Towards Automated, Resilient, and Robust AIoT

Leming Shen
The Hong Kong Polytechnic University, University College London
leming.shen@connect.polyu.hk,leming.shen@ucl.ac.uk

Abstract

Artificial Intelligence of Things (AIoT) enables pervasive, intelligent mobile systems by leveraging advanced AI techniques to process vast amounts of data generated from various IoT devices. This vision, however, is heavily bottlenecked by high development barriers, severe edge heterogeneity, and emerging cyber-physical security threats. To overcome these limitations, we design and implement a series of systems across three pillars: automated agentic AIoT application development, heterogeneity-aware, resilient on-device federated learning, and robust embodied AI systems.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; • **Computer systems organization** → **Embedded and cyber-physical systems**.

Keywords

LLM Agent, Code Generation, Federated Learning, Embodied AI

ACM Reference Format:

Leming Shen. 2026. Towards Automated, Resilient, and Robust AIoT. In *The 24th Annual International Conference on Mobile Systems, Applications and Services (MobiSys '26), June 21–25, 2026, Cambridge, United Kingdom.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3812835.3814830

1 Introduction

Artificial Intelligence of Things (AIoT) is an emerging paradigm that leverages advanced AI techniques to process a vast amount of data generated by diverse IoT and mobile devices. This technology brings a new level of mobile intelligence and automation to various applications, such as healthcare, smart sensing, and autonomous driving.

Though promising, in real-world scenarios and applications, however, we find it challenging to consistently deliver robust, performant AIoT systems, which manifests in three main aspects. ❶ *Developing various AIoT applications demands extensive human effort and dedicated domain knowledge.* Developers should master in-depth knowledge of AI model internals, heterogeneous mobile devices, as well as runtime environment customization and optimization methods. They also need to iteratively fine-tune the AIoT applications to deliver functioning and robust systems. ❷ *Mobile devices typically have heterogeneous data distributions and resource constraints.* One-size-fits-all solutions may impose excessive burden on resource-constrained mobile devices, while conversely leading to resource underutilization on more capable hardware. ❸

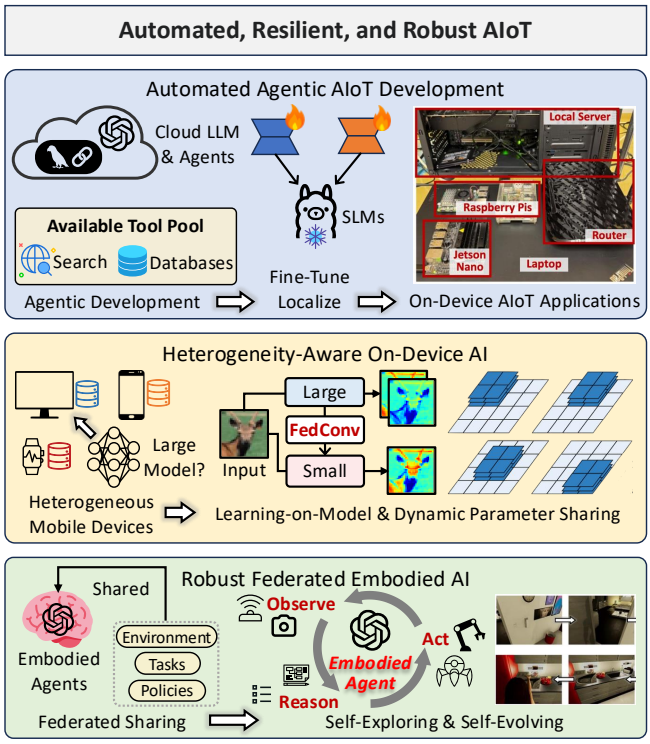

Figure 1: Automated, Resilient, and Robust AIoT

Mobile/embodied agents often struggle to generalize to diverse environments. Due to limited amounts of robotic data collected from the real world, existing embodied agents have constrained capabilities to adapt to new system configurations and environments.

To address these challenges, my research (Fig. 1) implements a series of mobile systems across three key pillars: 1) automated agentic AIoT application development via Large Language Models (LLMs); 2) heterogeneity-aware on-device AI via federated learning; and 3) robust federated embodied AI via multi-robot collaboration and sharing. Together, these efforts foster the vision of delivering automated, resilient, and secure AIoT applications and mobile systems.

2 Automated Agentic AIoT Development

Recent advances in large language models (LLMs) have fundamentally changed the way we interact with AI, exhibiting remarkable language processing and even code generation capabilities. By integrating LLMs with various external tools such as search engines and knowledge databases, we can build agentic systems to automate the entire AIoT application development process with minimal user intervention.

We first propose AutoIOT [5] that exploits agents to automatically synthesize executable programs to process IoT data, rather

MobiSys '26, Cambridge, United Kingdom

ACM ISBN 979-8-4007-2711-5/26/06
https://doi.org/10.1145/3812835.3814830

than feeding raw sensor data into LLMs for interpretation. By meticulously crafting prompts and chaining background knowledge retrieval, code generation, and iterative refinement, AutoIOT can synthesize performant programs for various IoT applications. Extensive evaluation results demonstrate that the code generated by AutoIOT can sometimes even outperform state-of-the-art algorithms.

Furthering AutoIOT, we reveal that existing programming agentic systems require labor-intensive design in agentic workflows and may even send sensitive user data to cloud LLMs. To deal with these issues, we develop GPIoT [3] and IoTCoder [7]. We first construct two large-scale code generation datasets tailored for the IoT domain. Then, we propose a *parameter-efficient co-tuning* (PECT) paradigm capable of collaboratively fine-tuning multiple local small language models (SLMs), facilitating knowledge transfer among distinct SLMs and significantly enhancing the performance of the generated IoT programs.

More importantly, we propose a self-exploring and self-evolving paradigm [4] that overcomes the inflexibility of fixed agentic workflows for AIoT development. We leverage the reasoning capabilities LLMs to dynamically construct adaptive workflows on the fly by autonomously select next appropriate operations based on the current context. To ensure robustness and adaptability in complex environments, we employ a hierarchical coordinator-actor architecture to efficiently manage tasks, a dynamic agentic recovery mechanism to mitigate cascaded errors, and a continuous reinforcement fine-tuning strategy for ongoing self-evolution.

3 Heterogeneity-Aware On-Device AI

In real-world scenarios, one-size-fits-all AI models suffer severe performance degradation on mobile devices due to data heterogeneity (intrinsic statistical differences across datasets) and resource heterogeneity (diverse computational constraints). My research addresses these challenges by dynamically generating AI models of varying sizes tailored to specific device resource constraints.

To overcome the heterogeneity issues, we propose FedConv [1, 2], which adopts a learning-on-model paradigm to distill diverse sub-models from a large global model. We find that by taking the parameters of a large global model as input, we can train a meta-model to learn and generate the parameters of sub-models with distinct sizes. These models are then assigned to appropriate mobile devices to ensure resource constraints are met. Evaluation results demonstrate that the generated sub-models can preserve the large model's feature-extraction capabilities and ultimately achieve comparable performance. Similarly, FedDM [6] adopts a dynamic model parameter sharing strategy, where sub-models dynamically share different parts of the parameters of the large model.

4 Robust Federated Embodied AI

Embodied AI (EAI) agents often struggle to adapt to diverse, dynamic real-world environments due to limited real-world data, heterogeneous environmental contexts, and rigid, predefined workflows. To address these challenges, my research envisions a robust federated embodied AI ecosystem where agents continuously learn, adapt, and evolve through collaboration and knowledge sharing.

To tackle environmental and task heterogeneity, we propose FEAI [8] that enables multiple federated EAI agents to share their experiences collaboratively. Rather than transmitting raw sensor data, FEAI shares and constructively aggregates environment semantic maps, task templates, and action-reward rules among federated agents. A central server generates aggregated representations, providing agents with enhanced collective situational awareness and generalizable manipulation policies for unfamiliar settings.

FSEAI [9] further overcomes rigid agentic workflows by exploiting LLM reasoning to dynamically construct workflows on the fly. We distill EAI tasks into three atomic operations, *i.e.*, observe, reason, and act, allowing agents to adaptively determine their next steps. Additionally, FSEAI utilizes continuous federated tuning via an ensemble mutual-actor-critic strategy to enable self-evolution. Ultimately, this collaborative approach achieves up to a 42.6% higher task success rate and significantly reduces token costs across heterogeneous EAI environments.

5 Closing Remarks & Looking Ahead

In summary, my research advances practical AIoT by synergizing automated application synthesis, heterogeneity-aware federated learning, and robust cyber-physical security. Looking ahead, my future work will focus on three directions: 1) automating on-device AI model deployment to close the loop in agentic AIoT development; 2) building self-evolving agents that can autonomously make next decisions on the fly; and 3) enhancing the generalizability of existing embodied agents across heterogeneous scenarios.

Leming Shen is currently a PhD student at The Hong Kong Polytechnic University, advised by Prof. Yuanqing Zheng. He is also a visiting PhD student at University College London, advised by Prof. Chris Xiaoxuan Lu. His research interests include LLMs, AI agents, embodied AI, and edge computing. He served as a TPC member in multiple conferences and journals, including artifact evaluations of MobiCom, MobiSys, SenSys, CCS, NDSS, S&P, and USENIX Security. He received the Distinguished Artifact Reviewer Award at MobiSys, SenSys, and CSS. Personal website: https://lemingshen.github.io/

Acknowledgments

We sincerely thank all anonymous reviewers for their valuable suggestions. This work is supported by Hong Kong GRF under Grant No. 15206123 and No. 15211924.

References

[1] Leming Shen, Qiang Yang, Kaiyan Cui, Yuanqing Zheng, Xiao-Yong Wei, Jianwei Liu, and Jinsong Han. 2024. Fedconv: A learning-on-model paradigm for heterogeneous federated clients. In *ACM MobiSys*. 398–411.

[2] Leming Shen, Qiang Yang, Kaiyan Cui, Yuanqing Zheng, Xiao-Yong Wei, Jianwei Liu, and Jinsong Han. 2025. Hierarchical and heterogeneous federated learning via a learning-on-model paradigm. *IEEE TMC* (2025).

[3] Leming Shen, Qiang Yang, Xinyu Huang, Zijing Ma, and Yuanqing Zheng. 2025. Gpiot: Tailoring small language models for iot program synthesis and development. In *ACM SenSys*. 199–212.

[4] Leming Shen, Qiang Yang, Xinyu Huang, Zijing Ma, Yuanqing Zheng, and Chris Xiaoxuan Lu. 2026. One Workflow Doesn't Fit All: Adaptive Workflows for Edge AI Development. *GetMobile* 30, 1 (2026), 11–15.

[5] Leming Shen, Qiang Yang, Yuanqing Zheng, and Mo Li. 2025. Autoiot: Llm-driven automated natural language programming for aiot applications. In *ACM MobiCom*. 468–482.

[6] Leming Shen and Yuanqing Zheng. 2023. FedDM: data and model heterogeneity-aware federated learning via dynamic weight sharing. In *IEEE ICDCS*.

[7] Leming Shen and Yuanqing Zheng. 2024. Iotcoder: A copilot for iot application development. In *ACM MobiCom*. 1647–1649.

[8] Leming Shen and Yuanqing Zheng. 2025. Poster: Towards federated embodied AI with FEAI. In *ACM MobiSys*. 599–600.

[9] Leming Shen and Yuanqing Zheng. 2026. Federated Self-Evolving Embodied AI Agents. In *IEEE INFOCOM EIN*.