LLM Agents for Internet of Things (IoT) Applications

Akshat Bhat *akshatb5@illinois.edu*
akshatb5

Aishee Mondal *aisheem2@illinois.edu*
aisheem2

Aniket Tripathy *anikett2@illinois.edu*
anikett2

Abstract

The rapid proliferation of Internet of Things (IoT) technologies has transformed numerous sectors by enabling real-time monitoring, automation, and data-driven insights. IoT systems still struggle with challenges stemming from raw, heterogeneous data, limited contextual understanding, and rigid, rule-based analytics. Meanwhile, Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language understanding, reasoning, and human-machine interaction. Integrating large-language models with the Internet of Things (IoT) represents a rapidly evolving research area with significant potential to enhance the adaptability, intelligence, and user experience of multi-user connected IoT systems. This survey paper aims to explore the various applications of LLMs in IoT, including reviewing their role in recent advancements that integrate LLMs into IoT workflows from edge intelligence and smart interfaces to dynamic task execution and generative interaction. Furthermore, the study will examine key challenges, such as computational overhead, latency, and data privacy concerns associated with deploying LLMs in resource-constrained IoT environments. Through a comprehensive review of the existing literature and recent advancements, this paper seeks to highlight emerging trends and provide insights into the future development of LLM-driven IoT applications. We analyze prominent use cases, architectural trends, and optimization techniques, and we discuss the technical, ethical, and operational challenges that arise. Finally, we outline promising future directions for developing scalable, efficient, and context-aware IoT systems empowered by LLM agents.

1 Introduction

The Internet of Things (IoT) has significantly impacted various sectors, facilitating real-time monitoring, automation, and extensive data collection from diverse interconnected devices across healthcare, manufacturing, smart cities, and personal computing environments. Despite these advancements, IoT-generated data is typically raw, heterogeneous, and challenging to analyze, limiting the actions that automated IoT systems can perform. Traditional IoT systems often rely on rule-based analytics or domain-specific models, restricting their adaptability and contextual understanding in dynamic environments, thus limiting scalability and responsiveness (Zhao et al., 2025; De Vito et al., 2025).

Recently, Large Language Models (LLMs) such as GPT-4 and BERT have significantly advanced their abilities in sophisticated natural language processing, contextual reasoning, and extensive knowledge interpretation (Zhao et al., 2025; Zong et al., 2024). These models have demonstrated exceptional proficiency in language-centric tasks, significantly improving data interpretability, summarization, and interaction (Li et al., 2024). Notably, implementing agentic workflows using LLMs (LLM agents) enables complex task execution and intelligent interactions through natural language instructions, making them particularly suitable for IoT domains (Cui et al., 2024; Zong et al., 2024).

This intersection between LLMs and IoT is timely and crucial, driven by the complexity of modern IoT environments, the urgent need for efficient edge intelligence, and demands for intuitive human-machine interfaces (De Vito et al., 2025; Guan et al., 2024). For example, specialized prompting methods for generative IoT demonstrate significant efficiency improvements in constrained edge devices (Xiao et al., 2024). Additionally, recent research output indicates growing interest in using LLMs to address IoT-specific challenges like data management, interoperability, security, and scalability, identifying methods to optimize computational resource usage and conform to ethical considerations in IoT applications (De Vito et al., 2025; Zong et al., 2024).

Motivated by these opportunities and challenges, this survey reviews recent strategies, advancements, and limitations for integrating LLM agents into IoT domains. In particular, we explore the capability for LLM agents to autonomously in their IoT environments and the extent of human assistance required. We analyze key use cases, provide system architecture considerations, outline challenges and limitations, and propose future directions to facilitate robust, efficient, and intelligent IoT LLM agents.

2 Background

2.1 Large Language Models (LLMs)

LLMs have advanced considerably due to transformer architectures, instruction tuning, and effective tool-use mechanisms. Transformers utilize self-attention to process text, significantly improving the understanding of context and semantics (Naveed et al., 2024). Fine-tuning methods, such as instruction tuning, allow models to perform specific tasks effectively, enhancing their reasoning and interaction capabilities with limited additional training data (Parthasarathy et al., 2024). Recent advancements in tool use demonstrate LLMs' ability to interact with external APIs or software, enabling them to perform practical tasks beyond mere text generation (Zhao et al., 2025).

2.2 Internet of Things (IoT)

IoT architectures typically comprise three primary layers: edge, fog, and cloud (Domínguez-Bolaño et al., 2022). Edge computing involves processing data near the data source, reducing latency, and improving efficiency for real-time applications (Ahmed et al., 2023). Fog computing acts as an intermediary, handling computation closer to end devices to balance the computational load and improve responsiveness. Cloud computing serves as a central repository and powerful analytical engine, managing extensive computational tasks and data storage (Domínguez-Bolaño et al., 2022).

Sensors and actuators are fundamental IoT components, providing real-time environmental monitoring and enabling responsive actions based on data-driven decisions. IoT networks commonly use various communication protocols, including MQTT, HTTP, and CoAP, which facilitate effective communication between sensors, actuators, and central processing units (Domínguez-Bolaño et al., 2022).

2.3 Integration of LLM Agents in IoT

Integrating LLM agents within IoT architectures offers substantial improvements in data interpretation, predictive analytics, and system control. LLM agents can process raw sensor data, converting it into human-readable summaries and actionable insights, enhancing decision-making processes (Guan et al., 2024; Cui et al., 2024). Additionally, the multimodal capabilities of LLMs, which combine textual, numerical, and sensory inputs, significantly enrich contextual understanding, vital for effective IoT applications such as anomaly detection, predictive maintenance, and intelligent interfaces (Zong et al., 2024).

Moreover, IoT provides LLMs with real-time, multimodal data inputs, refining their context sensitivity and specificity. Examples include energy management recommendations in smart buildings based on occupancy and temperature data, and patient monitoring systems in healthcare that utilize real-time physiological data to detect health anomalies (Guan et al., 2024).

Overall, this survey comprehensively addresses these integrations, synthesizing recent research and analyzing optimal approaches for the practical deployment of LLM agents into IoT systems.

3 Related Work

Existing surveys focus on how LLMs can be integrated into existing paradigms for IoT systems to improve their performance and enable greater interactivity. Kök et al. (2024) investigate how LLMs can be integrated at the edge, fog, and cloud deployment levels of IoT systems to enhance functionalities such as by optimizing resource usage and enhancing real-time processing abilities. They focus on how this integration transforms the LLM into more perceptive systems by giving them access to IoT sensors, enabling their generative ability for IoT-specific prompts and leveraging their reasoning capabilities for IoT tasks. Zong et al. (2024) explore how LLMs can make IoT networks more intelligent and responsive for enhancing the system's security, enabling macroprogramming frameworks that treat the system of devices as a single entity, and overcoming hardware limitations for storing IoT data. Our survey builds on these works by exploring the capability for fully autonomous LLM agents that can not only improve current IoT systems, but introduce novel paradigms as well for solving IoT problems by delegating tasks to the agents.

Other surveys have investigated the capability for LLM agents tailored for specialized IoT domains. Li et al. (2024) explore the the capability for Personal LLM Agents, which have access to a user's personal devices such as smartphones and smart watches to help users obtain information and achieve goals. They focus on architectures and capabilities of these agents as well as addressing security concerns that arise from working with personal data. Ferrara (2024) covers trends and challenges for implementing LLM agents in systems of wearable sensors, covering the capability for personalized health-related applications such as coaching and physiotherapist agents. Our survey expands on these works by examining the potential for LLM agents in other IoT ecosystems that may use different types of devices and have different levels of stakes involved, introducing the unique considerations different domains incur while also synthesizing broader conclusions.

4 LLM Agents in IoT: Applications

In this section we examine and compare how LLMs can be effectively integrated into systems for different IoT domains. We focus on agentized solutions that provide automation and improved decision-making for a variety of IoT tasks.

4.1 Smart Homes

Smart homes make a natural domain for implementing LLM agents due to the existing paradigm of voice-controlled smart home assistants that act as agents by controlling various IoT devices such as smart lights and televisions to satisfy user commands.

The smart home LLM assistant Sasha leverages the natural language proficiency of LLMs to better interpret imprecise commands such as "help me sleep better," utilizing a framework for decomposing the task to clarify goals, determine appropriate devices, and produce action plans in JSON format (King et al., 2024). Sasha leverages planning and reasoning frameworks to significantly improve the raw performance of the LLM, allowing it to support complex multi-part task commands. Sasha's incorporated mechanism for iteratively improving plans from user feedback suggests that the human-in-the-loop paradigm may actually be a reasonable expectation for LLM agents in smart homes, as unlike in other domains, the human is typically near the agent and requires little effort to supervise its generated actions for new commands.

Progressing beyond generating action plans, the centralized smart home LLM assistant SAGE executes tasks using API calls to control any smart device's full functionality without needing to know device-specific code (Rivkin et al., 2025). SAGE dynamically constructs a tree framework to determine which action to take next, whether the action was successful, and determine when the task is completed. SAGE uses a long-term memory mechanism to infer user preferences over time for handling commands such as “put the game on the TV over the dresser” and generates condition code to handle persistent tasks such as “remind me to throw out the milk when I open the fridge.” Coupled with its generalization for different smart devices, SAGE’s multi-modal capability for recognizing devices from photos of the room demonstrates the LLM agent’s enhanced ability to not only perform tasks, but ease the integration of new devices and setups with minimal user configuration.

In addition to handling direct user commands, Khelifi & Morris (2024)’s framework for LLM agents in smart spaces can use context from sensors to infer user desires to provide services autonomously without human direction (Khelifi & Morris, 2024). Rather than a centralized agent for the smart space, each smart device acts as its own agent, distributing the workload for detecting when an action should be performed. A benefit of having the agents infer their own tasks is the automatic implementation of universal values, such as energy conversation by turning off the lights when on one is in the room. Furthermore, the agents can act in hybrid smart spaces implemented with augmented reality and virtual reality, synchronizing between the different realities and providing more immersive user experiences. Though further away from being deployed in the real world, these agents show the potential for smart home LLM agents that do not require humans for task recognition.

4.2 Industrial IoT

Delegating industrial IoT tasks such as data analysis, energy optimization, and real-time monitoring and decision-making to LLM agents has less precedence due to the lack of non-LLM agent parallels, requiring creativity in designing multi-agent systems for domain-specific solutions.

CityGPT can perform user-requested spatiotemporal data learning and analysis tasks by employing multiple LLM agents that are given access to specialized models tailored for their tasks Guan et al. (2024). A benefit of agentizing these tasks is to enable interactivity with complex models, broadening the range of applications the system can be used for. CityGPT’s conversational interface helps bridge the gap between complex IoT-generated data and human-interpretable knowledge, enabling the agents to work alongside humans to improve the efficacy of crucial tasks such as evaluating and predicting air quality.

Beyond collaboration with humans, CASIT facilitates a system of LLM agents that can run by itself for assessing abnormalities in IoT sensor data, enabling automation for tasks such as maintaining livable conditions in natural habitats (Zhong et al., 2024). CASIT’s multiple agent framework significantly reduces data transmission volumes from the environment’s sensors to the data servers by using a Chairman Agent to choose the necessary information from Data Analyst agents to transmit and the appropriate means for transmission, reducing costs and alleviating bandwidth limitations for massive IoT data. This reveals the capacity for fully autonomous LLM agents in distant environments from humans, as the agents can make remote communication more manageable.

More progress toward fully autonomous industrial IoT agents can be achieved using the GPT-in-the-loop approach for drawing from environmental feedback to enhance decision making and adaptability (Nascimento et al., 2023). The iterative improvements from GPT-in-the-loop enables streetlight agents to make decisions for optimizing energy usage while providing adequate lighting at a human-level performance after a few attempts without the need for extensive training. Thus, the integration of LLM agents for industrial IoT applications can have varying levels of reliance on humans and is highly task-dependent, but ultimately supports working alongside humans as well as executing remote tasks.

4.3 Task Scheduling and Smart Devices

LLMs have achieved impressive results in textual and visual domains, but still struggle to replicate those results in the physical world, revealing a gap in their understanding of specific outputs that follow physical constraints. To replicate human cognition in complex tasks where perception is fundamental to reasoning, LLMs can be augmented with enhanced perception abilities using IoT sensor data and pertinent knowledge for IoT task reasoning in the physical domain. (An et al., 2024)

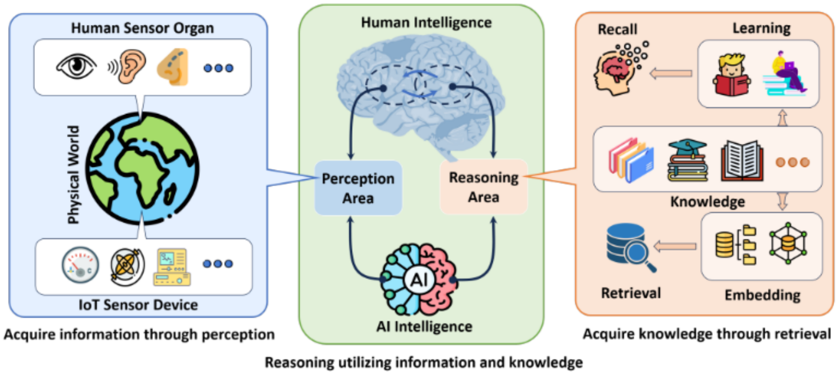

Figure 1: Augmenting LLMs with physical world perception from IOT data (An et al., 2024)

Goal-oriented communication is an important aspect of existing smart IoT devices, which have become limited in their capacity to handle complex tasks, and especially in their interactions with humans. LLMind is an LLM-based task-oriented AI agent framework that enables effective collaboration among IoT devices, with humans communicating high-level verbal instructions to perform complex tasks. (Cui et al., 2024) The exceptional logical reasoning and linguistic capabilities of LLMs can be leveraged to achieve seamless coordination among diverse IoT devices and perform complex tasks in an overall advanced intelligence IoT system, thus enabling human users to control multiple IoT devices simultaneously through various media, including text, voice, video, and virtual reality.

In the case of smart devices, the LLM agent engages in conversation with users by means of a chatting interface or verbal commands to generate contextual information regarding the tasks for the system. The LLM takes over the planning tasks and generates control scripts utilizing specialized AI models and interacting with IoT devices by sending control commands over network connections. To ensure efficient control of diverse IoT devices in LLMind across multi-modal inputs, manufacturers must provide well-documented API functions. These functions should encapsulate the necessary functionality, enabling the AI agent to initiate specific actions or retrieve information from the IoT devices. (Cui et al., 2024) This design then enables the system to adapt to different AI modules and IoT devices without requiring extensive modifications to the LLM itself, enhancing scalability and interoperability.

The rise of LLMs has also led to an increase in deployment of LLMs-based agents on personal IoT devices. These agents perform daily tasks by understanding user intentions, gathering information, making decisions, and taking autonomous actions. On smart devices, these agents could autonomously operate mobile applications, call different APIs, and use various sensors to perform tasks. However, this integration poses a challenge in effectively displaying information during task execution and ensuring users are informed about the operations and the outcomes of the tasks they desire. Wen et al. (2024) proposes a system consisting of a task planning agent, a UI navigating agent, and a UI reassembling agent. Given the high level task that may involve several applications, the task planner agent divides it into sub-tasks. Then the UI navigating agent can complete each sub-task on one application, and record the important UI elements for display. Finally, the UI reassemble agent constructs a user-friendly UI layout based on these recorded UI elements. Thus, LLM agents here can be utilized as a personalization agent to customize UI elements for apps on your own smartphone.

LLM's reasoning ability can also be enhanced using IoT data, into the novel unified framework IoT-LLM for task reasoning (An et al., 2024). IoT-LLM is composed of three steps tailored for IoT reasoning: designing an LLM-friendly data format, activating knowledge by chain-of-thought prompting, and automatic IoT-oriented Retrieval-Augmented Generation (RAG) based on LLMs' in-context learning capability. We can also determine if LLMs truly understand and solve the task they were assigned to by generating analytical processes and analyzing the reasonableness of the analytics. The analysis generated by IoT-LLM indicates that LLMs can provide a reasonable process of solving simple tasks, but their efficacy diminishes in more specialized domains like heartbeat anomaly detection. This performance disparity is attributable to the complexity of data and the limited domain-specific knowledge inherent in LLMs.

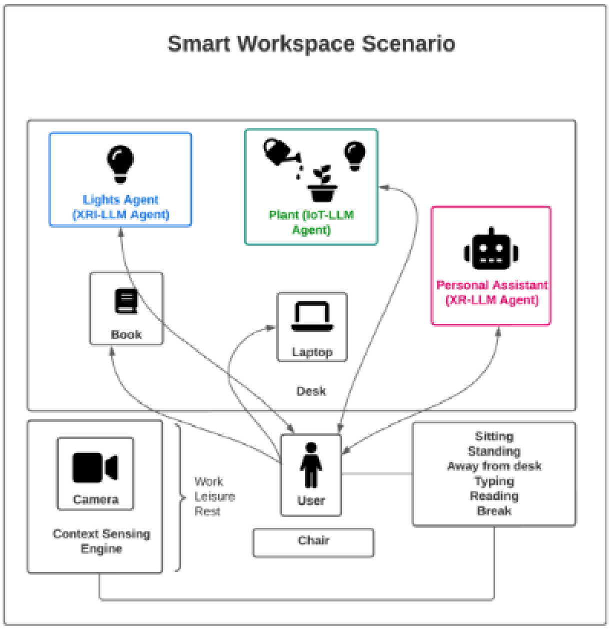

Figure 2: Smart Workspace Scenario
(Khelifi & Morris, 2024)

Khelifi & Morris (2024) explore the opportunity to design Smart Space (SS) systems that combine Extended Reality-based human-computer interactions with IoT and LLM-based task planning and control, and using contextual information to adapt to both the physical and virtual world's behavior and objects. The solution in this framework usually consists of several components as shown in Figure 2. The user which has several states, tasks, and goals; the Context Engine, which is the environment; and then the Goal-Driven Embodied Agents.

This framework shows how the intersection of Extended Reality, LLMs, and IoT technologies can provide intelligent, context-aware, and immersive interactions between the user and IoT devices that also take into account the user contexts and enhance user experience within Smart Spaces.

4.4 Healthcare

IoT-based sensors play a crucial role in continuous health monitoring, enabling the early detection of medical conditions and the management of chronic diseases. These sensors utilize advanced technologies to detect potential health issues and trigger timely alerts, operating on the principle of preventive healthcare, ensuring patient safety and improved health management. Further, by analyzing data from wearable motion and physiological sensors, systems can classify various physical activities, which is valuable in fitness tracking (especially for athlete training progress and preventing injuries), rehabilitation, and elder care.

LLMs trained on diverse datasets exhibit potential for handling complex health wearables sensor data, transforming raw sensor data into structured information to extract meaningful insights for analysis. This process is critical in Human Activity Recognition (HAR), health monitoring, and behavior analysis (Ferrara, 2024). The raw data, particularly in the case of multiple and combinations of physiological sensors, tends to be high dimensional. Dimensionality reduction techniques, such as Principal Component Analysis (PCA) and feature selection methods, are employed to manage this complexity and make the data more manageable for analysis.

Recent promising applications of LLMs in the analysis of wearable sensor data involve the models processing and analyzing multi-modal data, including text, audio, and sensor signals, offering a more comprehensive understanding essential for generating accurate healthcare decisions. Ferrara (2024) introduces the health monitoring system PhysioLLM, which integrates physiological data from wearables with contextual information to provide personalized health insights. It has been instrumental in chronic disease management, enhancing users' understanding of health data, supporting actionable health goals by continuously monitoring vital signs and alerting users and healthcare providers about potential health issues, thus facilitating early intervention and improving patient outcomes. In the case of HAR, HARGPT leverages LLMs in classifying human activities based on sensor data collected about the user's movements, body positions, and physiological responses, and outperforms traditional machine learning models in recognizing activities from raw IMU data, achieving high accuracy, even on unseen data.

Additional improvements in real-time feedback and intervention mechanisms in wearable devices, such as biofeedback-enabled wearables that monitor physiological parameters, can provide immediate feedback to users and promote healthier behaviors and overall well-being. LLMs are also used to improve the accuracy of activity recognition systems by detecting small variations in movement patterns, which is crucial in applications like rehabilitation and physical therapy.

Xin Liu (2023) presents the Health-LLM framework, which evaluates various LLM architectures for health prediction tasks using wearable sensor data. The study highlights the effectiveness of LLMs in predicting health-related outcomes such as heart rate variability, stress levels, and sleep patterns. Health-LLM utilizes prompting and fine-tuning techniques to adapt the models to specific health tasks, providing comprehensive and personalized insights.

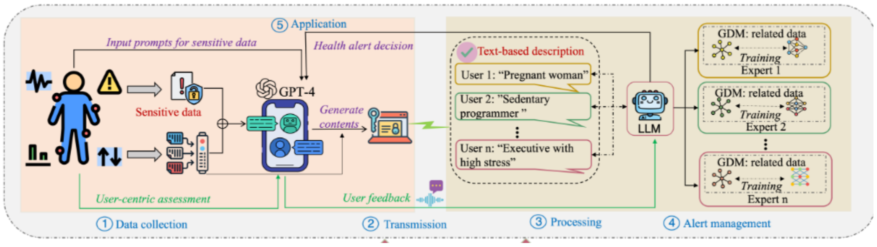

Figure 3: Architecture of the proposed LLM-HAS framework
(Gao et al., 2024)

Gao et al. (2024) introduces LLM-HAS, a framework that incorporates large language models into healthcare alert systems to improve accuracy, ensure user privacy, and enhance personalized health service, while maintaining the subjective quality of experience (QoE) of users. Healthcare Alert Systems take advantage of LLMs like GPT-4 on the user side and incorporate an LLM-enabled Mixture of Experts (MoE) framework on the edge server, specifically to reduce the probability of missed alerts (MA) and false alerts (FA).

LLM agents integrate deep reinforcement learning (DRL) in these healthcare alert systems to analyze extensive datasets and learn from diverse inputs, including patient behaviorial patterns, enabling the models to not only be reactive, but also predictive, thus providing a more nuanced decision-making approach by analyzing previous alert outcomes to reduce inaccuracies and personalize health management strategies. (Gao et al., 2024)

Since privacy protection and integrity of user health data is an important aspect of LLM-enabled Healthcare Alerts, the LLM-HAS utilizes a robust approach of collecting data from the wearable sensors, uploading mixed-content data to the edge server (the hub node), and then discerning which data should be reconstructed - such as sensitive health data (removing any personal identifying markers) by GPT-4 from prompts (Gao et al., 2024). Healthcare alert systems like LLM-HAS can also process conversational user feedback, enhancing the accuracy of health alert decisions by finetuning the algorithm based on conversational user feedback.

Building on this foundation, future goals of development could be an autonomous LLM-enabled smart healthcare system that can leverage advanced capabilities of LLMs to more efficiently understand user intentions. This model would retain its ability to make informed decisions about health alerts, adapt to new data, deliver timely health interventions, and also align with the evolving needs of smart healthcare infrastructure.

4.5 Cybersecurity and Privacy

The integration of LLMs within IoT systems presents significant advancements for cybersecurity and privacy, effectively addressing increasingly sophisticated and dynamic threats. Recent research highlights innovative methods employing transformer-based architectures, notably leveraging BERT, to enhance cybersecurity management and data privacy in IoT ecosystems.

Ferrag et al. (2024) introduces *SecurityBERT*, a lightweight, privacy-preserving BERT-based model specifically tailored for IoT and Industrial IoT (IIoT) environments. SecurityBERT utilizes a novel Privacy-Preserving Fixed-Length Encoding (PPFLE) technique alongside Byte-level Byte-Pair Encoding (BBPE) for tokenization, significantly enhancing detection accuracy and reducing computational overhead. The model was validated against the Edge-IIoTset cybersecurity dataset, achieving an accuracy of 98.2%. A notable benefit of SecurityBERT is its suitability for deployment on resource-constrained edge devices, owing to its reduced model size and efficient inference time. However, a limitation is the trade-off between model complexity and computational requirements, particularly impacting the depth of anomaly detection capabilities when compared with larger, more resource-intensive models.

Worae et al. (2024) presents a comprehensive framework combining context-driven LLMs for IoT management with a fine-tuned anomaly detection component based on BERT. This unified framework leverages Retrieval-Augmented Generation (RAG) to dynamically integrate contextual knowledge from IoT administrative documents, significantly enhancing the precision and reliability of responses to IoT administrative queries. The anomaly detection module demonstrated exceptional accuracy (99.87%) using the Edge-IIoTset dataset, effectively identifying subtle network anomalies and potential security threats. The primary benefit of this approach lies in its ability to provide contextually accurate and reliable administrative support while proactively managing cybersecurity threats. Nevertheless, its dependence on external knowledge repositories for context augmentation introduces potential latency issues and raises concerns about data consistency and availability.

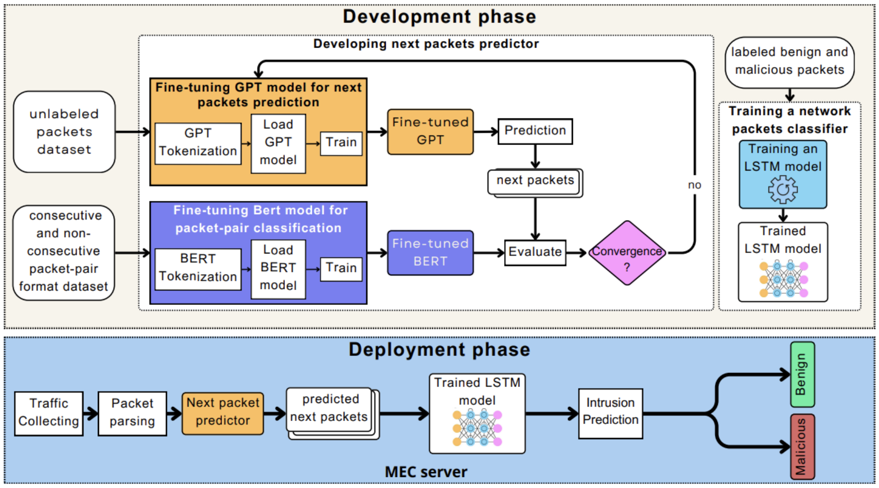

Figure 4: Workflow of the intrusion prediction framework leveraging GPT and BERT models (Diaf et al., 2024).

Further advancing predictive capabilities in IoT cybersecurity, Diaf et al. (2024) explores the use of BERT-based models for the prediction of cyber-attacks within IoT networks. Their predictive approach enhances proactive cybersecurity strategies by recognizing and forecasting cyber-attack patterns, enabling preemptive measures against threats. The main benefit of their method is the substantial reduction in response latency during cyber-incidents, significantly enhancing system resilience. However, the predictive accuracy heavily depends on historical data availability and quality, potentially limiting its effectiveness in rapidly evolving or entirely novel threat scenarios. The workflow of their predictive model, shown in Figure 4, illustrates the integration of fine-tuned GPT and BERT models to predict network traffic and classify packet pairs, respectively.

Adjewa et al. (2024) proposes an optimized federated intrusion detection system employing a BERT-based architecture designed specifically for 5G IoT ecosystems. This federated approach facilitates decentralized learning among edge devices, ensuring data privacy compliance while achieving robust accuracy (97.79%). One key advantage is the method's compatibility with resource-limited devices due to linear quantization techniques, leading to substantial model size reductions. Nevertheless, federated learning introduces complexity in managing heterogeneous and non-identically distributed (non-IID) data, potentially affecting model convergence and overall performance.

Hasan et al. (2024) investigates distributed threat intelligence at the edge using large language models, showcasing the benefits of decentralizing security threat detection across edge devices. Their approach significantly improved real-time responsiveness and reduced network traffic by processing threat intelligence locally. However, a notable limitation is the inherent variability in edge device capabilities, which can result in inconsistent threat detection performance across different devices.

Collectively, these studies underscore pivotal advancements in the use of LLMs for cybersecurity and privacy within the IoT, emphasizing efficient, contextually aware, and privacy-preserving methodologies. Despite the limitations described, the ability of these models to dynamically integrate contextual information, predict threats proactively, and operate within decentralized frameworks represents a critical evolution toward secure and resilient IoT ecosystems.

4.6 Edge Computing

LLMs are resource-intensive and pose significant challenges for deployment on constrained edge devices. Recently, several works have explored strategies to efficiently run LLMs on edge devices by distributing workloads and optimizing model deployments.

Zhang et al. (2024a) introduces *EdgeShard*, a framework designed to optimize the inference of LLMs by partitioning models into shards deployed across heterogeneous edge devices and cloud servers. EdgeShard employs adaptive device selection and dynamic model partitioning algorithms based on device capabilities and network conditions. Their approach significantly reduces inference latency and improves throughput by strategically leveraging device heterogeneity and parallel computation (Fig. 5). Experimentally, EdgeShard achieved up to 50% latency reduction and a two-fold throughput improvement compared to baseline methods. Despite these advantages, the system's effectiveness relies heavily on stable network connectivity between edge and cloud components, which may introduce challenges in environments with highly variable network conditions.

To facilitate efficient cross-domain knowledge transfer at the edge, Zhou et al. (2024) proposes GenG, a generic time-series data generation method leveraging LLMs and diffusion models for edge intelligence. GenG decomposes the generation task into abstract textual understanding via fine-tuned LLMs and detailed time-series synthesis via diffusion models. It employs a two-stage generation process, ensuring consistency and controllability of generated data through abstract and detailed guidance signals. This method significantly enhances data fidelity and generation efficiency in resource-constrained edge scenarios. Nevertheless, GenG's complexity, including dual-stage processes and cross-domain transfers, may pose additional computational overhead, potentially restricting real-time applicability on extremely constrained edge nodes.

In addressing cybersecurity at edge devices, Hasan et al. (2024) developed a decentralized threat intelligence approach leveraging LLM-driven lightweight models deployed directly onto edge devices. These models analyze local network traffic and system logs in real-time, detecting and mitigating cybersecurity threats efficiently. The decentralized framework enhances privacy by processing data locally and improves responsive-

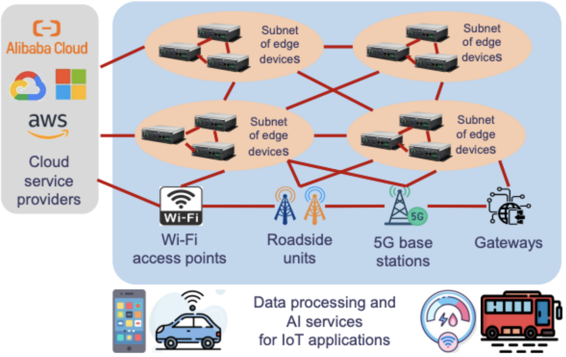

Figure 5: Collaborative edge computing integrating geo-distributed edge devices and cloud servers for efficient LLM inference (Zhang et al., 2024a).

ness by reducing latency. Additionally, the system incorporates peer-to-peer secure communication among devices, facilitating dynamic threat mitigation. However, one significant limitation is the variability in the capabilities of edge devices, potentially leading to inconsistent threat detection and mitigation effectiveness across devices.

Collectively, these studies highlight critical advancements toward effective deployment and utilization of LLMs in edge environments. They illustrate methods for overcoming resource limitations, enhancing real-time performance, and improving security and privacy. Nonetheless, achieving a balance between computational efficiency, real-time responsiveness, and robustness remains a persistent challenge, particularly given the inherent constraints and heterogeneity of edge computing environments.

5 System Architecture Considerations

The integration of LLM agents into IoT systems necessitates careful consideration of deployment architectures, tool integrations, model optimization techniques, communication protocols, and performance trade-offs.

5.1 Deployment Architectures: Edge, Cloud, and Hybrid Approaches

LLM agents can be deployed across various architectures:

- **Edge Computing**: Deploying LLMs directly on IoT devices ensures low latency and real-time responsiveness. However, this approach is constrained by the limited computational resources and energy availability of edge devices (Rondanini et al., 2025).

- **Cloud Computing**: Utilizing cloud servers offers substantial computational power and storage, facilitating the deployment of large-scale LLMs. The trade-off includes increased latency and potential privacy concerns due to data transmission over networks (Kök et al., 2024).

- **Hybrid Models**: Combining edge and cloud computing leverages the strengths of both approaches. Lightweight models can handle immediate tasks on the edge, while more complex processing is offloaded to the cloud, optimizing resource utilization and performance. (Kök et al., 2024)

5.2 Integration of APIs and Tools

The deployment of LLM agents in IoT systems is facilitated by various tools and frameworks:

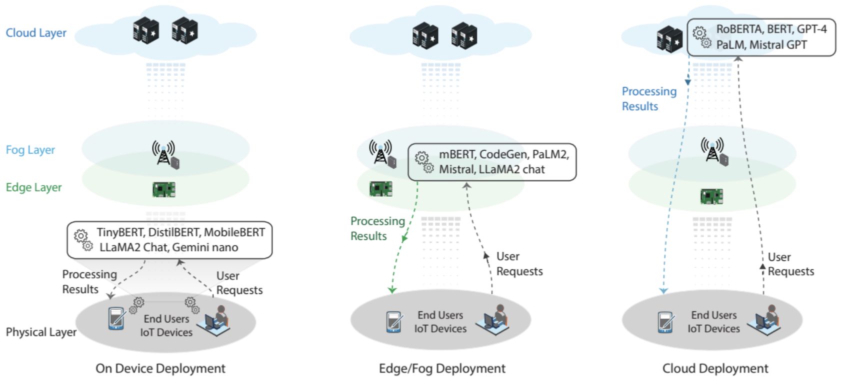

Figure 6: Integration of LLMs within IoT architectures encompassing edge, fog, and cloud computing paradigms (Kök et al., 2024).

- **LangChain**: An open-source framework that enables the development of applications powered by LLMs through composability. It allows for the integration of various data sources and APIs, streamlining the creation of complex applications.

- **OpenAgents**: A platform that provides a suite of tools for building and managing LLM-based agents, offering functionalities such as memory management, tool usage, and multi-agent collaboration (Xie et al., 2023).

5.3 Model Optimization: Lightweight Models and Quantization

To accommodate the resource constraints of edge devices, model optimization techniques are employed:

- **Quantization**: Reducing the precision of model parameters (e.g., from 32-bit to 8-bit) decreases model size and computational requirements, enabling deployment on devices with limited resources. EdgeQAT introduces entropy and distribution-guided quantization-aware training to optimize lightweight LLMs for edge deployment (Shen et al., 2024).

- **Pruning and Distillation**: Removing redundant parameters (pruning) and transferring knowledge from larger models to smaller ones (distillation) help in creating efficient models suitable for edge deployment (Zheng et al., 2025).

5.4 Communication Protocols: MQTT, RESTful APIs, and Federated Learning

Effective communication is vital for the functionality of LLM agents in IoT systems:

- **MQTT (Message Queuing Telemetry Transport)**: A lightweight, publish-subscribe network protocol that transports messages between devices, ideal for low-bandwidth, high-latency, or unreliable networks. The SDFLMQ framework utilizes MQTT to facilitate semi-decentralized federated learning at the edge (Ali-Pour & Gascon-Samson, 2025).

- **RESTful APIs**: Representational State Transfer (REST) APIs provide a standardized way for LLM agents to interact with web services, facilitating interoperability and scalability.

- **Federated Learning**: A decentralized approach where models are trained across multiple devices without exchanging data, enhancing privacy and reducing communication overhead (Ali-Pour & Gascon-Samson, 2025).

5.5 Performance Trade-offs

Deploying LLM agents in IoT environments involves balancing various performance aspects:

- **Latency vs. Accuracy**: Edge deployments offer low latency but may compromise on model complexity and accuracy. Cloud deployments provide higher accuracy at the cost of increased latency (Kök et al., 2024).
- **Energy Consumption vs. Computational Power**: Edge devices have limited energy resources, necessitating efficient models, whereas cloud servers can handle energy-intensive computations (Rondanini et al., 2025).
- **Privacy vs. Data Accessibility**: Processing data on the edge enhances privacy but limits the availability of comprehensive datasets that can be leveraged in the cloud for improved model performance (Kök et al., 2024).

6 Challenges and Limitations

Challenges for LLMs, specifically in the context of IoT, include biases in the lack of up-to-date information, privacy concerns, high computational power and memory consumption requirements, and the need for task specificity. (Zong et al., 2024)

6.1 Need for Task Specificity

Currently, a general LLM cannot handle highly specialized tasks like object detection or facial recognition, which can already be performed well by AI models in these domains. (Cui et al., 2024) Thus, there exist challenges in combining the power of specialized but fragmented AI models within the overall general LLM to perform complex tasks. It is also extremely difficult to build IoT-specific LLM architecture that efficiently utilizes domain-specific data while managing the unique dynamic states of the environment and addressing the constraints of the IoT devices (Kök et al., 2024). An architecture such as this could enhance performance by improving both the accuracy and efficiency of data processing and decision-making within IoT systems.

6.2 Performance in Dynamic Environment

IoT environments are also inherently dynamic, and models must be able to adjust to fluctuating data, device conditions, and network states. Kök et al. (2024) As discussed in the survey paper, LLMs have shown strong performance in real-world IoT datasets and virtual representations with domain-specific knowledge, but struggle to adapt to rapidly changing environments and states. A significant challenge in these domains involve efficient control of IoT devices through task-oriented communications by the LLM agent based on verbal user commands.(Cui et al., 2024) The LLM must accurately learn all the diverse functionalities and operational characteristics of various IoT devices, to better analyze and predict device behavior according to user-specified and environment constraints.

6.3 Computational Power and Memory Cost

One of the biggest challenges in integrating LLMs in smart devices is hardware limitations, which make it difficult to handle the large model size (Kök et al., 2024). Current solutions like edge-cloud collaborative devices for model partitioning introduce latency issues which affect real-time effectiveness. In addition, there are limited applications of advanced intelligence algorithms applied to edge devices, likely due to hardware limitations and computational costs.

According to Li et al. (2024), in the context of LLMs for sensor macro-processing, the significant computational cost of processing large amounts of raw and complex sensor data interferes with the limited capacity of the LLM to understand prompts which guide sensor data in plain text. Cui et al. (2024) suggests that due to LLMs' slow inference speed and high computational costs, future research needs to focus on figuring out methods to enhance the system response speed and overall efficiency.

6.4 Security and Privacy Concerns

Despite significant progress in deploying LLM-based systems for cybersecurity within IoT environments, multiple challenges remain, particularly concerning latency, real-time responsiveness, and resource constraints. Intrusion detection systems and predictive security models leveraging large language models often encounter difficulties in maintaining real-time threat detection and response due to the high computational demands inherent LLM inference (Ferrag et al., 2024; Adjewa et al., 2024). These latency concerns become especially pronounced in resource-limited edge devices, which may lack sufficient memory and processing capabilities to execute computationally intensive models efficiently, potentially compromising responsiveness and leaving systems vulnerable during critical security incidents (Hasan et al., 2024; Zhang et al., 2024a).

Deploying LLM-based cybersecurity at the edge faces major energy and memory constraints, as most models demand significant resources for continuous threat detection (Zhou et al., 2024; Zhang et al., 2024a). These limitations restrict model complexity and degrade performance across heterogeneous devices, especially in federated settings (Adjewa et al., 2024).

Security and privacy risks also persist. LLMs are prone to adversarial attacks, hallucinations, and data leaks, which threaten system trust and reliability. Federated learning, while privacy-friendly, increases complexity and risk of data exposure due to distributed data handling (Ferrag et al., 2024; Diaf et al., 2024; Hasan et al., 2024).

7 Open Problems and Future Directions

7.1 Edge-Native LLMs

Deploying LLMs at or near the source of data generation is a crucial paradigm for achieving effective LLM agents for IoT. Relying on centralized cloud servers for LLM procedures typically does not meet the latency requirements for many applications, especially for industrial IoT systems that have distributed multi-agent frameworks, are computationally intensive, and need real-time decision-making. Another potential benefit is the increased security and privacy that is enabled by processing data locally, which is particularly important for healthcare and personal LLM agents that work with sensitive data. The main challenge of deploying LLMs closer to the sensors is limited resources of smaller devices in contrast to the massive amount of parameters in traditional LLMs. Current implementations of LLM agents for smaller devices employ edge-cloud collaboration techniques for distributing model resources, however this approach lessens the capacity for real-time responsiveness (Kök et al., 2024). Developing edge-native LLMs can provide a more effective and robust solution. One ongoing attempt at a compact LLM that can run on edge devices is the open-source TinyLlama project, which aims to pretrain a 1.1 billion parameter Llama model (Zhang et al., 2024b).

7.2 Standard Evaluation Protocols

Developing standard methods for evaluating the performance of LLM agents in various IoT domains is ambitious due to the specific nature of the tasks current agents are designed for. However, some domains such as smart homes lend more natural opportunities for creating evaluation methods due the universal requirements of different implementations, leading to new benchmarks being created for challenging smart home tasks (Rivkin et al., 2025). Creating standard benchmarks for other IoT domains can help accelerate unified progress in the field by clearly defining end goals for the capabilities of the agents and enabling the assessment of different frameworks and techniques. Additionally, standard evaluation protocols may encourage the development of more general agents that can adapt to different IoT tasks, which could be a more robust paradigm.

7.3 Cost-Effective Agent Implementations

Current implementations of LLM agents often leverage frameworks for dividing tasks into sub-tasks to improve performance on complex tasks, however this significantly increases the amount of LLM inferences required. Additionally, several implementations leverage API calls to avoid needing to learn device-specific code for each device in the system, incurring additional costs. These costs add up quickly, especially for agents that need to persistently execute tasks and perform real-time monitoring. One solution being explored is finetuning the LLM for learning more specific tasks within the appropriate IoT domain to reduce the amount of tokens needed for in-context learning, which has the additional benefit of alleviating current failure modes of the agent failing to respect given instructions due to lengthy prompts (Rivkin et al., 2025).

7.4 LLM Agents for Smart Cities

An additional IoT domain that has gained recent traction is integrating LLM agents into smart cities. For instance, LLM agents have the potential to act as dispatcher agents to plan and coordinate the paths of autonomous vehicles, such as firetrucks and goods transportation Chen et al. (2024). These implementations are inherently more difficult to develop due to the infeasibility of testing real-world applications outside of lower-stake simulations, which cannot entirely emulate logistic and safety concerns. However, findings from LLM agents in other IoT domains suggest that integrating them into real-world smart cities can also improve energy efficiency improve traffic flow through real-time monitoring and predicting.

8 Conclusion

The emergence of LLMs presents new opportunities for enhancing IoT applications, displaying the potential to revolutionize the way IoT tasks are executed through the introduction of LLM agents. In this paper, we investigate key applications of LLMs into different IoT domains by comprehensively reviewing recent literature to understand the current capability of LLM agents to transform paradigms for solving IoT tasks. By exploring use cases in different IoT environments such as smart homes, industries, personal devices, and healthcare, we provide a comprehensive comparison of key implementation strategies as well as the unique challenges each domain introduces. In addition, we summarize common system architecture considerations across the different domains to reveal broader trends for LLM agents in IoT. Finally, we conclude the paper by identfying current challenges and limitations for deploying LLM agents in IoT and suggest open problems and directions for future research.

References

Frederic Adjewa, Moez Esseghir, and Leila Merghem-Boulahia. Efficient federated intrusion detection in 5g ecosystem using optimized bert-based model, 2024. URL `https://arxiv.org/abs/2409.19390`.

Shams Forruque Ahmed, Shanjana Shuravi, Shaila Afrin, Sabiha Jannat Rafa, Mahfara Hoque, and Amir H. Gandomi. The power of internet of things (iot): Connecting the dots with cloud, edge, and fog computing, 2023. URL `https://arxiv.org/abs/2309.03420`.

Amir Ali-Pour and Julien Gascon-Samson. Sdflmq: A semi-decentralized federated learning framework over mqtt, 2025. URL `https://arxiv.org/abs/2503.13624`.

Tuo An, Yunjiao Zhou, Han Zou, and Jianfei Yang. Iot-llm: Enhancing real-world iot task reasoning with large language models, Oct 2024. URL `https://arxiv.org/abs/2410.02429`.

Ruiqing Chen, Wenbin Song, Weiqin Zu, ZiXin Dong, Ze Guo, Fanglei Sun, Zheng Tian, and Jun Wang. An llm-driven framework for multiple-vehicle dispatching and navigation in smart city landscapes. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2147–2153, May 2024. doi: 10.1109/icra57147.2024.10610578.

Hongwei Cui, Yuyang Du, Qun Yang, Yulin Shao, and Soung Chang Liew. Llmind: Orchestrating ai and iot with llm for complex task execution. *IEEE Communications Magazine*, pp. 1–7, 2024. doi: 10.1109/MCOM.002.2400106.

Gabriele De Vito, Fabio Palomba, and Filomena Ferrucci. The role of large language models in addressing iot challenges: A systematic literature review. *Future Generation Computer Systems*, pp. 107829, 2025. ISSN 0167-739X. doi: https://doi.org/10.1016/j.future.2025.107829. URL `https://www.sciencedirect.com/science/article/pii/S0167739X25001244`.

Alaeddine Diaf, Abdelaziz Amara Korba, Nour Elislem Karabadji, and Yacine Ghamri-Doudane. Beyond detection: Leveraging large language models for cyber attack prediction in iot networks, 2024. URL `https://arxiv.org/abs/2408.14045`.

Tomás Domínguez-Bolaño, Omar Campos, Valentín Barral, Carlos J. Escudero, and José A. García-Naya. An overview of iot architectures, technologies, and existing open-source projects. *Internet of Things*, 20:100626, 2022. ISSN 2542-6605. doi: https://doi.org/10.1016/j.iot.2022.100626. URL `https://www.sciencedirect.com/science/article/pii/S254266052200107X`.

Mohamed Amine Ferrag, Mthandazo Ndhlovu, Norbert Tihanyi, Lucas C. Cordeiro, Merouane Debbah, Thierry Lestable, and Narinderjit Singh Thandi. Revolutionizing cyber threat detection with large language models: A privacy-preserving bert-based lightweight model for iot/iiot devices, 2024. URL `https://arxiv.org/abs/2306.14263`.

Emilio Ferrara. Large language models for wearable sensor-based human activity recognition, health monitoring, and behavioral modeling: A survey of early trends, datasets, and challenges. Jul 2024. doi: 10.20944/preprints202407.0970.v1.

Yulan Gao, Ziqiang Ye, Ming Xiao, Yue Xiao, and Dong In Kim. Guiding iot-based healthcare alert systems with large language models. 2024.

Qinghua Guan, Jinhui Ouyang, Di Wu, and Weiren Yu. Citygpt: Towards urban iot learning, analysis and interaction with multi-agent system, 2024. URL `https://arxiv.org/abs/2405.14691`.

Syed Mhamudul Hasan, Alaa M. Alotaibi, Sajedul Talukder, and Abdur R. Shahid. Distributed threat intelligence at the edge devices: A large language model-driven approach. In *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 1496–1497. IEEE, July 2024. doi: 10.1109/compsac61105.2024.00206. URL `http://dx.doi.org/10.1109/COMPSAC61105.2024.00206`.

Syrine Khelifi and Alexis Morris. Mixed reality iot smart environments with large language model agents. In *2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS)*, pp. 1–7, 2024. doi: 10.1109/ICHMS59971.2024.10555610.

Evan King, Haoxiang Yu, Sangsu Lee, and Christine Julien. Sasha: Creative goal-oriented reasoning in smart homes with large language models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–38, March 2024. ISSN 2474-9567. doi: 10.1145/3643505. URL `http://dx.doi.org/10.1145/3643505`.

İbrahim Kök, Orhan Demirci, and Suat Özdemir. When iot meet llms: Applications and challenges. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 7075–7084, 2024. doi: 10.1109/BigData62323.2024.10825187.

Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, and et al. *Personal LLM Agents: Insights and survey about the capability, efficiency and security.* 2024.

Nathalia Nascimento, Paulo Alencar, and Donald Cowan. Gpt-in-the-loop: Adaptive decision-making for multiagent systems, 2023. URL `https://arxiv.org/abs/2308.10435`.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models, 2024. URL `https://arxiv.org/abs/2307.06435`.

Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities, 2024. URL `https://arxiv.org/abs/2408.13296`.

Dmitriy Rivkin, Francois Hogan, Amal Feriani, Abhisek Konar, Adam Sigal, Xue Liu, and Gregory Dudek. Aiot smart home via autonomous llm agents. *IEEE Internet of Things Journal*, 12(3):2458–2472, 2025. doi: 10.1109/JIOT.2024.3471904.

Christian Rondanini, Barbara Carminati, Elena Ferrari, Antonio Gaudiano, and Ashish Kundu. Malware detection at the edge with lightweight llms: A performance evaluation, 2025. URL `https://arxiv.org/abs/2503.04302`.

Xuan Shen, Zhenglun Kong, Changdi Yang, Zhaoyang Han, Lei Lu, Peiyan Dong, Cheng Lyu, Chih hsiang Li, Xuehang Guo, Zhihao Shu, Wei Niu, Miriam Leeser, Pu Zhao, and Yanzhi Wang. Edgeqat: Entropy and distribution guided quantization-aware training for the acceleration of lightweight llms on the edge, 2024. URL `https://arxiv.org/abs/2402.10787`.

Hao Wen, Wenjie Du, Yuanchun Li, and Yunxin Liu. Poster: Enabling agent-centric interaction on smartphones with llm-based ui reassembling. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*, MOBISYS '24, pp. 706–707, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705816. doi: 10.1145/3643832.3661432. URL `https://doi.org/10.1145/3643832.3661432`.

Daniel Adu Worae, Athar Sheikh, and Spyridon Mastorakis. A unified framework for context-aware iot management and state-of-the-art iot traffic anomaly detection, 2024. URL `https://arxiv.org/abs/2412.19830`.

Bin Xiao, Burak Kantarci, Jiawen Kang, Dusit Niyato, and Mohsen Guizani. Efficient prompting for llm-based generative internet of things. *IEEE Internet of Things Journal*, pp. 1–1, 2024. ISSN 2372-2541. doi: 10.1109/jiot.2024.3470210. URL `http://dx.doi.org/10.1109/JIOT.2024.3470210`.

Tianbao Xie, Fan Zhou, Zhoujun Cheng, Peng Shi, Luoxuan Weng, Yitao Liu, Toh Jing Hua, Junning Zhao, Qian Liu, Che Liu, Leo Z. Liu, Yiheng Xu, Hongjin Su, Dongchan Shin, Caiming Xiong, and Tao Yu. Openagents: An open platform for language agents in the wild, 2023. URL `https://arxiv.org/abs/2310.10634`.

Geza Kovacs Isaac Galatzer-Levy Jacob Sunshine Jiening Zhan Ming-Zher Poh Shun Liao Paolo Di Achille Shwetak Patel Xin Liu, Daniel McDuff. Large language models are few-shot health learners. 2023.

Mingjin Zhang, Jiannong Cao, Xiaoming Shen, and Zeyang Cui. Edgeshard: Efficient llm inference via collaborative edge computing, 2024a. URL `https://arxiv.org/abs/2405.14371`.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model, 2024b. URL `https://arxiv.org/abs/2401.02385`.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2025. URL `https://arxiv.org/abs/2303.18223`.

Yue Zheng, Yuhao Chen, Bin Qian, Xiufang Shi, Yuanchao Shu, and Jiming Chen. A review on edge large language models: Design, execution, and applications, 2025. URL `https://arxiv.org/abs/2410.11845`.

Ningze Zhong, Yi Wang, Rui Xiong, Yingyue Zheng, Yang Li, Mingjun Ouyang, Dan Shen, and Xiangwei Zhu. Casit: Collective intelligent agent system for internet of things. *IEEE Internet of Things Journal*, 11(11):19646–19656, 2024. doi: 10.1109/JIOT.2024.3366906.

Xiaomao Zhou, Qingmin Jia, Yujiao Hu, Renchao Xie, Tao Huang, and F. Richard Yu. Geng: An llm-based generic time series data generation approach for edge intelligence via cross-domain collaboration. In *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 1–6, 2024. doi: 10.1109/INFOCOMWKSHPS61880.2024.10620716.

M. Zong, A. Hekmati, M. Guastalla, Y. Li, and B. Krishnamachari. Integrating large language models with internet of things applications, 2024. URL `https://arxiv.org/abs/2410.19223`.