Towards Automated Mobile Model Deployment

Leming Shen[1,3], Qiang Yang[2], Xinyu Huang[1], Zijing Ma[1],
Yuanqing Zheng[1], Chris Xiaoxuan Lu[3]
[1]The Hong Kong Polytechnic University, [2]University of Cambridge, [3] University College London

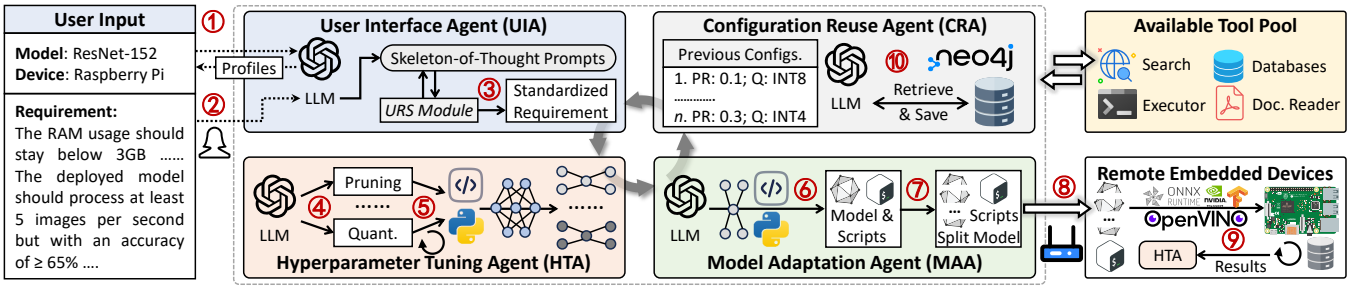

Figure 1: The system overview and workflow of AutoDeploy with four specialized agents.

Abstract

We present AutoDeploy that deploys AI models on heterogeneous mobile devices by fully automating model optimization and adaptation with Large Language Model (LLM) agents. AutoDeploy not only significantly reduces human effort via automation but also enhances inference efficacy by integrating LLMs with device-specific knowledge.

ACM Reference Format:
Leming Shen[1,3], Qiang Yang[2], Xinyu Huang[1], Zijing Ma[1],, Yuanqing Zheng[1], Chris Xiaoxuan Lu[3] . 2026. Towards Automated Mobile Model Deployment. In *The 8th UK Mobile, Wearable and Ubiquitous Systems Research Symposium (MobiUK '26), June 25–26, 2026, Cambridge, United Kingdom.* ACM, New York, NY, USA, 1 page.

1 Introduction

Mobile AI enables the deployment of AI models directly on mobile devices and facilitates onboard data processing of various sensors for low-latency decision-making [1, 2, 4].

To develop a mobile AI application, developers can download a model from open-source platforms and deploy it on target mobile devices. However, most models cannot be directly deployed for two main reasons. 1) These models often contain a huge amount of parameters, making them hard to fit into resource-constrained mobile devices. 2) These models are primarily available in PyTorch format, which may not be supported by mobile devices with heterogeneous execution environments. As a result, developers need to perform multiple iterations of model optimization and adaptation (O&A), including model compression, translation, and environment adaptation, which are labor-intensive and error-prone.

In this paper, we propose AutoDeploy, a multi-agent-assisted model O&A system that leverages agents' coding abilities [3, 5] for heterogeneous mobile AI applications.

2 System Overview

Fig. 1 shows the overall workflow of AutoDeploy, which coordinates four agents to automatically execute the O&A process and present the final deployment results. Specifically, *User Interface Agent* **(UIA)** first constructs detailed profiles of the model and the device (①), serving as references for the user to elicit her deployment requirements (②). UIA then uses the *user requirement standardization* module to refine the requirement into a standardized format (③). *Hyperparameter Tuning Agent* **(HTA)** then generates a list of optimizable hyperparameters (④) and synthesizes a script for automated hyperparameter tuning (⑤). After iterative optimization, multiple optimized models that meet resource constraints are generated. *Model Adaptation Agent* **(MAA)** translates them into portable models and generates a script to split the models into multiple parts via an *operator-aware model partition* module (⑥). MAA then synthesizes a script that can be executed on the target device to perform inference using a *cross-engine inference pipeline* (⑦). The partitioned model and the script are then transmitted to the target device (⑧), which monitors the real-time performance on its local data and sends the results to HTA as feedback to start a new optimization round (⑨). *Configuration Reuse Agent* **(CRA)** further optimizes HTA by reusing previous hyperparameters (⑩) during model optimization, thereby enhancing overall efficiency rather than from scratch.

3 Conclusion

AutoDeploy is a multi-agent automatic model O&A system for heterogeneous mobile AI applications. We believe AutoDeploy can unlock the potential to drive the practical deployment of open-source AI models across diverse and distributed mobile devices in real-world everyday applications.

References

[1] Leming Shen. 2026. Towards Automated, Resilient, and Robust AIoT. In *ACM MobiSys*.
[2] 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*. 1–14.
[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, Yuanqing Zheng, and Mo Li. 2025. Autoiot: Llm-driven automated natural language programming for aiot applications. In *ACM MobiCom*. 468–482.
[5] Leming Shen and Yuanqing Zheng. 2024. Iotcoder: A copilot for iot application development. In *ACM MobiCom*. 1647–1649.