LLMind 2.0

Yuyang Du, *Student Member, IEEE*, Soung Chang Liew, *Fellow, IEEE*,

***Abstract*—Recent advances in large language models (LLMs) have sparked interest in their application to IoT and automation systems, particularly for facilitating device management through natural language instructions. However, existing centralized approaches face significant scalability challenges when managing and coordinating the collaboration between IoT devise of diverse capabilities in large-scale heterogeneous IoT systems. This paper introduces LLMind 2.0, a distributed IoT automation framework that addresses the scalability challenges through lightweight LLM-empowered device agents via natural language-based machine-to-machine (M2M) communication. Unlike previous LLM-controlled automation systems that rely on a centralized coordinator to generate device-specific code to be executed on individual devices, LLMind 2.0 distributes intelligence across individual devices through lightweight LLMs embedded in IoT devices. The central coordinator translates human instructions into simple subtasks described in natural human language, which are then processed by device-specific agents to generate device-specific code locally at the associated devices. This approach transcends device heterogeneity barriers by using natural language as a unified communication medium, enabling seamless collaboration between devices from different manufacturers. The system incorporates several key innovations: a Retrieval-Augmented Generation (RAG) mechanism for accurate subtask-to-API mapping, fine-tuned lightweight LLMs for reliable code generation, and a finite state machine-based task execution framework. Experimental validation in multi-robot warehouse scenarios and real-world WiFi network deployments demonstrate significant improvements in scalability, reliability, and privacy protection compared to the centralized approach. The distributed architecture enables parallel processing, reduces coordinator computational burden, and facilitates natural collaborative behaviors among IoT devices.**

***Index Terms*—Random access, packet detection, false alarm, missed detection, machine-type communication**

I. Introduction

Recent breakthroughs in generative AIs have attracted significant interest in applying [1], [2] large language models (LLMs) to facilitate efficient system automation and device management [3]–[5] in complex IoT systems consisting of collaborative IoT devices of diverse capabilities. In this line research, LLM serves as the central system coordinator, translating human command into machine executable codes and assigning the codes to IoT devices to execute complex tasks involving the collaborative efforts of multiple IoT devices. Although this technical solution has proven effective for small-scale IoT systems, it has inherent scalability limitations for larger-scale systems with device heterogeneity.

This paper explores the latest technical trend in lightweight embedded LLMs and attempts to address the scalability problem by introducing distributed device agents into IoT management and automation systems empowered by LLMs. The integration of device agents opens new possibilities for the communication methods between the system coordinator and distributed devices: specifically, a new type of machine-to-machine (M2M) communication that leverages natural language as the unified interaction medium between the coordinator and IoT devices. As a replacement for the rigid, code-based interaction pattern – where the coordinator generates device-specific code for the diverse devices to execute – the natural language-based M2M interface transcends the barriers of device heterogeneity, allowing devices from different manufacturers (or even supported by different programming languages) to communicate and collaborate using a unified human language. Code generation is entrusted to device agents that know the languages of the associated devices intimately. Furthermore, as the natural language-based M2M interface shifts language-code transformation tasks from the central coordinator to distributed device agents, the coordinator's workload could be significantly reduced so that it can easily manage a larger-scale system. Additionally, the reliability of code generation is also enhanced with LLM-supported device agents that are specifically fine-tuned for the code generation task. Meanwhile, compared with previous solutions where a device's information is uploaded to a cloud-based LLM coordinator, the device's privacy and safety risk can be eliminated with the new natural language-based M2M interface, given that all device agents are locally deployed.

In short, this paper explores M2M communication using human natural language rather than machine languages. An added advantage is transparency, in which humans can easily understand machine interactions in automated systems when needed through logs written in natural language.

The rest of the introduction elaborates the above arguments with an overview. Subsection A articulates the limitations of conventional systems, and subsection B introduces our solution. Novel challenges arising from the implementation of the proposed system, as well as how they are addressed in our system, are highlighted in subsection C. This is followed by a summary of our contributions in subsection D.

A. Limitations of existing systems

Previously reported approaches relied heavily on an all-knowing LLM coordinator to translate natural language in-

Manuscript received 6 March 2024; revised 22 September 2024, and 2 January 2025; accepted 20 Feberuary 2025. Date of publication xx xxx 2025; date of current version xx xxx 2025. The review of this article was coordinated by Dr. Mugen Peng.

This work was supported in part by the General Research Funds (Project No. 14200221) established under the University Grant Committee of the Hong Kong SAR, China, and in part by CUHK Direct Grant (Project Code: 405524).

The research work described in this paper was conducted in the JC STEM Lab of Advanced Wireless Networks for Mission-Critical Automation and Intelligence funded by The Hong Kong Jockey Club Charities Trust.

Y. Du and S. C. Liew are with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China (e-mail: {dy020, soung}@ie.cuhk.edu.hk). S. C. Liew is the corresponding author.

structions, such as those provided by human users, into device-executable code. In [3], authored by us, proposed a task-oriented human-to-machine communication framework, where an LLM acts as a centralized coordinator to orchestrate IoT devices in executing complex tasks. This framework, known as LLMind, uses a language-code transformation approach, with the LLM translating verbal instructions into a finite-state machine (FSM) representation as an intermedia, which is then transformed into executable codes for diverse devices, possibly from different manufacturers. The components in the frameworks of other studies could also potentially serve as the centralized code generator. For example, [6] developed an instruction-tuning dataset to train an open-source LLM to execute human instructions through application APIs, while [7] created a dataset of API documentation to fine-tune the LLM, improving its understanding of API structures and syntax. Additionally, [8] fine-tuned a LLaMA model to generate accurate API function calls and adapted to test-time document changes by integrating a document retriever.

While effective for small-scale systems, the reliance on a centralized coordinator and code-based M2M interface introduces significant scalability challenges as the diversity and number of devices increases:

1) API Adaptation Complexity: Proprietary APIs from diverse devices are often written in different programming languages and incorporate device-specific specifications. This heterogeneity makes it increasingly complex for the LLM coordinator to generate accurate executable code as device diversity grows. Additionally, integrating new devices requires ensuring compatibility with existing ones, further exacerbating system complexity.
2) Coordinator Computation Bottleneck: As the number of devices grows, the centralized coordinator must generate device-specific executable code for many devices simultaneously. This creates significant computational strain, leading to inefficiencies and potential delays.

B. Our Solution: LLMind 2.0

We propose to leverage lightweight LLM-empowered IoT device agents, which use natural language as M2M communication interfaces, to address the scalability problem. In this approach, a conventional large-scale LLM acts as the central coordinator that translates natural language instructions - such as those provided by a human - into a series of subtasks, also specified in natural language. The coordinator then transmits these natural language subtask specifications, rather than executable code, to device-specific agents. These agents, realized with embedded lightweight LLMs in IoT devices (also known as embedded LLMs), are responsible for interpreting the natural language-based subtask instructions and generating executable code tailored to their respective devices' proprietary APIs.

This distributed approach fundamentally changes the computation and communication paradigm in existing systems: rather than relying on the LLM coordinator to handle device-specific code generation for all devices, the devices themselves, through their device agents, become active participants in the translation process. Using natural language as the medium of communication between the central coordinator and device agents gives rise to several key advantages:

1) Enhanced Scalability: The computational burden is offloaded from the centralized LLM coordinator, as device-specific agents handle the translation from natural language to executable code. This reduces the computation strain on the central coordinator as the number of devices grows.
2) Improved Flexibility: Device-specific agents are designed to interpret and process subtasks specific to their associated devices only, making it easier to integrate new devices with proprietary APIs into the system without requiring extensive updates to the LLM coordinator.
3) Facilitation of Human-Readable M2M Interactions: By using natural language as the communication medium, this approach paves the way for devices to interpret and even share instructions in human-readable form, enabling more transparent M2M communication.
4) Cost-Efficient Device Agents: The device agents do not need to rely on costly, general-purpose large language models. Instead, smaller, task-specific language models can be employed as agents since they only need expertise specific to their respective devices, rather than maintaining general knowledge. This reduces computational requirements and operational costs for individual devices.
5) Privacy: In previous systems [3, 6-8], code generations require device-specific information such as API interface or detailed descriptions of the environment that the device works in. By shifting the duty from cloud-based large-scale LLMs (such as GPT-4 used in LLMind [3]) to locally deployed lightweight device agents, the system's safety and privacy concerns can be well eliminated.
6) Parallel Code generation: the distributed architecture enables the parallel translation of multiple natural language subtasks by different device agents into device-executable code, significantly enhancing system efficiency.

Through this distributed agent-based approach, we aim to overcome the scalability limitations of centralized IOT automation systems while enabling efficient, flexible, and privacy-protecting IoT device management with the natural language-based M2M interactions between devices and the central coordinator. Building upon the original LLMind framework in [3], which is henceforth referred to as LLMind 1.0 for distinction, this paper puts forth LLMind 2.0, a fundamentally redesigned successor with the integration of all the above features. Fig. 1 illustrates the LLMind 2.0 system. At its core is a central coordinator containing an LLM that interprets human instructions and translates them into natural language subtasks. These subtasks are then executed by various devices through their respective agents.

II. Single-antenna Packet Detection: Analysis, Simulation, and Discussion

III. Benchmarking RP-S&C with Conventional S&C

IV. Multi-antenna Packet Detection: Analysis and Optimizations

V. Multi-antenna Packet Detection: Discussion and Experiments in a Distributed Antenna System

VI. Conclusion

Appendix A
How CFO influences the autocorrelation term

Appendix B
Finding the optimal packet-detection threshold

References

[1] V. Ashish, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, p. I, 2017.

[2] Z. Deng, W. Ma, Q.-L. Han, W. Zhou, X. Zhu, S. Wen, and Y. Xiang, "Exploring deepseek: A survey on advances, applications, challenges and future directions," *IEEE/CAA Journal of Automatica Sinica*, vol. 12, no. 5, pp. 872–893, 2025.

[3] H. Cui, Y. Du, Q. Yang, Y. Shao, and S. C. Liew, "Llmind: Orchestrating ai and iot with llm for complex task execution," *IEEE Communications Magazine*, 2024.

[4] B. Xiao, B. Kantarci, J. Kang, D. Niyato, and M. Guizani, "Efficient prompting for llm-based generative internet of things," *IEEE Internet of Things Journal*, 2024.

[5] İ. Kök, O. Demirci, and S. Özdemir, "When iot meet llms: Applications and challenges," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024, pp. 7075–7084.

Yuyang Du (Student Member, IEEE) is a Ph.D. candidate in the Department of Information Engineering at The Chinese University of Hong Kong (CUHK), Hong Kong SAR, China. He received his B.S. degree in Electronics from Peking University, Beijing, China, in 2019. Currently, he is a visiting scholar at the Systems + Theory group at Harvard University. Prior to joining CUHK, he worked as a communication engineer at 2012 Lab, Huawei, where he participated in the power-efficient optimization of the Kirin 9000 5G SoC. His research interests include Ultra-Reliable Low-Latency Communications (URLLC), nonlinear amplifiers, and the application of generative AI in wireless networks.

Soung Chang Liew (Fellow, IEEE) is a Choh-Ming Li Professor of Information Engineering at The Chinese University of Hong Kong (CUHK). He received his S.B., S.M., E.E., and Ph.D. degrees from the Massachusetts Institute of Technology. From 1984 to 1988, he was at the MIT Laboratory for Information and Decision Systems, where he investigated Fiber-Optic Communications Networks. From March 1988 to July 1993, he was at Bellcore (now Telcordia), where he engaged in Broadband Network Research. Since 1993, he has been a Professor at the Department of Information Engineering, CUHK. Prof. Liew is currently a Co-Director of the Institute of Network Coding at CUHK. His research interests include wireless networks, Internet of Things, intelligent transport systems, machines learning, Internet protocols, multimedia communications, and packet switch design. Prof. Liew is the recipient of the first Vice Chancellor Exemplary Teaching Award in 2000 and the Research Excellence Award in 2013 at CUHK. Prof. Liew is a Fellow of IEEE, IET, Hong Kong Institution of Engineers, and Hong Kong Academy of Engineering Sciences.