

Are We There Yet? A Measurement Study of Efficiency for LLM Applications on Mobile Devices

Xiao Yan

The University of Texas at Dallas
Richardson, USA
xiao.yan@utdallas.edu

Abstract

Recent advancements in large language models (LLMs) have prompted interest in deploying these models on mobile devices to enable new applications without relying on cloud connectivity. However, the efficiency constraints of deploying LLMs on resource-limited devices present significant challenges. In this paper, we conduct a comprehensive measurement study to evaluate the efficiency tradeoffs between mobile-based, edge-based, and cloud-based deployments for LLM applications. We implement AutoLife-Lite, a simplified LLM-based application that analyzes smartphone sensor data to infer user location and activity contexts. Our experiments reveal that: (1) Only small-size LLMs (<4B parameters) can run successfully on powerful mobile devices, though they exhibit quality limitations compared to larger models; (2) Model compression is effective in lower the hardware requirement, but may lead to significant performance degradation; (3) The latency to run LLMs on mobile devices with meaningful output is significant (>30 seconds), while cloud services demonstrate better time efficiency (<10 seconds); (4) Edge deployments offer intermediate tradeoffs between latency and model capabilities, with different results on CPU-based and GPU-based settings. These findings provide valuable insights for system designers on the current limitations and future directions for on-device LLM applications.

Keywords

On-Device Learning, Foundation Model, Real-Time

1 Introduction

In recent years, large language models (LLMs) and vision language models (VLMs) have been greatly advanced and applied in various domains. Given their impressive reasoning and generating capacities, some recent work has explored the LLMs and VLMs to perceive the environment through smartphone sensors and conduct inferences, such as SHARE [44], PenetrativeAI [37], and AutoLife [36]. However, in most existing work, LLMs and VLMs are running remotely on the cloud server, which incurs some potential limitations: dependence on stable network access, uncontrolled latencies due to network and server status, and privacy risks in uploading data. Therefore, deploying the LLMs and VLMs on

Yi Ding

The University of Texas at Dallas
Richardson, USA
yi.ding@utdallas.edu

local mobile devices can enable new applications in broader scenarios by overcome these limitations.

In the mobile computing community, research work have been done from different perspectives to enable training and prediction on mobile devices (latency [11, 14, 39, 43], memory [8, 16, 38], and energy [13, 21, 30]). The solutions can be categorized as model compression and selection [8, 16, 30, 38], federated learning [5, 31], and heterogeneous computing [13]. However, the existing work is mostly focused on the traditional deep-learning framework, which cannot address all the challenges in deploying LLMs and VLMs. For the topic of LLMs on mobile and edge devices, survey and position papers published recently laid the foundation of this emerging direction [4, 45]. Specific works focus on the different perspectives, including privacy and security [19, 42], model customization and personalization [28, 46], and model/system optimization [7, 35, 40]. However, most work does not provide a horizontal comparison of the system efficiency across different mobile, edge, and cloud platforms.

In this paper, by conducting an experimental study on the efficiency of different LLM deployments (i.e., mobile, edge, and cloud) for a mobile application, we aim to provide a holistic comparison and discussion to enhance the community's understanding of potential tradeoffs of different deployment settings. Specifically, we implemented a simplified version of AutoLife [36], a recent work that uses multi-modality sensor data on smartphones to infer human locations and activities and generate a diary for the users. We deploy the system with three different settings: mobile-based, edge-based, and cloud-based, and compare the memory consumption and the latency of the system.

The major observations and conclusions we have include: (1) We successfully deployed four small-size LLMs (Gemma2B, Gemma2-2B, Llama3.2-1B, and Llama3.2-3B) on an Android phone with GPU and 8GB RAM, but only one model (Gemma2-2B) provides meaningful answers. Meanwhile, all LLMs with 7B+ parameters can provide meaningful answers, but these models cannot be deployed on mobile devices due to limited memory. This indicates the design space is limited for systems that require running LLMs locally. (2) We identified the drawbacks of model compression. The compressed versions of Llama3.2 deployed on mobile devices

fail to generate meaningful answers, while the original versions work on edge servers. (3) The latency to run LLMs on mobile devices with meaningful output is significant (i.e., >30 seconds), while using a cloud service with API is more time-efficient (i.e., <10 seconds). For all deployments, the latency positively correlates with the model size regardless of model series (e.g., DeepSeek or Llama), but different model series with the same model sizes have different latencies. (4) A typical one-GPU-based edge server shows much higher efficiency (i.e., higher model output speed, lower latency with smaller variance) than a typical 8-core-CPU-based edge server, indicating the superiority of GPU-based servers for LLM-based applications. (5) The model's speed (i.e., number of tokens output per second) negatively correlates with the model size, indicating the difficulties of adopting large-size LLMs for real-time applications.

The contribution of the work is three-fold: (1) We deployed an LLM-based mobile application in three different settings (i.e., mobile, edge (CPU-based and GPU-based), and cloud) and measured the latency and memory consumption to provide an aligned comparison. The code used in the paper will be published so that the researchers can use it to conduct the following work. (2) We conducted thorough experiments with different model versions (Gemma, Llama, DeepSeek, Qwen, GPT, Claud) and different model sizes (e.g., 0.5B, 1B, 2B, 3B, 4B, 7B, 8B) to evaluate the system efficiency. (3) We obtained some interesting observations and conclusions from the experiments, which can help enhance the community's understanding of the practicability and potential challenges of mobile LLM applications.

2 LLM Application

The adoption of LLMs has been an emerging topic in the CPS-IoT community [3]. Applications explored include spatiotemporal data mining [15, 25], mobile tasking [6, 17, 20, 32, 41], and mobile sensing and reasoning [2, 6, 36, 37]. In this work, we choose to implement a key component in AutoLife [36], a novel application to use smartphone sensors (i.e., GPS, Wi-Fi, IMU, Barometer) to infer user location/context and compile daily diaries. The motivation to implement AutoLife are two folds: (1) The application's task is complex enough to unveil the potential of mobile LLM and simple enough to be solved with small LLMs (e.g., Gemma2-2B). Actually, our results indicate that the task in AutoLife is at the boundary of solvable (at least 2B model needed) and executable (at most 3B can be deployed on smartphones with 8 GB memory) using mobile LLM. (2) Multiple types of sensor data are collected and processed to illustrate LLMs' capacity in sensor data understanding and reasoning, which can further motivate future work in LLM for perception.

We implement AutoLife-Lite (Figure 1), a lite version of the location and motion context fusion part in AutoLife [36]

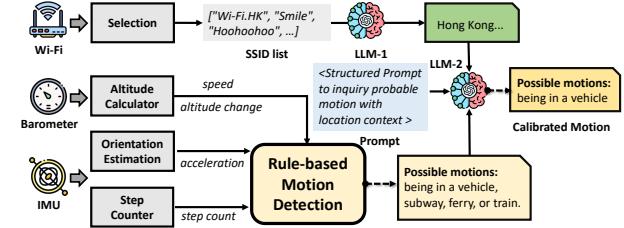


Figure 1: AutoLife-Lite

without the VLM module because there is no available VLM to deploy on mobile devices. The subtask implemented in AutoLife-Lite is to use the Wi-Fi, barometer, and IMU data to infer the location and motion of the user, which is a key module in AutoLife. Specifically, the Rule-based Motion Detection collects sensor data by registering listeners with Android's sensor system. When a sensor reports new data, the detector processes it through callback functions that update internal variables. This collected data is combined in a detection algorithm that uses predefined thresholds to classify the user's current motion state, such as walking, running, or riding in a vehicle.

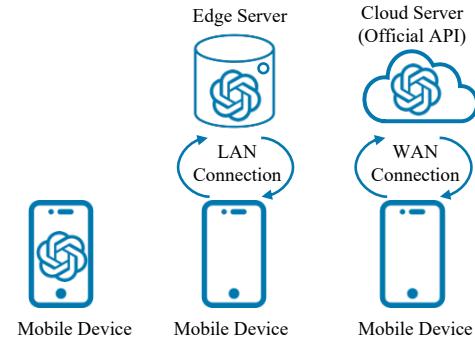


Figure 2: Mobile-, Edge-, and Cloud-based Deployment

3 Application Deployment

We envisioned three typical deployment settings of using LLMs in mobile applications as illustrated in Figure 2: **Mobile-based**, **Edge-based**, and **Cloud-based**. In **Mobile-based** deployment, the LLM models (usually a small-sized or compressed model like 2B) are completely on mobile devices to ensure data privacy and usability without a network connection. In the **Edge-based** deployment, LLM models (a medium-sized model like 8B) are deployed on the edge server, where mobile devices can connect to the server with a local area network (LAN) (e.g., Wi-Fi, cellular). A potential scenario is a group of drones as mobile devices connect to the edge server on the cellular station using a cellular network. In the **Cloud-based** deployment, the LLM models are deployed by commercial companies on cloud servers, and APIs are provided to access the service through wide area network (WAN) (e.g., Internet). Most of the existing mobile APPs are

deployed in this fashion because (1) availability of latest models; (2) minimum deployment effort; (3) minimum memory consumption on mobile devices needed.

In this work, we deployed AutoLife-Lite on all three settings to provide a holistic observation of the performance and efficiency (i.e., latency, memory) of different deployments.

3.1 Mobile-based Deployment

Hardware: We use a Google Pixel 8 [9] as the mobile device, which is equipped with 8GB RAM, a Google Tensor G3 processor [33] (with a 10-Core GPU Immortalis-G715s MC1 and Nona-core CPU), and a customized TPU.

Software and Framework: We leverage Google MediaPipe [10] (recently rebranded as LiteRT) to deploy Gemma-2B and Gemma2-2B and PyTorch Executorch [26] to deploy the Llama3.2-1B and Llama3.2-3B on the smartphone.

3.2 Edge-based Deployment

Hardware: We use the Xen-virtualized server [1] the AWS platform provides as the edge server. We use different settings to simulate a CPU-based and a GPU-based server. For the CPU-based server, we used 800% CPU resources with 500 GB memory, which represents eight full cores of processing power on an Amazon P3 instance with Intel Xeon E5-2686 v4 processors. For the GPU-based server, we used one Tesla V100-SXM2-16GB GPU with 16GiB memory.

Software and Framework: We use Ollama [24], a light-weight and extensible framework for building and running language models on the local machine (e.g., macOS, Linux, and Windows). The LLMs hosted include Llama (3.2-3B, 3.2-1B, 3.1-8B), DeepSeek-R1 (1.5B, 8B), Qwen (0.5B, 1.8B, 4B, 7B), Gemma (2-2B, 2B, 7B).

3.3 Cloud-based Deployment

We use Wi-Fi to access the Internet, and HTTP requests are used to call the APIs provided by the companies. The LLMs used include Claude (claude-3-sonnet-20240229-70B, claude-3-haiku-20240307-20B), OpenAI GPT (gpt-4-turbo-preview-1700B, gpt-4o-mini-8B), and Google Gemini (gemini-1.5-pro-1500B, gemini-2.0-flash-40B).

4 Evaluation

We evaluate the efficiency performance (i.e., memory consumption, latency) of the three deployments (i.e., mobile-based, edge-based, and cloud-based) of AutoLife-Lite. We did not measure the accuracy of model output as it's outside the scope of this paper, but we have conducted manual validation of the output to identify cases in which the model failed to provide the output or provided hallucinated output.

4.1 Measurement Methodology

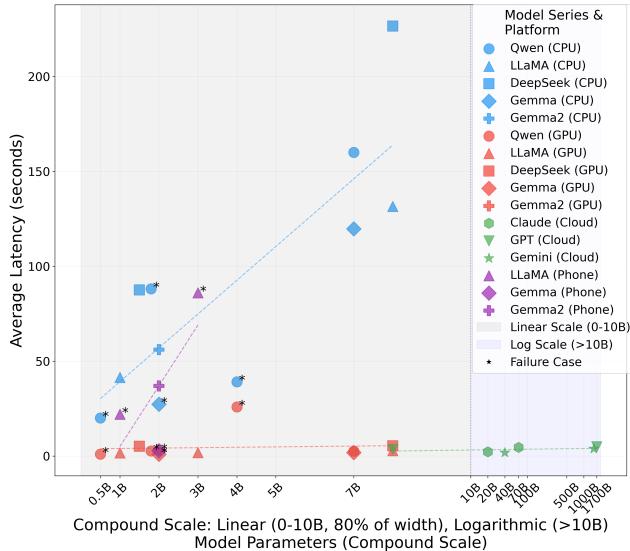
Memory Consumption Measurement. We measured the smartphone memory consumption in all three deployments

to provide an aligned comparison. We use resident set size (RSS) [34] to directly access the process status file (`/proc/<pid>/status`) and measure the total physical memory held in RAM. This metric represents the actual memory footprint of our application in the Android system. All the results are based on the setting that AutoLife-Lite is the only APP running and runs in the foreground. We also measured the memory consumption of the edge server in the edge-based deployment. For memory on the edge server, we employ a dual-phase approach to measure memory utilization during LLM inference. For CPU memory (RAM), we capture baseline measurements before inference begins and peak measurements upon completion, with the difference representing the model's RAM footprint. Simultaneously, we track GPU memory using NVIDIA's SMI tools, measuring allocated and actively used memory for each GPU. A GPU is considered "active" when its utilization or memory usage exceeds a certain threshold (5% in our work based on empirical observation). Throughout the inference process, a dedicated monitoring thread samples all resources at 500ms intervals, providing visibility into memory allocation patterns, load distribution across multiple GPUs, and transient resource spikes. This comprehensive approach enables detailed analysis of memory requirements across different model architectures and sizes.

Latency Measurement. We measure the latency to evaluate if LLM-based applications like AutoLife-Lite can be achieved in real-time and what the impact of different deployment settings and model sizes on the latency. Specifically, in the mobile-based deployment, the latency is measured as the duration between the time that sensor data is collected and the time the local model on the smartphone generates the inference results. We use `System.currentTimeMillis()` to capture the timestamps for consistent measurement across all operations. In the edge-based deployment, the latency is measured as the duration between the time that the sensor data is received on the edge server and the time the local model on the edge server generates the inference results. The time is obtained from the log of Ollama. Note that the data transmission time is not included since (1) the transmission time is negligible compared to the model processing time; (2) the transmission time may vary significantly in different settings. In the cloud-based deployment, the latency is measured as the duration between the time that sensor data is collected and the time the inference results are transmitted back to the device. Note that the data transmission time is included in the cloud since it's difficult to isolate it. Moreover, other unknown processing times (e.g., query queuing time on the cloud server) are also included.

Structured Prompt. To achieve a fair comparison between different models in different deployments, we use a structured prompt to enforce unified output (Figure 3).

```
Return JSON response matching this structure:
{
  "confidence": 0.0-1.0,
  "motion": {
    "type": ["stationary", "limited motion", "jogging/running",
    "walking", "cycling", "vehicle/subway/ferry/train", "escalator/elevator"],
    "direction": string?,
    "speed": string?,
    "attributes": string[]
  },
  "location": {
    "place": string,
    "type": ["indoor", "outdoor", "transit", "unknown"],
    "coordinates": {"latitude": number?, "longitude": number?},
    "attributes": string[]
  },
  "contextDescription": "max 50 words"
}
```

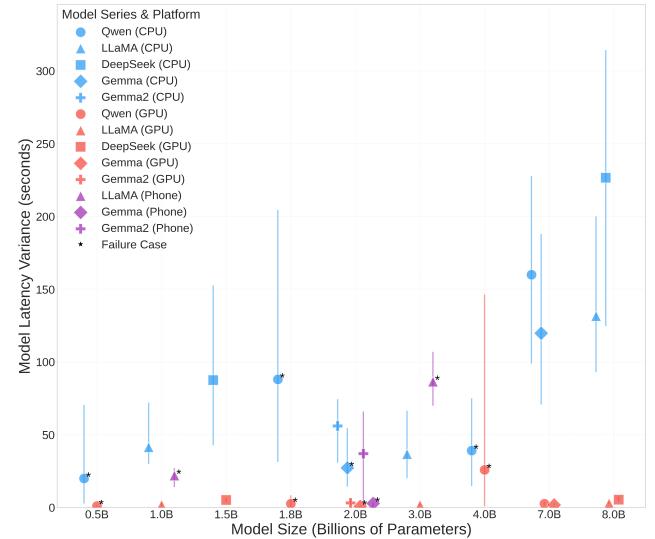
Figure 3: Structured Prompt to Inquiry Motion**Figure 4: Model Size v.s. Latency**

4.2 Evaluation Results

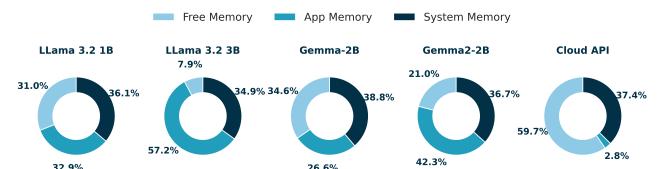
Overall Results. To provide the overall result panoramically, we use Figure 4 to illustrate the relation of model size, average latency, and deployment. We have the following observations: (1) Cloud-based deployments have the minimum latency (<10s), while mobile-based deployments have the maximum latency (>30s for meaningful output), indicating that cloud-based deployment is still the best way for near-real-time applications. (2) GPU-server has smaller and more consistent latency compared to CPU-server, indicating the importance of GPU in LLM applications. (3) Only smaller models (<4B) can be deployed on mobile devices, but most (3 out of 4) failed to generate meaningful answers. More detailed results and analysis are as follows.

Model Output Quality. Although we didn't conduct thorough experiments to validate the accuracy of the results, we manually evaluated the results and observed consistent

failures for some model-platform combinations. Specifically, two types of failures are observed: (1) Nothing generated: Gemma-2B (Mobile, CPU, GPU). (2) Failing to generate a reasonable result (i.e., unfinished description or copy-paste of the input): Qwen-0.5B (Mobile, CPU, GPU), Qwen-1.8B (Mobile, CPU, GPU), Qwen 4B (CPU, GPU), Llama3.2-1B(Mobile), Llama3.2-3B(Mobile). Note that Llama3.2-1B and Llama3.2-3B do not fail on CPU- and GPU-based deployment because although they are the same model with the same number of parameters, Llama3.2 on mobile is an optimized version with techniques like LoRA and quantization [22]. This implies the drawbacks and risks of model compression: for the same model with the same number of parameters, the non-compressed version works on the edge server, but the compressed version fails on mobile.

**Figure 5: Model Size v.s. Latency Variance**

Latency. As shown in Figure 5, the CPU-based edge server has a much higher average latency and larger variance compared with the GPU-based edge server. A potential reason is that the GPU's specialized memory design makes it efficient for both LLM training and inference. This result implies the necessity of GPU-based settings for applications with latency requirements.

**Figure 6: Memory Consumption on Mobile Device**

Memory Consumption on Mobile Devices. Figure 6 shows a comparison of the memory consumption on mobile devices with mobile-based (left four) and cloud-based (last one)

deployment. Given that the OS consumes around 35%-40% memory, a Llama-3B model would consume 57% memory, with only around 8% free memory. We use a Google Pixel with 8GB memory in the experiment, and some recent work has successfully deployed 7B models on smartphones with 16GB memory [27]. As we have verified that models with 7B+ parameters can always produce reasonable output, smartphone memory becomes the bottleneck for local LLM deployment to wider applications.

Table 1: Memory Usage in Edge-based Deployment

| Model | CPU-based (GB) | GPU-based (GB) |
|-----------------|-------------------|-------------------|
| Qwen-0.5B | 3.04 | 1.88 |
| Llama3.2-1B | 3.80 | 2.59 |
| DeepSeekR1-1.5B | 2.50 | 2.00 |
| Qwen-1.8B | 4.54 | 3.24 |
| Gemma-2B | 3.38 | 2.82 |
| Gemma2-2B | 3.88 | 3.41 |
| Llama3.2-3B | 4.90 | 3.76 |
| Qwen-4B | 7.08 | 6.02 |
| Qwen-7B | 10.28 | 9.06 |
| Gemma-7B | 8.40 | 9.37 |
| Llama3.1-8B | 6.71 | 6.52 |
| DeepSeek-R1-8B | 7.01 | 6.52 |

Memory Consumption on Edge Servers. Memory consumption is similar in CPU- and GPU-based deployment (Table 1). GPU-based deployment has less memory usage, as we only measure the usage on GPU memory. Some operations still use CPU even in the GPU-based deployment, and it's hard to count that part.

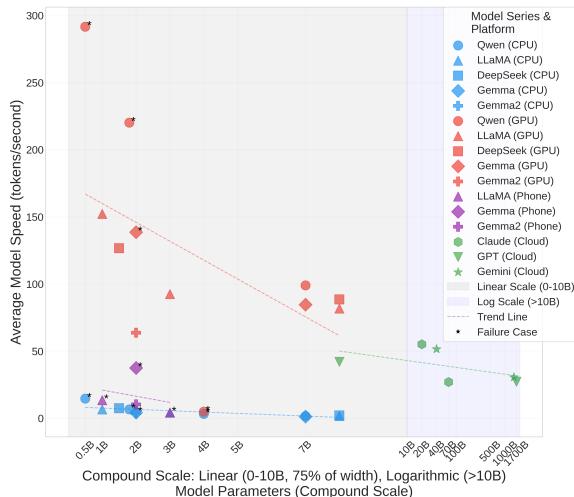


Figure 7: Model Size and Speed on Edge Server

Model Speed on Edge Server. We measure the speed of LLMs running on the edge server and depict its relation with the model size in Figure 7. Here, the speed is measured as

the ratio of the number of token outputs and time. The result verifies a straightforward intuition that models with a large number of parameters tend to be slower. This indicates the challenges of deploying or using large-size LLMs (e.g., >100B) to pursue output quality in a time-efficient way.

5 Related Work

LLM in CPS-IoT. The adoption of LLMs has been an emerging topic in the CPS-IoT community [3]. Applications explored include spatiotemporal data mining [15, 25], mobile tasking [6, 17, 18, 20, 32, 41], and mobile sensing and reasoning [2, 6, 36, 37]. The technical challenges solved can be categorized as: benchmark construction [2, 12, 29, 41], data representation [2, 32, 36] models [2, 41], prompting [25], memory [18, 32].

LLM on Mobile/Edge Devices. Several survey and position papers have been published recently to lay the foundation of this emerging direction [4, 45]. Some work focuses on the benefit of privacy/security provided by the on-device LLM [19, 42]. Some work focuses on the model customization and personalization [28, 46]. Some work focuses on improving the performance through model or system optimization [7, 35, 40]. However, most work does not provide a horizontal comparison of the system efficiency across different mobile, edge, and cloud platforms. This paper provides a holistic comparison and discussion to help the community understand the potential challenges and tradeoffs when choosing deployment settings for a specific application.

6 Discussion

Limitations and Future Work. (1) We only measure the system performance on one mobile device and one edge server, which may not be typical in other scenarios (e.g., sensor networks), but the work can still shed light on similar systems with powerful mobile devices and edge servers. (2) We only measured the efficiency (memory and latency) but did not measure the accuracy. However, we conducted a sanity check of the output and identified some failure cases. (3) Due to the frameworks and models available, we only test the performance of LLMs but not VLMs. On-device VLM is an interesting problem since many mobile devices are equipped with cameras, and videos embed rich information to help understand the environment and human behavior. (4) As new technologies developed later may cause our results to appear less accurate, but we still believe the thorough experimental results reported here can benefit the community in identifying the gaps.

Benchmark and Standards. In this work, we use AutoLife-Lite as a benchmark task to test the LLM efficiency across different deployments. We are aware of other CPS-IoT benchmarks proposed in the community and plan to study them in the following work. [2, 12, 29, 41]. At the same time, the

heterogeneity in hardware and operating systems brings significant difficulties in comparing different designs. Platforms like NVIDIA Jetson [23] have the potential to become a standard edge computing platform for AI deployment in resource-constrained environments.

7 Conclusion

The experimental study reveals significant challenges in deploying LLMs on mobile devices, including the deployability, output quality, latency, and memory consumption. These findings suggest that while on-device LLM applications show promise, significant work are needed from different perspectives (e.g., hardware, model, system, application) before they become practical for real-time applications.

References

- [1] Amazon Web Services. 2017. Introducing Amazon EC2 P3 Instances. <https://aws.amazon.com/about-aws/whats-new/2017/10/introducing-amazon-ec2-p3-instances/>. Accessed: 2025-02-08.
- [2] T. et al. An. 2024. IoT-LLM: Enhancing Real-world IoT Task Reasoning with Large Language Models. *arXiv preprint arXiv:2410.02429* (2024).
- [3] O. et al. Baris. 2025. Foundation Models for CPS-IoT: Opportunities and Challenges. *arXiv preprint arXiv:2501.16368* (2025).
- [4] H. et al. Chen. 2025. Towards Edge General Intelligence via Large Language Models: Opportunities and Challenges. *IEEE Network* (2025).
- [5] Z. et al. Chen. 2022. FedSEA: A Semi-Asynchronous Federated Learning Framework for Extremely Heterogeneous Devices. In *ACM SenSys'22*. 1–14.
- [6] J. et al. Cosentino. 2024. Towards a Personal Health Large Language Model. *arXiv preprint arXiv:2406.06474* (2024).
- [7] Y. et al. Ding. 2024. Enhancing On-device LLM Inference with Historical Cloud-based LLM Interactions. In *ACM SIGKDD'24*. 597–608.
- [8] I. Gim and J. Ko. 2022. Memory-efficient DNN Training on Mobile Devices. In *ACM MobiSys'22*. 464–476.
- [9] Google. 2025. Google Pixel 8 Specifications.
- [10] Google AI. 2025. MediaPipe Solutions Guide.
- [11] P. et al. Guo. 2021. Mistify: Automating DNN Model Porting for On-Device Inference at the Edge. In *USENIX NSDI'21*. 705–719.
- [12] S.A. et al. Imran. 2024. LLaSA: Large Multimodal Agent for Human Activity Analysis through Wearable Sensors. *arXiv preprint arXiv:2406.14498* (2024).
- [13] F. et al. Jia. 2022. CoDL: Efficient CPU-GPU Co-execution for Deep Learning Inference on Mobile Devices. In *ACM MobiSys'22*. 209–222.
- [14] S. et al. Jiang. 2021. Flexible High-resolution Object Detection on Edge Devices with Tunable Latency. In *ACM MobiCom'21*. 559–572.
- [15] M. et al. Jin. 2023. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. *arXiv preprint arXiv:2310.01728* (2023).
- [16] B. Kim and S. Lee. 2023. On-NAS: On-Device Neural Architecture Search on Memory-Constrained Intelligent Embedded Systems. In *ACM SenSys'23*. 152–166.
- [17] S. et al. Lee. 2023. Explore, Select, Derive, and Recall: Augmenting LLM with Human-like Memory for Mobile Task Automation. *arXiv preprint arXiv:2312.03003* (2023).
- [18] S. et al. Lee. 2024. MobileGPT: Augmenting LLM with Human-like App Memory for Mobile Task Automation. In *ACM MobiCom'24*. 1119–1133.
- [19] Q. et al. Li. 2024. Governing Open Vocabulary Data Leaks Using an Edge LLM through Programming by Example. *Proc. ACM IMWUT* 8, 4 (2024), 1–31.
- [20] K. et al. Liu. 2024. Tasking Heterogeneous Sensor Systems with LLMs. In *ACM SenSys'24*. 901–902.
- [21] S. et al. Liu. 2018. On-demand Deep Model Compression for Mobile Devices: A Usage-driven Model Selection Framework. In *ACM MobiSys'18*. 389–400.
- [22] Meta AI. 2025. Llama 3.2 Model Card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md. Accessed: 2025-02-08.
- [23] NVIDIA. 2025. NVIDIA Embedded Systems for Autonomous Machines.
- [24] Ollama. 2025. Ollama: Open-Source LLMs for Local Use.
- [25] X. Ouyang and M. Srivastava. 2024. LLMSense: Harnessing LLMs for High-level Reasoning Over Spatiotemporal Sensor Traces. *arXiv preprint arXiv:2403.19857* (2024).
- [26] PyTorch. 2025. Executorch Overview.
- [27] PyTorch Contributors. 2025. XNNPACK Delegate README. GitHub: [pytorch/executorch](https://github.com/pytorch/executorch). Accessed: 2025-02-08.
- [28] R. et al. Qin. 2024. Enabling On-device Large Language Model Personalization with Self-supervised Data Selection and Synthesis. In *ACM/IEEE DAC'24*. 1–6.
- [29] P. et al. Quan. 2024. SensorBench: Benchmarking LLMs in Coding-Based Sensor Processing. *arXiv preprint arXiv:2410.10741* (2024).
- [30] M.M. et al. Rastikerdar. 2024. CACTUS: Dynamically Switchable Context-aware Micro-Classifiers for Efficient IoT Inference. In *ACM MobiSys'24*. 505–518.
- [31] S. et al. Wang. 2023. EEFL: High-Speed Wireless Communications Inspired Energy Efficient Federated Learning over Mobile Devices. In *ACM MobiSys'23*. 1–13.
- [32] H. et al. Wen. 2024. Autodroid: LLM-powered Task Automation in Android. In *ACM MobiCom'24*. 543–557.
- [33] Wikipedia contributors. 2025. Google Tensor.
- [34] Wikipedia contributors. 2025. Resident Set Size.
- [35] D. et al. Xu. 2025. Fast On-device LLM Inference with NPUs. In *ACM ASPLOS'25*. 445–462.
- [36] H. et al. Xu. 2024. AutoLife: Automatic Life Journaling with Smartphones and LLMs. *arXiv preprint arXiv:2412.15714* (2024).
- [37] H. et al. Xu. 2024. Penetrative AI: Making LLMs Comprehend the Physical World. In *ACM SenSys'23*. 1–7.
- [38] L. et al. Yang. 2022. Rep-net: Efficient On-device Learning via Feature Reprogramming. In *IEEE/CVF CVPR'22*. 12277–12286.
- [39] R. et al. Yi. 2023. Boosting DNN Cold Inference on Edge Devices. In *ACM MobiSys'23*. 516–529.
- [40] Z. et al. Yu. 2024. Edge-LLM: Enabling Efficient Large Language Model Adaptation on Edge Devices via Unified Compression and Adaptive Layer Voting. In *ACM/IEEE DAC'24*. 1–6.
- [41] J. et al. Yuan. 2024. Mobile Foundation Model as Firmware. In *ACM MobiCom'24*. 279–295.
- [42] Y. et al. Yuan. 2024. WIP: An On-device LLM-based Approach to Query Privacy Protection. In *Workshop Edge Mobile Foundation Models'24*. 7–9.
- [43] L. et al. Zhang. 2021. nn-Meter: Towards Accurate Latency Prediction of Deep-Learning Model Inference on Diverse Edge Devices. *arXiv preprint arXiv:2106.12550* (2021).
- [44] X. et al. Zhang. 2023. Unleashing the Power of Shared Label Structures for Human Activity Recognition. In *ACM CIKM'23*. 3340–3350.
- [45] Y. et al. Zheng. 2024. A Review on Edge Large Language Models: Design, Execution, and Applications. *arXiv preprint arXiv:2410.11845* (2024).
- [46] Y. et al. Zhuang. 2024. LiteMoE: Customizing On-device LLM Serving via Proxy Submodel Tuning. In *ACM SenSys'24*. 521–534.