process, it still incurs large runtime memory footprint, which acts as a key bottleneck for memory-limited IoT devices [129].

Memory-Efficient Fine-Tuning (MEFT): Motivated by the limitation of PEFT, MEFT [69] focuses on reducing the memory footprint during fine-tuning. These MEFT methods reduce the memory footprint by avoiding storing large input vectors, such as activations [129, 69]; utilizing an optimizer that requires less memory [77]; or by combining the gradient calculation and updating operations [74]. Although MEFT can address the shortcomings of PEFT to reduce the runtime memory, it may take longer to complete the fine-tuning process, resulting in higher energy consumption. Additionally, the performance of the fine-tuned generative models may be reduced, which largely limits its use on IoT devices.

Data-Efficient Fine-Tuning (DEFT): Different from PEFT and MEFT, DEFT achieves efficient fine-tuning from a data-centric perspective [126]. Recent work such as LIMA [135] and AlpaGasus [19] demonstrate that by only using a small fraction of the data, one 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 manual selection of the small set of data for fine-tuning [135, 19], which is difficult to accomplish on IoT devices. Therefore, an automated data selection scheme would allow IoT devices to benefit more from DEFT.

F. Federated Learning

Data captured by IoT devices may contain private user information that is privacy-sensitive. As a privacy-preserving machine learning paradigm, federated learning (FL) emerges as a solution that can improve the quality of the generative models through the personal data while keeping the data on the IoT devices which substantially mitigates privacy risks. While FL has been intensively studied in recent years [55, 113, 133, 4], most of the proposed techniques have been developed for models with much smaller scales. The emergence of billion-parameter generative models presents new challenges in designing FL frameworks that were not previously encountered.

Federated Learning for Generative Models: The rise of large generative models drives the need for training them through federated methods. Nevertheless, most current FL techniques are designed for training compact models that can be entirely accommodated within IoT devices. Unfortunately, the scale of large generative models precludes their complete storage within IoT devices due to resource limitations. How to enable federated training for large generative models on IoT devices is a key challenge. To address this challenge, we envision that the opportunities lie at exploring partial training (PT)-based approaches where each IoT device trains a smaller sub-model extracted from the large generative model hosted on the cloud server, and this server model is updated by aggregating those trained sub-models. For example, Wen *et al*. [116] propose Federated Dropout which extract smaller sub-models from the large server model in a random manner. Horvath *et al*. [49] propose FjORD where sub-models are always extracted from a designated part of the large server model. Alam *et al*. [3] refine this process by introducing FedRolex, which extracts sub-models 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. Lastly, Dun *et al*. [29] propose AsyncDrop, which tackles this problem in an asynchronous manner.

Generative AI-assisted Federated Learning: On the other hand, generative models can play a crucial role in enhancing federated training itself. A key advantage offered by generative models lies in their ability to produce high-quality, diverse synthetic data. For example, Zhang *et al*. [134] propose GPT-FL, a generative model-assisted FL framework that harnesses the power of generative models pre-trained on extensive datasets to generate authentic synthetic data to facilitate the federated training process. Through this approach, GPT-FL consistently surpasses state-of-the-art FL methods in terms of model test accuracy, communication efficiency, and client sampling efficiency.

G. Security

In addition to preserving the privacy of user data captured on IoT devices, ensuring the security of data storage, transmission, and processing for Generative AI applications in the context of IoT is also critical. Trusted Execution Environment (TEE) [96] provides a secure enclave within the processor of IoT devices to protect sensitive computations and data from unauthorized access or tampering. In the context of Generative AI, which involves complex machine learning models creating novel content, TEE becomes essential for safeguarding intellectual property, proprietary algorithms, and the confidentiality of generated outputs. By isolating the private data executed by Generative AI applications within a TEE on IoT devices, organizations can reduce the risk of data breaches and unauthorized interference, creating a secure and reliable environment for running these applications in the IoT landscape. This heightened security is especially important as IoT devices often operate in diverse and dynamic environments, where ensuring the integrity of AI processes is essential for maintaining user trust and system reliability.

H. Development Tools and Benchmarks

Development Tools: The implementation of Generative AI for IoT-related applications presents a wide range of unique challenges due to the unique characteristics of IoT devices and their deployment environments. To facilitate the implementation and widespread adoption of these applications, the design of development tools becomes essential. Existing generic development tools such as PyTorch and TensorFlow as well as LLM-focused development tools such as DeepSpeed [6] and Megatron [84], unfortunately, are not designed for IoT-oriented scenarios. At the same time, development tools such as TFLite, Torch Mobile, and ONNX have been developed to deploy models on IoT devices, but unfortunately do not provide dedicated supports for large generative models. To fill this gap, new development tools have recently been designed. For example, llama.cpp [38] was developed in C++ to enable native deployment of popular LLMs on resource-constrained devices. In addition, new advanced AI compilers such as OpenVino [41], TVM [21] (wrapper on TVM), and MLIR [62] have also been developed to support the efficient execution of Generative AI on diverse IoT platforms. However, these newly developed solutions are still in their infancy. We envision that further refinement on better supporting IoT-related tasks such as workload balancing, resource management, task scheduling, and efficient data processing is a promising opportunity.

Benchmarks: Lastly, the advancement of a field cannot be realized without established benchmarks. Popular benchmarks such as MMLU [48], GSM8K [24], and MMMU [124] are becoming standards to evaluate the accuracy of generative models for diverse tasks. There are also a few benchmarks that focus on metrics related to efficiency. For example, The HULK Benchmark [137] focuses on energy efficiency in pre-trained language models and evaluates efficiency across various tasks. The ELUE [71] framework enables a comprehensive comparison of methods with a focus on performance-efficiency trade-offs. However, there is still no benchmark that is specifically designed for Generative AI for IoT 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, and evaluation metrics such as latency, memory footprint, and energy consumption will become more and more critical and beneficial to the IoT community.

V. Concluding Remarks

Generative AI has shown immense promise in advancing the capabilities of IoT. In this article, we highlighted the key benefits and elaborated on some important applications of Generative AI for IoT. We also presented eight challenges that act as the key barriers to enabling Generative AI for IoT. We hope this article acts as a catalyst to spark further research on IoT in the era of Generative AI.

References

[1] Abebe Abeshu und Naveen Chilamkurti. Deep learning: The frontier for distributed attack detection in fog-to-things computing. *IEEE Communications Magazine*, 56 (2) S. 169–175, 2018.

[2] Ehsan Aghaei, Xi Niu, Waseem Shadid, und Ehab Al-Shaer. SecureBERT: A Domain-Specific Language Model for Cybersecurity. In *International Conference on Security and Privacy in Communication Systems*, S. 39–56. Springer, 2022.

[3] Samiul Alam, Luyang Liu, Ming Yan, und Mi Zhang. Fedrolex: Model-heterogeneous federated learning with rolling sub-model extraction. *Advances in Neural Information Processing Systems*, 35 S. 29677–29690, 2022.

[4] Samiul Alam, Tuo Zhang, Tiantian Feng, Hui Shen, Zhichao Cao, Dong Zhao, JeongGil Ko, Kiran Somasundaram, Shrikanth S Narayanan, Salman Avestimehr, *et al*. FedAIoT: A Federated Learning Benchmark for Artificial Intelligence of Things. *arXiv preprint arXiv:2310.00109*, 2023.

[5] Natasha Alkhatib, Maria Mushtaq, Hadi Ghauch, und Jean-Luc Danger. CAN-BERT do it? Controller Area Network Intrusion Detection System based on BERT Language Model. In *2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA)*, S. 1–8. IEEE, 2022.

[6] Reza Yazdani Aminabadi, Samyam Rajbhandari, Minjia Zhang, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Jeff Rasley, Shaden Smith, Olatunji Ruwase, und Yuxiong He. DeepSpeed Inference: Enabling Efficient Inference of Transformer Models at Unprecedented Scale, 2022.

[7] T. J. Anande und Mark Stephen Leeson. Generative Adversarial Networks (GANs): A Survey on Network Traffic Generation. *International Journal of Machine Learning and Computing*, 2022. URL: https://api.semanticscholar.org/CorpusID:253339791.

[8] Saira Bano, Pietro Cassará, Nicola Tonellotto, und Alberto Gotta. A Federated Channel Modeling System using Generative Neural Networks. *2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring)*, S. 1–5, 2023. URL: https://api.semanticscholar.org/CorpusID:258967882.

[9] Ranbir Singh Batth, Anand Nayyar, und Amandeep Nagpal. Internet of robotic things: driving intelligent robotics of future-concept, architecture, applications and technologies. In *2018 4th international conference on computing sciences (ICCS)*, S. 151–160. IEEE, 2018.

[10] Markus Bayer, Philipp Kuehn, Ramin Shanehsaz, und Christian Reuter. CySecBERT: A Domain-Adapted Language Model for the Cybersecurity Domain. *arXiv preprint arXiv:2212.02974*, 2022.

[11] Matan Ben Noach und Yoav Goldberg. Compressing Pre-trained Language Models by Matrix Decomposition. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, S. 884–889, Suzhou, China, December 2020. Association for Computational Linguistics. URL: https://aclanthology.org/2020.aacl-main.88.

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, und Dario Amodei. Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, und H. Lin, (Hrsg.), *Advances in Neural Information Processing Systems*, volume 33, S. 1877–1901. Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[13] Federico Busato und Jeff Pool. Exploiting NVIDIA Ampere Structured Sparsity with cuSPARSELt. https://developer.nvidia.com/blog/exploiting-ampere-structured-sparsity-with-cusparselt, December 2020. Accessed: 2023-12-13.

[14] Han Cai, Ji Lin, Yujun Lin, Zhijian Liu, Haotian Tang, Hanrui Wang, Ligeng Zhu, und Song Han. Enable deep learning on mobile devices: Methods, systems, and applications. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 27(3) S. 1–50, 2022.

[15] Shaoyu Cai und Kening Zhu. Multi-modal Transformer-based Tactile Signal Generation for Haptic Texture Simulation of Materials in Virtual and Augmented Reality. *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, S. 810–811, 2022. URL: https://api.semanticscholar.org/CorpusID:252275601.

[16] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, und Lichao Sun. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT, 2023.

[17] D. Chandramohan und B. V. Ramana Reddy. Enhanced capsule generative adversarial network for spectrum and energy efficiency of cooperative spectrum prediction framework in cognitive radio network. *Transactions on Emerging Telecommunications Technologies*, 34, 2023. URL: https://api.semanticscholar.org/CorpusID:256893709.

[18] Beidi Chen, Tri Dao, Eric Winsor, Zhao Song, Atri Rudra, und Christopher Ré. Scatterbrain: Unifying Sparse and Low-rank Attention. In A. Beygelzimer, Y. Dauphin, P. Liang, und J. Wortman Vaughan, (Hrsg.), *Advances in Neural Information Processing Systems*, 2021. URL: https://openreview.net/forum?id=SehIKudiIo1.

[19] Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, und Hongxia Jin. AlpaGa-

sus: Training A Better Alpaca with Fewer Data, 2023.

[20] Song Chen und Hai Liao. Bert-log: Anomaly detection for system logs based on pre-trained language model. *Applied Artificial Intelligence*, 36(1) S. 2145642, 2022.

[21] Tianqi Chen, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Yan, Meghan Cowan, Haichen Shen, Leyuan Wang, Yuwei Hu, Luis Ceze, Carlos Guestrin, und Arvind Krishnamurthy. TVM: An Automated End-to-End Optimizing Compiler for Deep Learning. In *Proceedings of the 13th USENIX Conference on Operating Systems Design and Implementation*, OSDI'18, S. 579–594, USA, 2018. USENIX Association. ISBN 9781931971478.

[22] Alexis Chevalier, Alexander Wettig, Anirudh Ajith, und Danqi Chen. Adapting Language Models to Compress Contexts, 2023.

[23] Alvaro Cintas-Canto, Jasmin Kaur, Mehran Mozaffari-Kermani, und Reza Azarderakhsh. ChatGPT vs. Lightweight Security: First Work Implementing the NIST Cryptographic Standard ASCON. *arXiv preprint arXiv:2306.08178*, 2023.

[24] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, und John Schulman. Training Verifiers to Solve Math Word Problems, 2021.

[25] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, und Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness, 2022.

[26] Kemal Davaslioglu und Yalin Evren Sagduyu. Generative Adversarial Learning for Spectrum Sensing. *2018 IEEE International Conference on Communications (ICC)*, S. 1–6, 2018. URL: https://api.semanticscholar.org/CorpusID:4560063.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, und Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019.

[28] Nan Ding, Yuchao Qin, Guoyin Yang, und et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5 S. 220–235, 2023. doi: 10.1038/s42256-023-00626-4. URL: https://doi.org/10.1038/s42256-023-00626-4.

[29] Chen Dun, Mirian Hipolito, Chris Jermaine, Dimitrios Dimitriadis, und Anastasios Kyrillidis. Efficient and Light-Weight Federated Learning via Asynchronous Distributed Dropout. In Francisco Ruiz, Jennifer Dy, und Jan-Willem van de Meent, (Hrsg.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, S. 6630–6660. PMLR, 25–27 Apr 2023.

[30] Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, S. 1–19. Springer, 2008.

[31] Qiang Fan und Nirwan Ansari. Application Aware Workload Allocation for Edge Computing-Based IoT. *IEEE Internet of Things Journal*, 5(3) S. 2146–2153, 2018. doi: 10.1109/JIOT.2018.2826006.

[32] Biyi Fang, Xiao Zeng, und Mi Zhang. NestDNN: Resource-Aware Multi-Tenant On-Device Deep Learning for Continuous Mobile Vision. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (MobiCom)*, S. 115–127, New Delhi, India, 2018.

[33] Weiwei Fang, Wenyuan Xu, Chongchong Yu, und Neal. N. Xiong. Joint Architecture Design and Workload Partitioning for DNN Inference on Industrial IoT Clusters. *ACM Trans. Internet Technol.*, 23(1), feb 2023. ISSN 1533-5399. doi: 10.1145/3551638. URL: https://doi.org/10.1145/3551638.

[34] Mohamed Amine Ferrag, Mthandazo Ndhlovu, Norbert Tihanyi, Lucas C Cordeiro, Merouane Debbah, und Thierry Lestable. Revolutionizing Cyber Threat Detection with Large Language Models. *arXiv preprint arXiv:2306.14263*, 2023.

[35] Elias Frantar und Dan Alistarh. SparseGPT: Massive Language Models Can Be Accurately Pruned in One-Shot. *arXiv preprint arXiv:2301.00774*, 2023.

[36] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, und Dan Alistarh. GPTQ: Accurate Post-training Compression for Generative Pretrained Transformers. *arXiv preprint arXiv:2210.17323*, 2022.

[37] Tianyu Gao, Adam Fisch, und Danqi Chen. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, S. 3816–3830, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.295. URL: https://aclanthology.org/2021.acl-long.295.

[38] Georgi Gerganov. llama.cpp. https://github.com/ggerganov/llama.cpp, 2023.

[39] Jiahui Gong, Qiaohong Yu, Tong Li, Haoqiang Liu, Jun Zhang, Hangyu Fan, Depeng Jin, und Yong Li. Demo: Scalable Digital Twin System for Mobile Networks with Generative AI. *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, 2023. URL: https://api.semanticscholar.org/CorpusID:259177820.

[40] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, und Yoshua Bengio. Generative Adversarial Networks, 2014.

[41] Yury Gorbachev, Mikhail Fedorov, Iliya Slavutin, Artyom Tugarev, Marat Fatekhov, und Yaroslav Tarkan. Openvino deep learning workbench: Comprehensive analysis and tuning of neural networks inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, S. 0–0, 2019.

[42] Roberto Gozalo-Brizuela und Eduardo C. Garrido-Merchán. A survey of Generative AI Applications, 2023.

[43] Luigi Alfredo Grieco, Alessandro Rizzo, Simona

Colucci, Sabrina Sicari, Giuseppe Piro, Donato Di Paola, und Gennaro Boggia. IoT-aided robotics applications: Technological implications, target domains and open issues. *Computer Communications*, 54 S. 32–47, 2014.

[44] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, und Philip Torr. A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models, 2023.

[45] Habib Hajimolahoseini, Walid Ahmed, Mehdi Rezagholizadeh, Vahid Partovinia, und Yang Liu. Strategies for applying low rank decomposition to transformer-based models. In *36th Conference on Neural Information Processing Systems (NeurIPS2022)*, 2022.

[46] Imtiaz Ali Halepoto, Umair Ali Khan, und Adnan Ahmed Arain. Retransmission Policies for Efficient Communication in IoT Applications. In *2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud)*, S. 197–202, 2018. doi: 10.1109/FiCloud.2018.00036.

[47] Hao Han, Ximing Wang, Fang lin Gu, Wen Li, Yuan Cai, Yifan Xu, und Yuhua Xu. Better Late Than Never: GAN-Enhanced Dynamic Anti-Jamming Spectrum Access With Incomplete Sensing Information. *IEEE Wireless Communications Letters*, 10 S. 1800–1804, 2021. URL: https://api.semanticscholar.org/CorpusID:236371972.

[48] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, und Jacob Steinhardt. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[49] Samuel Horvath, Stefanos Laskaridis, Mario Almeida, Ilias Leontiadis, Stylianos Venieris, und Nicholas Lane. Fjord: Fair and accurate federated learning under heterogeneous targets with ordered dropout. *Advances in Neural Information Processing Systems*, 34 S. 12876–12889, 2021.

[50] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, und Sylvain Gelly. Parameter-Efficient Transfer Learning for NLP, 2019.

[51] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, und Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, 2021.

[52] Ting Hua, Yen-Chang Hsu, Felicity Wang, Qian Lou, Yilin Shen, und Hongxia Jin. Numerical Optimizations for Weighted Low-rank Estimation on Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, S. 1404–1416, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.91. URL: https://aclanthology.org/2022.emnlp-main.91.

[53] Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Zhou Zhao, und Shinji Watanabe. AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head, 2023.

[54] Leila Ismail und Rajkumar Buyya. Metaverse: A Vision, Architectural Elements, and Future Directions for Scalable and Realtime Virtual Worlds. *ArXiv*, abs/2308.10559, 2023. URL: https://api.semanticscholar.org/CorpusID:261049832.

[55] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, *et al*. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2) S. 1–210, 2021.

[56] Andreas Kamilaris und Nicolo Botteghi. The penetration of Internet of Things in robotics: Towards a web of robotic things. *Journal of ambient intelligence and smart environments*, 12(6) S. 491–512, 2020.

[57] Diederik P Kingma und Max Welling. Auto-Encoding Variational Bayes, 2022.

[58] Constantinos Kolias, Georgios Kambourakis, Angelos Stavrou, und Jeffrey Voas. DDoS in the IoT: Mirai and other botnets. *Computer*, 50(7) S. 80–84, 2017.

[59] Xhafer Krasniqi und Edmond Hajrizi. Use of IoT technology to drive the automotive industry from connected to full autonomous vehicles. *IFAC-PapersOnLine*, 49 (29) S. 269–274, 2016.

[60] Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, und Wenhu Chen. ImagenHub: Standardizing the evaluation of conditional image generation models, 2023.

[61] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, und Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention, 2023.

[62] Chris Lattner, Mehdi Amini, Uday Bondhugula, Albert Cohen, Andy Davis, Jacques Pienaar, River Riddle, Tatiana Shpeisman, Nicolas Vasilache, und Oleksandr Zinenko. MLIR: Scaling compiler infrastructure for domain specific computation. In *2021 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, S. 2–14. IEEE, 2021.

[63] Jaejun Lee, Raphael Tang, und Jimmy Lin. What Would Elsa Do? Freezing Layers During Transformer Fine-Tuning, 2019.

[64] Yaniv Leviathan, Matan Kalman, und Yossi Matias. Fast Inference from Transformers via Speculative Decoding, 2023.

[65] Junnan Li, Dongxu Li, Silvio Savarese, und Steven Hoi. BLIP-2: Bootstrapping Language-Image Pretraining with Frozen Image Encoders and Large Language Models, 2023.

[66] Junyan Li, Li Lyna Zhang, Jiahang Xu, Yujing Wang, Shaoguang Yan, Yunqing Xia, Yuqing Yang, Ting Cao, Hao Sun, Weiwei Deng, Qi Zhang, und Mao Yang. Constraint-Aware and Ranking-Distilled Token Pruning

for Efficient Transformer Inference. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, S. 1280–1290, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599284. URL: https://doi.org/10.1145/3580305.3599284.

[67] Xiang Lisa Li und Percy Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, S. 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL: https://aclanthology.org/2021.acl-long.353.

[68] Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, und Xipeng Qiu. Unified Demonstration Retriever for In-Context Learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. 4644–4668, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.256. URL: https://aclanthology.org/2023.acl-long.256.

[69] Baohao Liao, Shaomu Tan, und Christof Monz. Make Your Pre-trained Model Reversible: From Parameter to Memory Efficient Fine-Tuning, 2023.

[70] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, und Graham Neubig. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.*, 55(9), jan 2023. ISSN 0360-0300. doi: 10.1145/3560815. URL: https://doi.org/10.1145/3560815.

[71] Xiangyang Liu, Tianxiang Sun, Junliang He, Jiawen Wu, Lingling Wu, Xinyu Zhang, Hao Jiang, Zhao Cao, Xuanjing Huang, und Xipeng Qiu. Towards efficient NLP: A standard evaluation and A strong baseline. *arXiv preprint arXiv:2110.07038*, 2021.

[72] Renze Lou, Kai Zhang, und Wenpeng Yin. Is Prompt All You Need? No. A Comprehensive and Broader View of Instruction Learning, 2023.

[73] Liangyi Luo, Kohei Ogawa, Graham Peebles, und Hiroshi Ishiguro. Towards a Personality AI for Robots: Potential Colony Capacity of a Goal-Shaped Generative Personality Model When Used for Expressing Personalities via Non-Verbal Behaviour of Humanoid Robots. *Frontiers in Robotics and AI*, 9 S. 728776, 2022.

[74] Kai Lv, Yuqing Yang, Tengxiao Liu, Qinghui Gao, Qipeng Guo, und Xipeng Qiu. Full Parameter Fine-tuning for Large Language Models with Limited Resources, 2023.

[75] Zhihan Lv. Generative Artificial Intelligence in the Metaverse Era. *Cognitive Robotics*, 2023. URL: https://api.semanticscholar.org/CorpusID:259431142.

[76] Xinyin Ma, Gongfan Fang, und Xinchao Wang. LLM-Pruner: On the Structural Pruning of Large Language Models, 2023.

[77] Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, und Sanjeev Arora. Fine-Tuning Language Models with Just Forward Passes, 2023.

[78] Aboli Marathe, Deva Ramanan, Rahee Walambe, und Ketan Kotecha. WEDGE: A multi-weather autonomous driving dataset built from generative vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, S. 3317–3326, 2023.

[79] Patrick McDaniel und Stephen McLaughlin. Security and privacy challenges in the smart grid. *IEEE security & privacy*, 7(3) S. 75–77, 2009.

[80] Bertalan Meskó und Eric J Topol. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digital Medicine*, 6(1) S. 120, 2023.

[81] Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Zeyu Wang, Rae Ying Yee Wong, Alan Zhu, Lijie Yang, Xiaoxiang Shi, Chunan Shi, Zhuoming Chen, Daiyaan Arfeen, Reyna Abhyankar, und Zhihao Jia. SpecInfer: Accelerating Generative Large Language Model Serving with Speculative Inference and Token Tree Verification, 2023.

[82] Microsoft. What is a Vector Database? https://learn.microsoft.com/en-us/semantic-kernel/memories/vector-db, November 2023. Accessed: 2023-12-13.

[83] Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, und Ahmed Awadallah. Orca: Progressive Learning from Complex Explanation Traces of GPT-4, 2023.

[84] Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, *et al.* Efficient large-scale language model training on gpu clusters using megatron-lm. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, S. 1–15, 2021.

[85] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, und Mark Chen. Point-E: A System for Generating 3D Point Clouds from Complex Prompts, 2022.

[86] Kannan Nova. Generative AI in Healthcare: Advancements in Electronic Health Records, facilitating Medical Languages, and Personalized Patient Care. *Journal of Advanced Analytics in Healthcare Management*, 7(1) S. 115–131, 2023.

[87] OpenAI. GPT-4 Technical Report, 2023.

[88] Bigi Varghese Philip, Tansu Alpcan, Jiong Jin, und Marimuthu Palaniswami. Distributed real-time IoT for autonomous vehicles. *IEEE Transactions on Industrial Informatics*, 15(2) S. 1131–1140, 2018.

[89] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, *et al.* Language models are unsupervised multitask learners. *OpenAI blog*, 1(8) S. 9, 2019.

[90] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, und Geoffrey Irving. Scaling Language Models: Methods, Analysis & Insights from Training Gopher, 2022.

[91] Abir Rahali und Moulay A Akhloufi. MalBERT: Malware Detection using Bidirectional Encoder Representations from Transformers. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, S. 3226–3231. IEEE, 2021.

[92] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, und Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, S. 8821–8831. PMLR, 2021.

[93] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, und Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022.

[94] Siyu Ren und Kenny Q. Zhu. Low-Rank Prune-And-Factorize for Language Model Compression, 2023.

[95] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, und Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, S. 10684–10695, 2022.

[96] Mohamed Sabt, Mohammed Achemlal, und Abdelmadjid Bouabdallah. Trusted Execution Environment: What It is, and What It is Not. In *2015 IEEE Trustcom/BigDataSE/ISPA*, volume 1, S. 57–64, 2015. doi: 10.1109/Trustcom.2015.357.

[97] Yunus Emre Seyyar, Ali Gökhan Yavuz, und Halil Murat Ünver. An attack detection framework based on BERT and deep learning. *IEEE Access*, 10 S. 68633–68644, 2022.

[98] Or Sharir und Anima Anandkumar. Incrementally-Computable Neural Networks: Efficient Inference for Dynamic Inputs, 2023.

[99] Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Daniel Y. Fu, Zhiqiang Xie, Beidi Chen, Clark Barrett, Joseph E. Gonzalez, Percy Liang, Christopher Ré, Ion Stoica, und Ce Zhang. FlexGen: High-Throughput Generative Inference of Large Language Models with a Single GPU, 2023.

[100] Alex Sherstinsky. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404 S. 132306, mar 2020. doi: 10.1016/j.physd.2019.132306. URL: https://doi.org/10.1016%2Fj.physd.2019.132306.

[101] Mitchell Stern, Noam Shazeer, und Jakob Uszkoreit. Blockwise Parallel Decoding for Deep Autoregressive Models. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, S. 10107–10116, Red Hook, NY, USA, 2018. Curran Associates Inc.

[102] Mingjie Sun, Zhuang Liu, Anna Bair, und Zico Kolter. A Simple and Effective Pruning Approach for Large Language Models. *arXiv preprint arXiv:2306.11695*, 2023.

[103] Shicheng Tan, Weng Lam Tam, Yuanchun Wang, Wenwen Gong, Shu Zhao, Peng Zhang, und Jie Tang. GKD: A General Knowledge Distillation Framework for Large-scale Pre-trained Language Model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, S. 134–148, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-industry.15. URL: https://aclanthology.org/2023.acl-industry.15.

[104] Tadahiro Taniguchi, Hiroshi Yamakawa, Takayuki Nagai, Kenji Doya, Masamichi Sakagami, Masahiro Suzuki, Tomoaki Nakamura, und Akira Taniguchi. A whole brain probabilistic generative model: Toward realizing cognitive architectures for developmental robots. *Neural Networks*, 150 S. 293–312, 2022.

[105] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, und Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, 2023.

[106] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Mar-

tinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, und Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023.

[107] Spyros G Tzafestas. Synergy of IoT and AI in modern society: The robotics and automation case. *Robot. Autom. Eng. J*, 31 S. 1–15, 2018.

[108] Sunil Vadera und Salem Ameen. Methods for Pruning Deep Neural Networks, 2021.

[109] S Venkatasubramanian. Ambulatory Monitoring of Maternal and Fetal using Deep Convolution Generative Adversarial Network for Smart Health Care IoT System. *International Journal of Advanced Computer Science and Applications*, 13(1), 2022.

[110] R Venkataswamy, Varaprasad Janamala, und Ravidranath Chowdary Cherukuri. Realization of Humanoid Doctor and Real-Time Diagnostics of Disease Using Internet of Things, Edge Impulse Platform, and ChatGPT. *Annals of Biomedical Engineering*, S. 1–3, 2023.

[111] Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, und Mi Zhang. Efficient Large Language Models: A Survey, 2023.

[112] Chi Wang, Susan Xueqing Liu, und Ahmed H. Awadallah. Cost-Effective Hyperparameter Optimization for Large Language Model Generation Inference, 2023.

[113] Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H Brendan McMahan, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, *et al*. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.

[114] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, und Hannaneh Hajishirzi. Self-Instruct: Aligning Language Models with Self-Generated Instructions, 2023.

[115] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, und William Fedus. Emergent Abilities of Large Language Models, 2022.

[116] Dingzhu Wen, Ki-Jun Jeon, und Kaibin Huang. Federated dropout—A simple approach for enabling federated learning on resource constrained devices. *IEEE wireless communications letters*, 11(5) S. 923–927, 2022.

[117] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, und Song Han. SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models. *arXiv*, 2022.

[118] Benfeng Xu, Quan Wang, Zhendong Mao, Yajuan Lyu, Qiaoqiao She, und Yongdong Zhang. kNN Prompting: Beyond-Context Learning with Calibration-Free Nearest Neighbor Inference. In *The Eleventh International Conference on Learning Representations*, 2023. URL: https://openreview.net/forum?id=fe2S7736sNS.

[119] Daliang Xu, Wangsong Yin, Xin Jin, Ying Zhang, Shiyun Wei, Mengwei Xu, und Xuanzhe Liu. LLMCad: Fast and Scalable On-device Large Language Model Inference, 2023.

[120] Minrui Xu, Dusit Niyato, Junlong Chen, Hongliang Zhang, Jiawen Kang, Zehui Xiong, Shiwen Mao, und Zhu Han. Generative AI-empowered simulation for autonomous driving in vehicular mixed reality metaverses. *arXiv preprint arXiv:2302.08418*, 2023.

[121] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, und Dahua Lin. PointLLM: Empowering Large Language Models to Understand Point Clouds. *arXiv preprint arXiv:2308.16911*, 2023.

[122] Hang Yang, Zekun Niu, Shilin Xiao, Jiafei Fang, Zhiyang Liu, David Fainsin, und Lilin Yi. Fast and Accurate Optical Fiber Channel Modeling Using Generative Adversarial Network. *Journal of Lightwave Technology*, 39 S. 1322–1333, 2020. URL: https://api.semanticscholar.org/CorpusID:229248300.

[123] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, und Lijuan Wang. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision), 2023.

[124] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, und Wenhu Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI, 2023.

[125] Xiao Zeng, Ming Yan, und Mi Zhang. Mercury: Efficient On-Device Distributed DNN Training via Stochastic Importance Sampling. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, SenSys '21, S. 29–41, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450390972. doi: 10.1145/3485730.3485930. URL: https://doi.org/10.1145/3485730.3485930.

[126] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, und Xia Hu. *Data-centric AI: Perspectives and Challenges*, S. 945–948. SIAM, 2023. doi: 10.1137/1.9781611977653.ch106. URL: https://epubs.siam.org/doi/abs/10.1137/1.9781611977653.ch106.

[127] Chao Zhang, Hang Zou, Samson Lasaulce, Walid Saad, Marios Kountouris, und Mehdi Bennis. Goal-Oriented Communications for the IoT and Application to Data Compression. *IEEE Internet of Things Magazine*, 5(4) S. 58–63, 2022. doi: 10.1109/IOTM.001.2200177.

[128] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, und Yu Su. MagicBrush: A Manually Annotated Dataset for Instruction-Guided Image Editing. In *Advances in*

Neural Information Processing Systems, 2023.

[129] Longteng Zhang, Lin Zhang, Shaohuai Shi, Xiaowen Chu, und Bo Li. LoRA-FA: Memory-efficient Low-rank Adaptation for Large Language Models Fine-tuning, 2023.

[130] Mi Zhang, Faen Zhang, Nicholas Lane, Yuanchao Shu, Xiao Zeng, Biyi Fang, Shen Yan, und Hui Xu. Deep Learning in the Era of Edge Computing: Challenges and Opportunities. In *Book chapter in Fog Computing: Theory and Practice, Wiley*, 2020.

[131] Qianqian Zhang, Aidin Ferdowsi, und Walid Saad. Distributed Generative Adversarial Networks for mmWave Channel Modeling in Wireless UAV Networks. *ICC 2021 - IEEE International Conference on Communications*, S. 1–6, 2021. URL: https://api.semanticscholar.org/CorpusID:231986833.

[132] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, und Luke Zettlemoyer. OPT: Open Pre-trained Transformer Language Models, 2022.

[133] Tuo Zhang, Lei Gao, Chaoyang He, Mi Zhang, Bhaskar Krishnamachari, und A Salman Avestimehr. Federated learning for the internet of things: Applications, challenges, and opportunities. *IEEE Internet of Things Magazine*, 5(1) S. 24–29, 2022.

[134] Tuo Zhang, Tiantian Feng, Samiul Alam, Mi Zhang, Shrikanth S Narayanan, und Salman Avestimehr. Gpt-fl: Generative pre-trained model-assisted federated learning. *arXiv preprint arXiv:2306.02210*, 2023.

[135] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, und Omer Levy. LIMA: Less Is More for Alignment, 2023.

[136] Qinhong Zhou, Zonghan Yang, Peng Li, und Yang Liu. Bridging the Gap between Decision and Logits in Decision-based Knowledge Distillation for Pre-trained Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. 13234–13248, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.738. URL: https://aclanthology.org/2023.acl-long.738.

[137] Xiyou Zhou, Zhiyu Chen, Xiaoyong Jin, und William Yang Wang. Hulk: An energy efficiency benchmark platform for responsible natural language processing. *arXiv preprint arXiv:2002.05829*, 2020.

[138] Zhi Zhou, Xu Chen, En Li, Liekang Zeng, Ke Luo, und Junshan Zhang. Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing. *Proceedings of the IEEE*, 107(8) S. 1738–1762, 2019. doi: 10.1109/JPROC.2019.2918951.

[139] Xunyu Zhu, Jian Li, Yong Liu, Can Ma, und Weiping Wang. A Survey on Model Compression for Large Language Models, 2023.

[140] Ivan Zyrianoff, Angelo Trotta, Luca Sciullo, Federico Montori, und Marco Di Felice. IoT Edge Caching: Taxonomy, Use Cases and Perspectives. *IEEE Internet of Things Magazine*, 5(3) S. 12–18, 2022. doi: 10.1109/IOTM.001.2200112.

[141] Marija Šakota, Maxime Peyrard, und Robert West. Fly-Swat or Cannon? Cost-Effective Language Model Choice via Meta-Modeling, 2023.