UDC 004.8

INFORMATION TECHNOLOGY FOR THE SYNTHESIS OF INTELLIGENT MOBILE SYSTEMS BASED ON LANGUAGE MODELS

Myroniuk M.V.[1], *Yahup K.V.[2],* *Kopp A.M.[3]*

[1] *Postgraduate Student of the SE & MIT Department, NTU «KhPI», Kharkiv, Ukraine*
[2] *Professor of the SE & MIT Department, Ph.D., Professor, NTU «KhPI», Kharkiv, Ukraine*
[3] *Head of the SE & MIT Department, Ph.D., Associate Prof., NTU «KhPI», Kharkiv, Ukraine*
maksym.myroniuk@cs.khpi.edu.ua

The rapid development of Artificial Intelligence systems, particularly Large Language Models (LLMs), has transformed human–computer interaction by enabling context-aware text generation, automated communication, and intelligent data processing. However, as the use of AI expands, there is a growing demand for adapting LLMs to mobile environments – smartphones, tablets, and embedded systems, which have become the dominant platform for real-time user interaction with digital services.

However, in practice, integrating LLMs into mobile systems remains a major challenge due to their high computational complexity, large memory footprint, energy consumption, and dependence on cloud infrastructure. These constraints limit on-device autonomy and data privacy, revealing a scientific and practical gap: the absence of a unified technology for synthesizing efficient, intelligent systems based on language models that can operate effectively under the constraints of mobile environments [1].

Therefore, the need to adapt language models to operate efficiently within mobile environments has emerged as a critical and timely challenge, requiring a comprehensive approach to their evaluation, optimization, and integration. Developing methods that ensure high performance, energy efficiency, and autonomy will help bridge the gap between large-scale AI capabilities and the limitations of mobile hardware, enabling the broader use of intelligent on-device solutions across education, healthcare, business, and other everyday technologies.

In response to the growing demand for deploying language models on mobile platforms, the scientific community has begun to explore and systematize approaches for integrating and evaluating such models within intelligent mobile systems. Recent research focuses on both architectural simplification and the creation of compact, specialized models suitable for real-time, on-device applications. Given the high computational demands of large language models, current studies emphasize four key optimization strategies – quantization, pruning, knowledge distillation, and low-rank factorization – which collectively enable the development of lightweight yet efficient models capable of functioning effectively in resource-constrained mobile environments [2].

Each optimization method offers unique advantages: quantization minimizes numerical precision to enhance speed and memory efficiency; pruning removes redundant parameters to simplify model structure; knowledge distillation transfers essential knowledge from larger models to smaller ones; and low-rank factorization decomposes complex matrices into compact forms. Together, these methods form the foundation for developing lightweight and high-performance language models for mobile environments.

Another promising direction in AI optimization research involves the development of compact language models – including Small, Tiny, and Super Tiny Language Models – which significantly reduce computational and energy demands while maintaining competitive performance. These lightweight architectures aim to achieve real-time efficiency and

compatibility with mobile hardware, though their practical stability, generalization ability, and large-scale applicability remain open questions for further scientific investigation.

As noted above, most research focuses on reducing the size and complexity of language models, while an equally promising direction involves developing context-oriented models trained for specific, well-defined tasks. Unlike general-purpose LLMs that require massive and diverse datasets, context-focused models leverage smaller, domain-specific data to achieve higher efficiency, faster performance, and lower resource consumption – qualities essential for mobile environments where memory, processing power, and energy are limited. This approach enables targeted optimization and ensures stable, high-quality operation within narrow functional boundaries, making it a key step toward sustainable on-device intelligence.

Another promising direction in adapting language models for mobile systems involves leveraging external data gathered by device sensors itself – such as accelerometers, gyroscopes, light and proximity sensors, GPS modules, microphones, Bluetooth, Wi-Fi, and contextual indicators like battery level, screen activity, or time of day – to enable dynamic, context-aware model behavior and personalized interaction based on the user's environment and activity [3].

Incorporating sensor data not only allows language models to respond to user inputs but also to anticipate them by recognizing behavioral patterns and contextual cues. This integration paves the way for predictive and proactive interaction between the user and the system. However, this area still remains underexplored and requires further theoretical and experimental study to determine its impact on model adaptability and performance.

Finally, a promising idea (see Fig. 1) for advancing mobile AI lies in designing language models that dynamically adapt to the hardware characteristics of each device – its CPU, GPU, RAM, and neural accelerators. By incorporating adaptive optimization strategies that account for the diverse configurations of modern smartphones, particularly across the Android ecosystem, such models could achieve consistent efficiency, stability, and performance regardless of hardware variability.

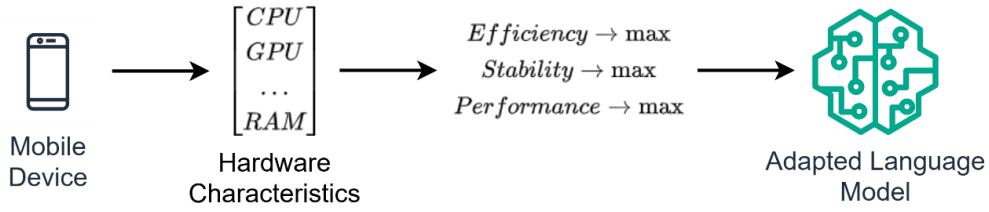

Figure 1 – Adapting language models to mobile device hardware characteristics

In summary, the integration of language models into mobile environments represents a crucial step toward making artificial intelligence more accessible, autonomous, and personalized. Developing adaptive, efficient, and context-aware models capable of operating directly on mobile devices will bridge the gap between large-scale AI technologies and everyday user needs, transforming smartphones into truly intelligent companions that enhance communication, learning, and decision-making in real time.

References:

1. *Wereszczyński K. Are Local LLMs on Mobile a Gimmick? The Reality in 2025. URL: https://www.callstack.com/blog/local-llms-on-mobile-are-a-gimmick (accessed: 30.10.2025).*

***2.** Jiajun Xu, Zhiyuan Li, Wei Chen, Qun Wang, Xin Gao, Qi Cai, Ziyuan Ling. (2024). On-Device Language Models: A Comprehensive Review. P. 38. DOI: 10.48550/arXiv.2409.00088.*

***3.** Nan Gao, Zhuolei Yu, Yue Xu, Chun Yu, Yuntao Wang, Flora D. Salim, Yuanchun Shi. (2023). Leveraging Large Language Models for Generating Mobile Sensing Strategies in Human Behavior Modeling. P. 8. DOI: 10.48550/arXiv.2311.05457.*