USENIX ATC'25

CLONE:

Customizing LLMs for Efficient Latency-Aware Inference at the Edge

Chunlin Tian, Xinpeng Qin, Kahou Tam, Li Li, Zijian Wang, Yuanzhe Zhao, Minglei Zhang, Chengzhong Xu

University of Macau



Boston, July 8, 2025

Agenda



Background & Motivation

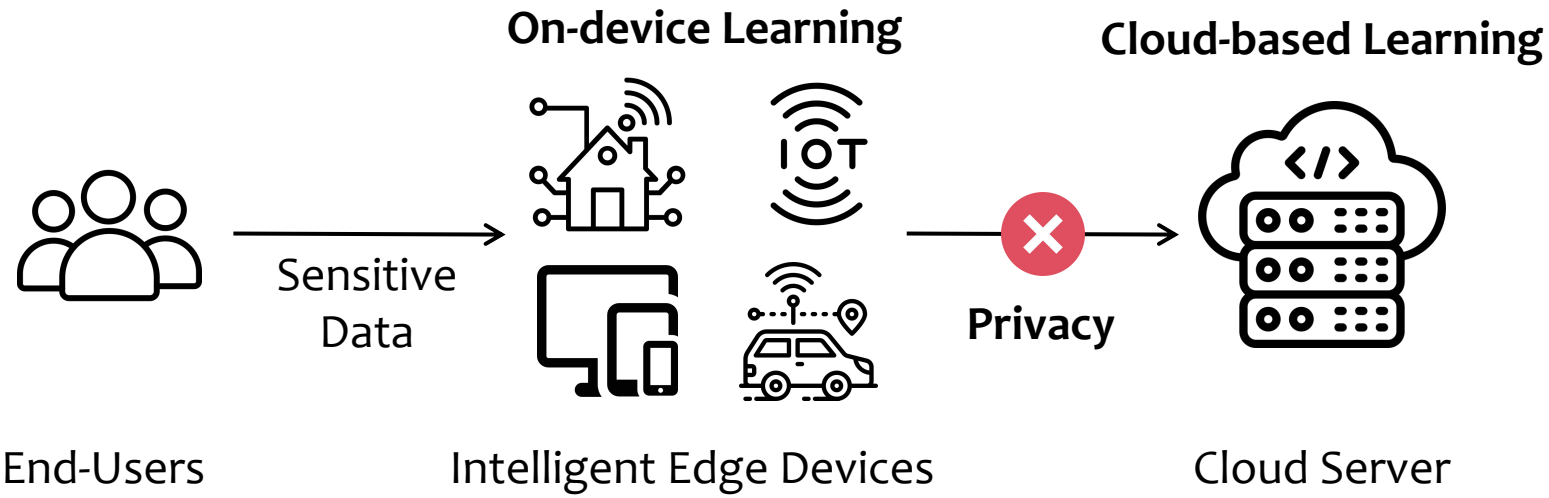
02 Design

03 Evaluation

04 Conclusion

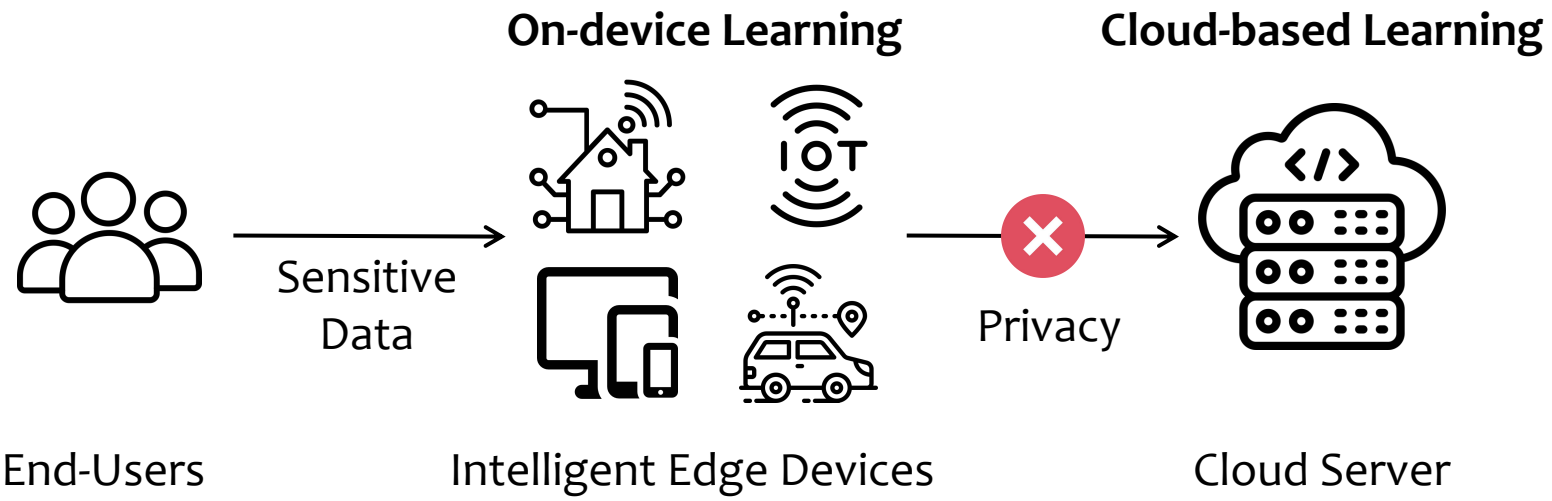
Why deploy LLMs on device?

Processing data near the source (edge) instead of the cloud.



Why deploy LLMs on device?

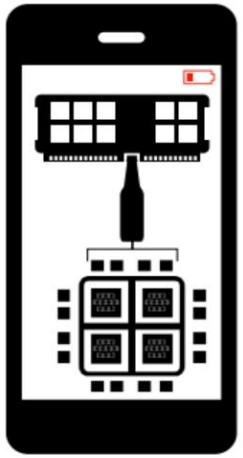
Processing data near the source (edge) instead of the cloud.



The last-mile of AI !

- **Low Latency**: Real-time decisions.
- **Privacy & Personalisation**: Keep data on the device.
- **Bandwidth**: Reduced cloud network dependency.
- **Sustainability and Cost-efficiency**.

On-device workload bottlenecks



1. High Memory Footprint.

(e.g.: Llama-7B, 14GB param. <FP16>, KV Cache: 0.5MB per token, 4096 tokens demands 2GB, **Out-of-Memory**)

2. High Compute-bound

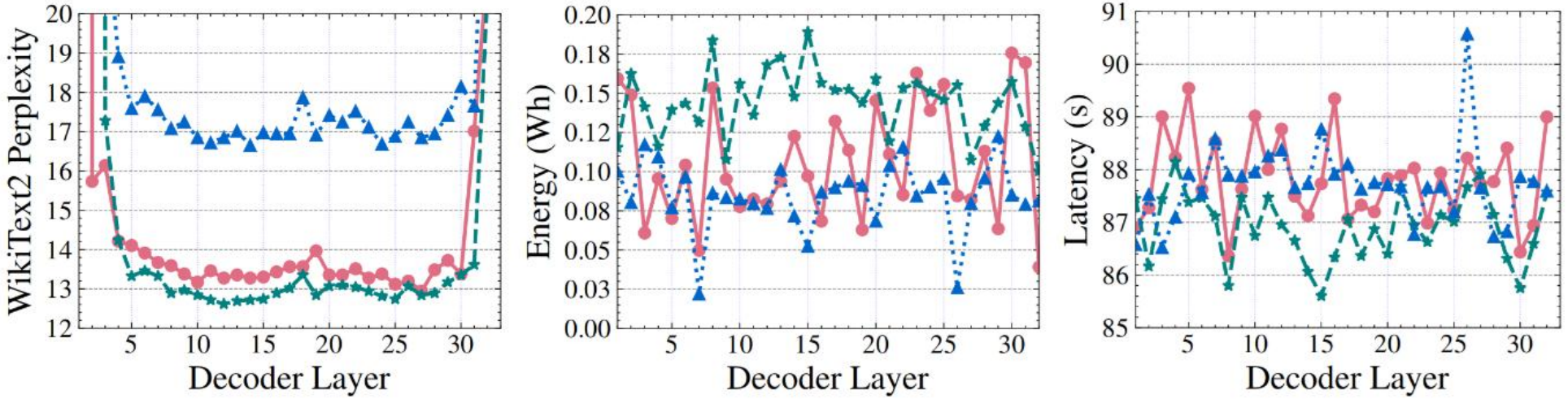
(e.g.: A prompt with 11 tokens on Llama2-7B requires approximately 14 TFLOPS , which is roughly **360×** the 39 GFLOPs needed by VGG-19 to process a single image)

3. High Energy-bound

(e.g.: GPT-3 consumes 300 J/response on an NVIDIA A100 GPU, which is **400×** the 0.75 J/response required by ResNet-5;
a Google search driven by a large AI model expends 8.9 watt-hours (Wh) of energy, approximately **30×** the 0.3 Wh required by a standard Google search)

Opportunities to deploy LLMs on device

1 Heterogeneity of Model Characteristics



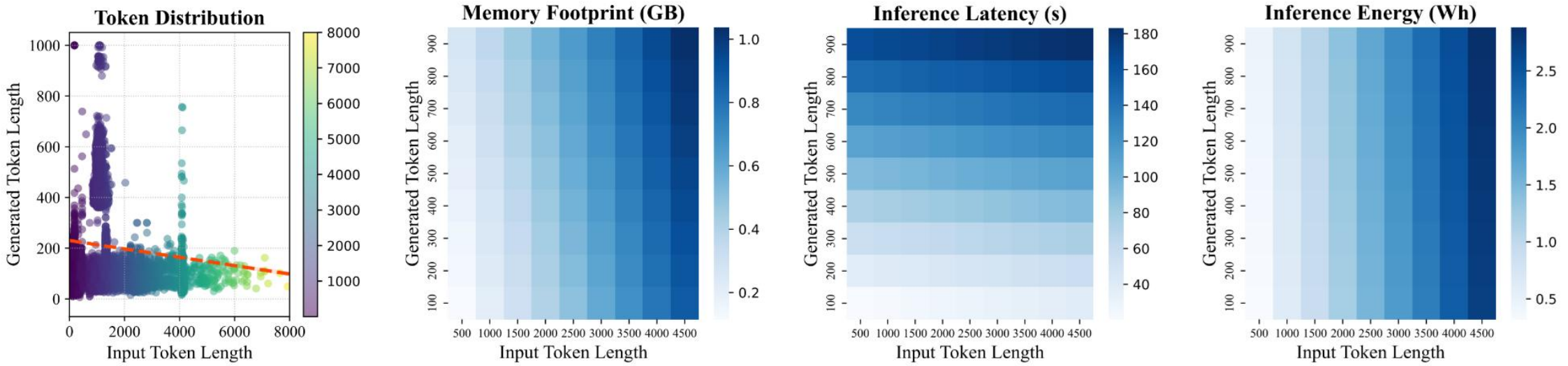
(a) Generative-ability (b) Energy (c) Latency

Figure: Layer sensitivity analysis: removing specific layers one-by-one to measure each layer's impact on Llama-7B (•), Llama2-7B (★), and Vicuna-7B (▲)

Homogeneous LLM layers contribute heterogeneously to both effectiveness and efficiency

Opportunities to deploy LLMs on device

2 Heterogeneity of Stochastic Input/Output



(a) Practical inference traces. (b) Memory consumption. (c) Latency analysis. (d) Energy consumption.

Figure: Comprehensive analysis of LLM inference system using Gemma-2B on NVIDIA Orin NX. (a) Analysis of practical LLM inference traces., while (b-d) highlight memory, latency, and energy consumption w.r.t. input and generated length.

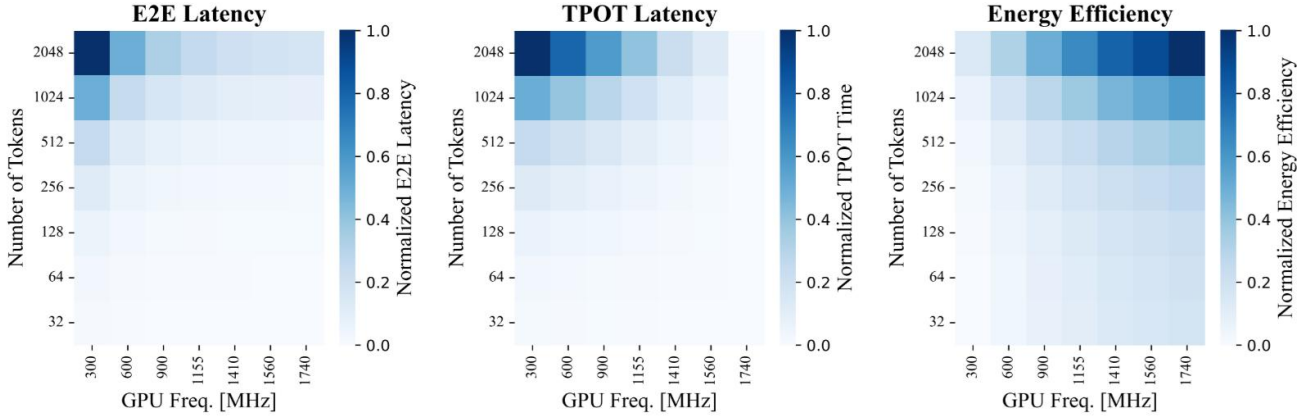
Stochastic I/O with mixed-task requests underscores the need for precise inference services

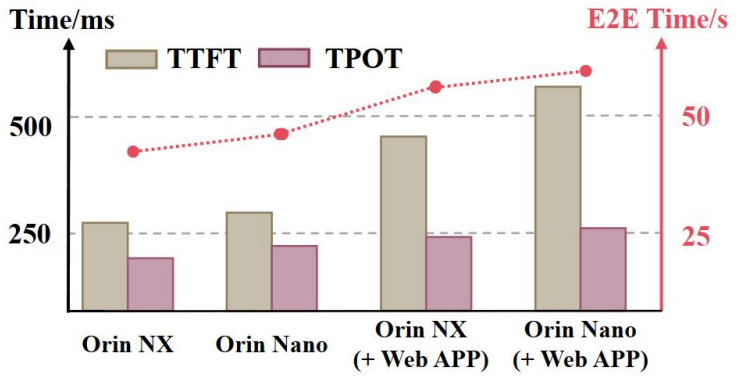
Opportunities to deploy LLMs on device

3 Heterogeneity of Runtime Platform

LLM Application	Service-Level Objective
Chatbot [5, 106]	Readable TTFT/TPOT
Search Engine [95]	Low TTFT, Medium TPOT
Event Logger [1, 58, 108]	Tolerable TTFT, Medium TPOT
Smart Reply [42, 96]	Low TTFT, Low TPOT
Code Generator [39]	Medium TTFT, High TPOT
Virtual Assistant [7, 41]	Low TTFT, Medium TPOT

SLOs of various LLM applications





Impact of different hardware devices (Nano/NX) and concurrently running apps (web search) on edge LLMs inference.

Hardware type, runtime varians and LLM application SLOs significantly impact end-user QoE.

Figure: Impact of different token lengths and GPU frequencies on a) end-toend (E2E) latency, b) time per output token (TPOT), and c) energy efficiency.

Agenda

01 Background & Motivation

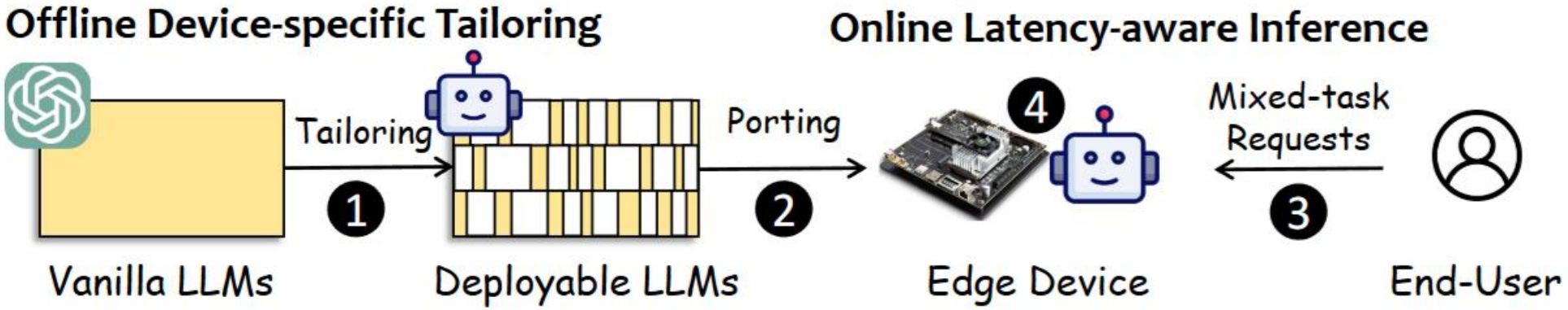


Design

03 Evaluation

04 Conclusion

CLONE Design

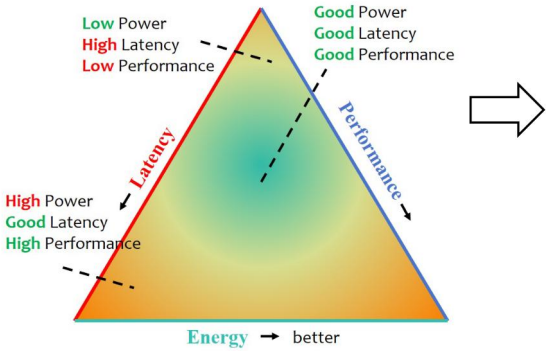


Overview

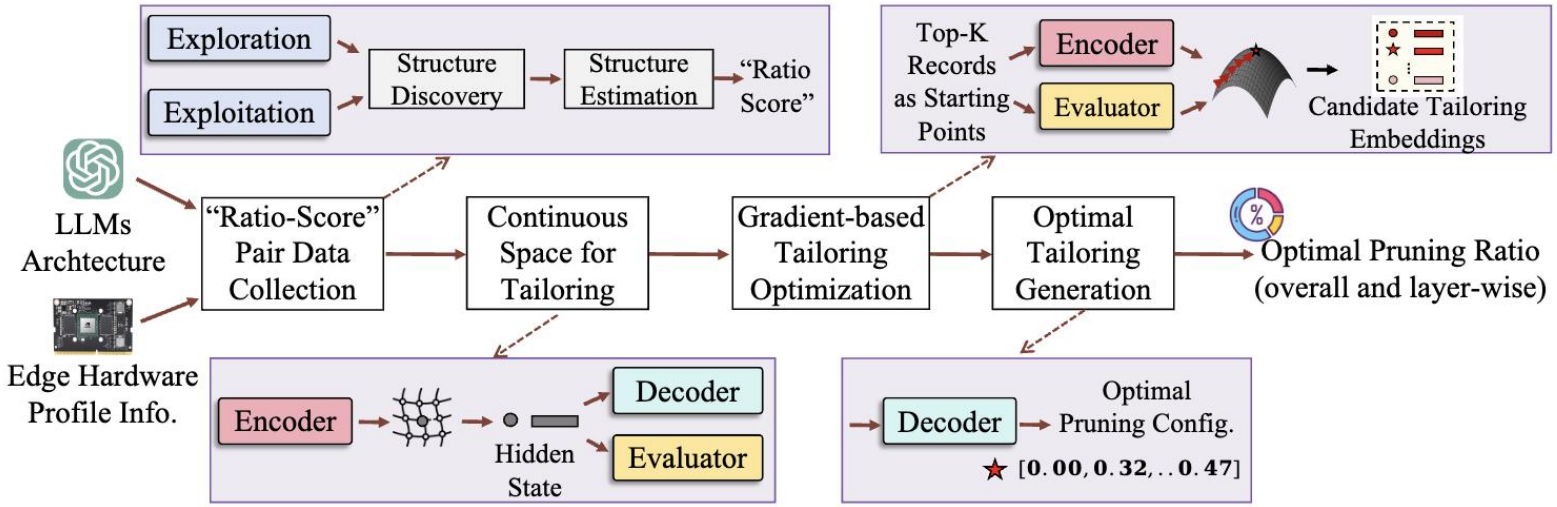
CLONE Design: Offline Device-specific Tailoring

Objective:

$$\frac{1}{ppl} \times \left(\frac{E}{e_i}\right)^{\mathbf{1}(E<e_i)\times\alpha} \times \left(\frac{T}{t_i}\right)^{\mathbf{1}(T<t_i)\times\beta}$$

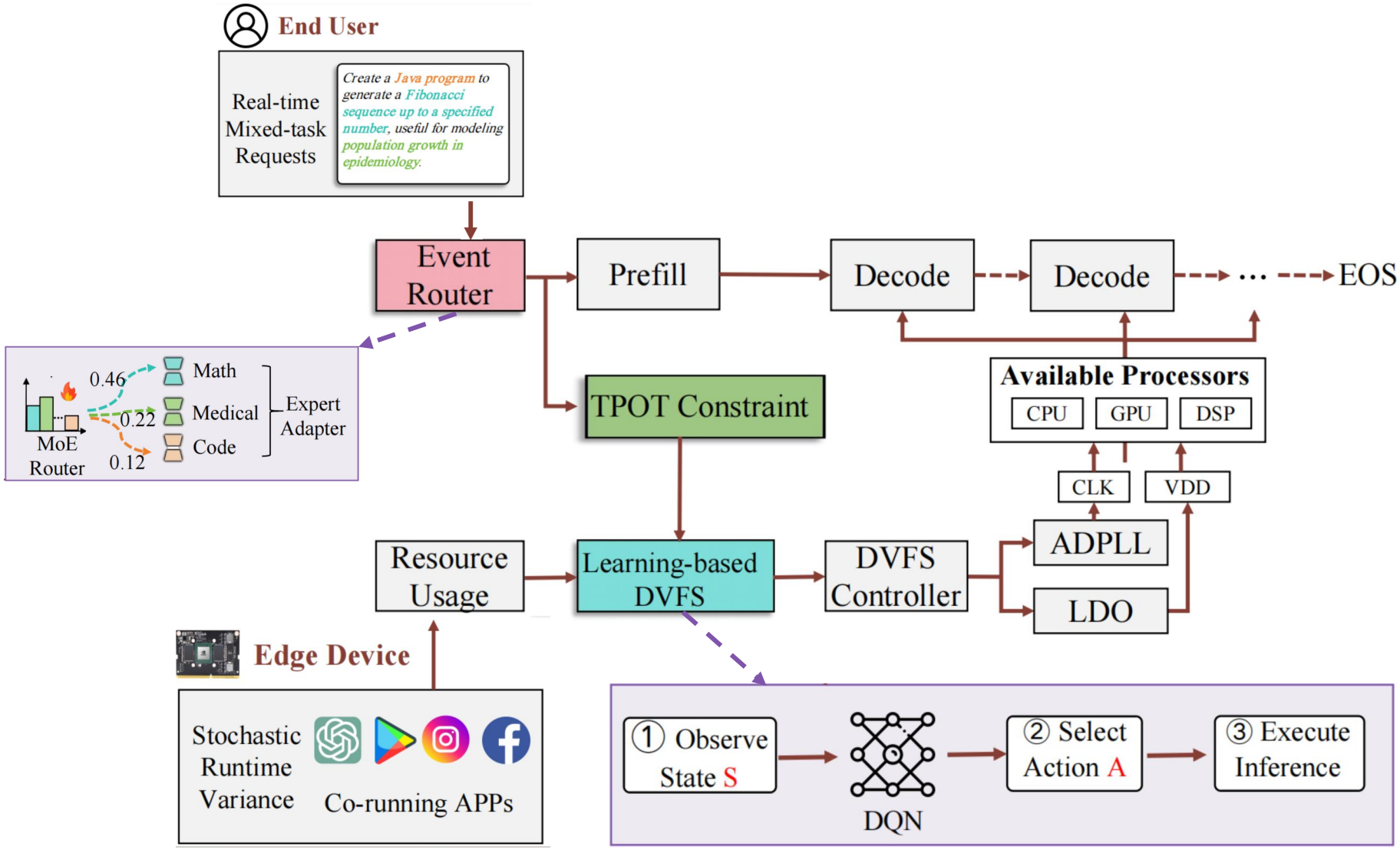


Optimization Space

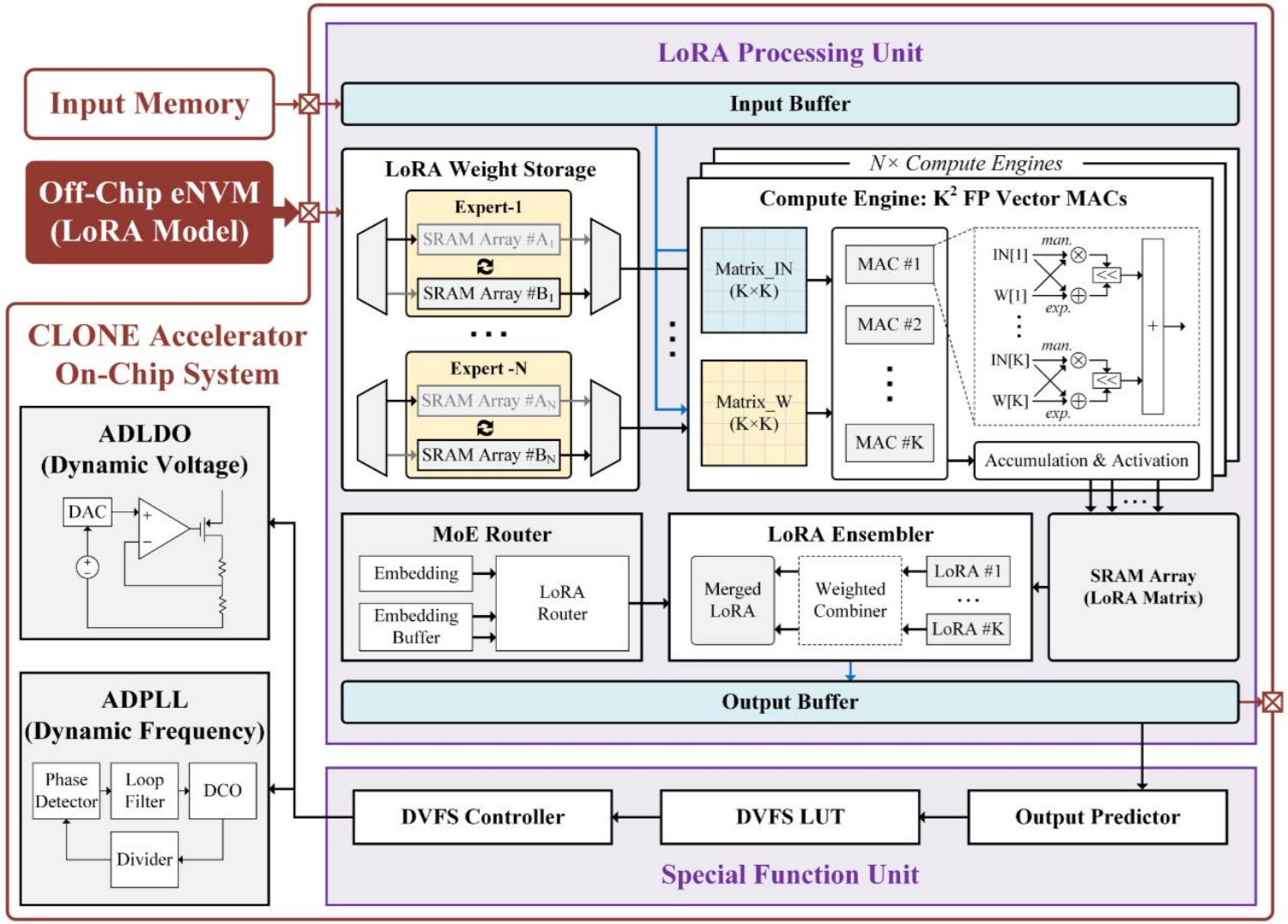


Encoder-evaluator-decoder Architecture

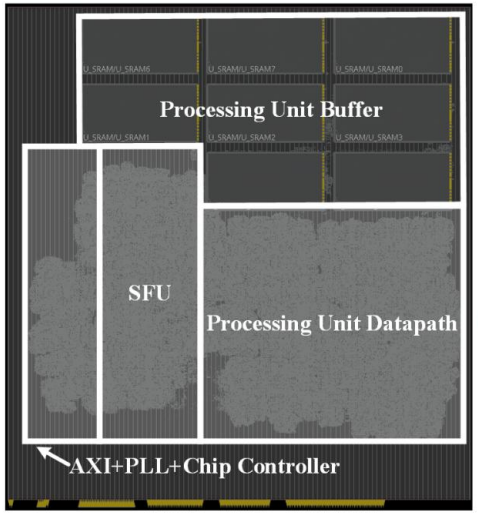
CLONE Design: Online Latency-aware Inference



CLONE Design: Online Latency-aware Inference



CLONE Design: Online Latency-aware Inference



Blocks	Area (mm^2)
PUB	0.621
PUD	0.529
SFU	0.252
Others	0.186
Total	1.588

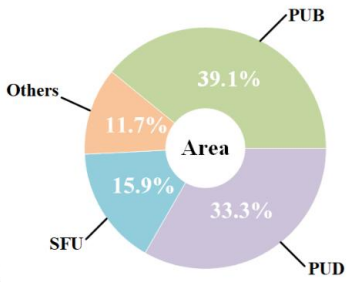


Figure: 28nm physical layout and area breakdown of the energy-optimal CLONE accelerator system.

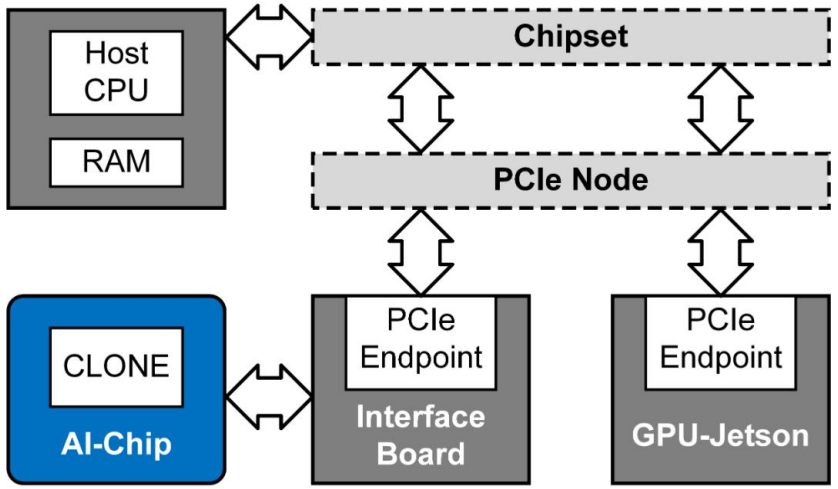


Figure: Test system. Integration of the CLONE accelerator into the Jetson platform via a PCIe interface to facilitate highperformance data exchange.

Agenda

01 Background & Motivation

02 Design



Evaluation

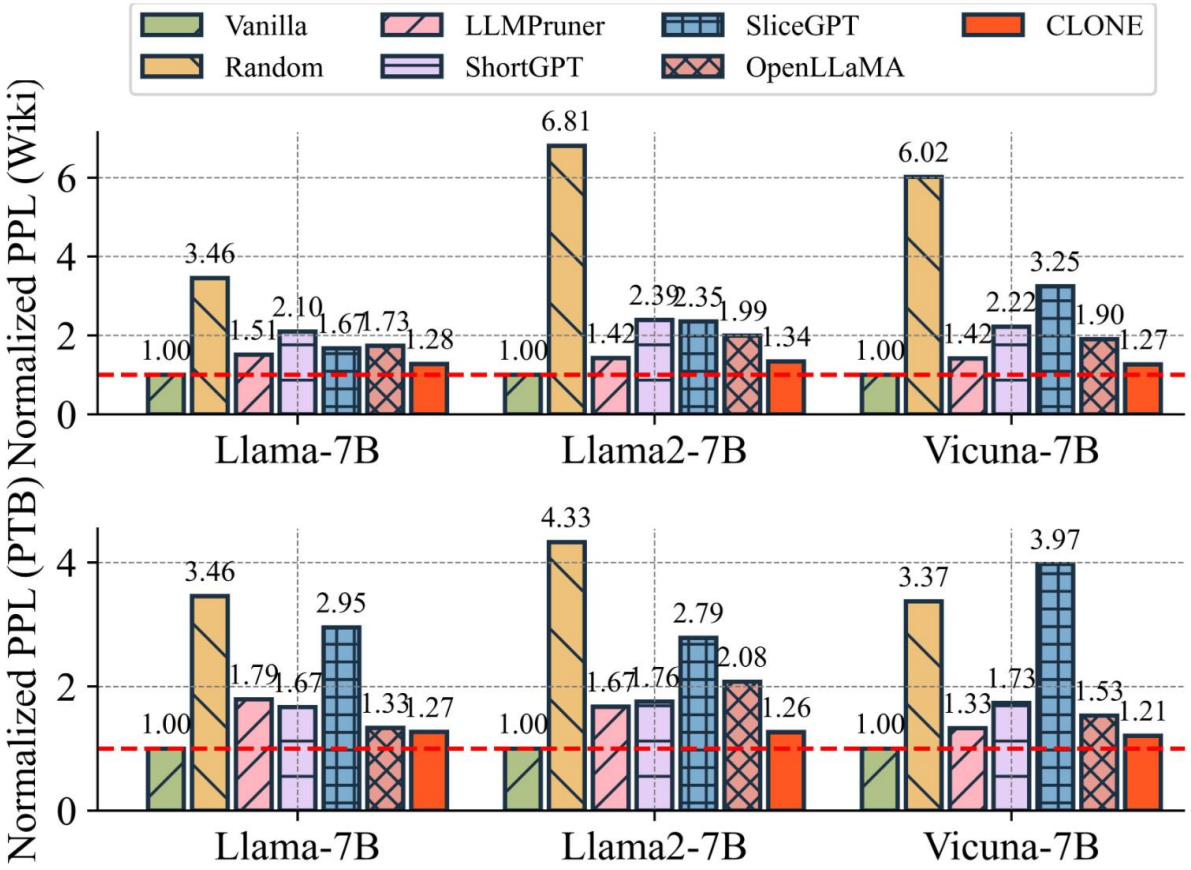
04 Conclusion

CLONE Evaluation

Infrastructure

- **Edge Devices**: Orin Nano & Orin NX;
- **Base Model**: Llama-7B, Llama2-7B, Vicuna-7B, Llama2-13B;
- **Fine-tuning Dataset**: Flanv2;
- **Test Dataset**:
  - **Generation ability**: WikiText2, PTB;
  - **General-purpose task solving ability**: BBH, MMLU, Commonsense reasoning.

CLONE Evaluation: Generation Ability



CLONE preserves the most generation ability.

- Using PPL as a pruning metric to construct a continuous pruning selection space;

- Utilizing gradient-based optimization to identify the optimal pruning configuration.

Figure: Generation ability comparison of various schemes on WikiText2 and PTB dataset across 3 representative LLMs.

CLONE Evaluation: Downstream Task Performance

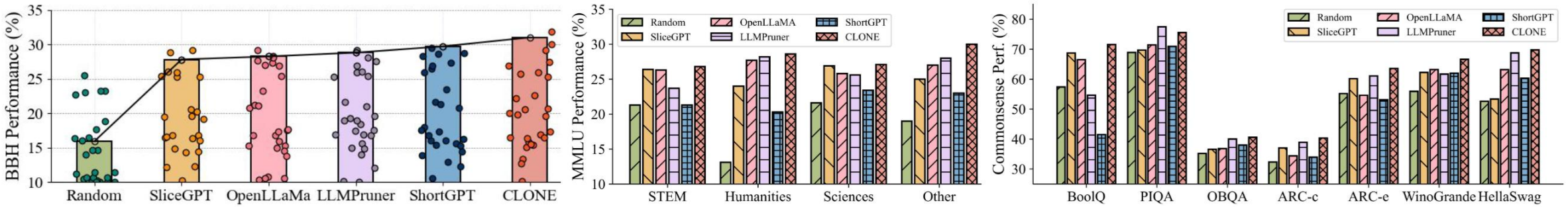


Figure: Downstream task performance of different customization approaches of Llama-7B on 3 representative benchmarks, including BBH (zero-shot), MMLU (3-shot), and Commonsense (zero-shot).

CLONE significantly outperforms other baselines across various domains and tasks.

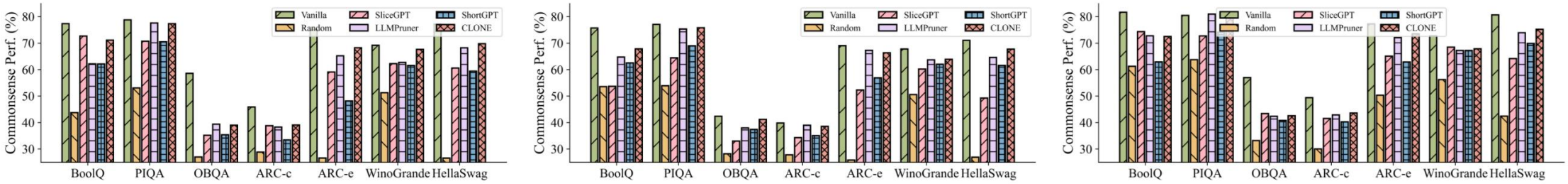
- Task-aware tailoring for specific devices;

- A request-wise MoE-based router.

CLONE Evaluation: Robustness and Scalability

Different types of models: Llama2-7B and Vicuna-7B

Different model sizes: Llama2-7B and Llama2-13B



(a) Llama2-7B (b) Vicuna-7B (c) Llama2-13B

CLONE is scalable to different LLMs in general, not limited to specific LLMs, it is not a one-off "software ASIC"

CLONE Evaluation: System Effectiveness

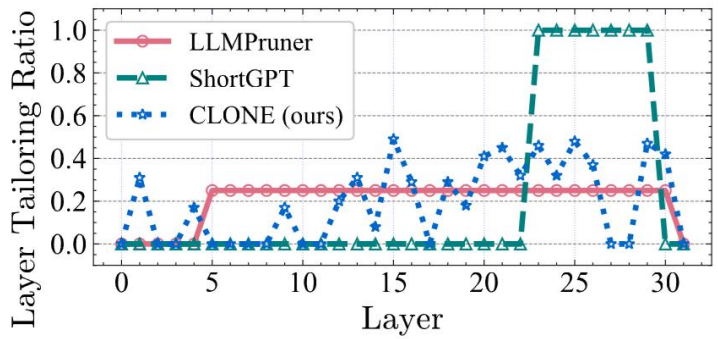
Method	Energy (Wh)		Latency (s)	
	Jetson NX	Jetson Nano	Jetson NX	Jetson Nano
Random	7.27	8.26	842.40	1145.07
SliceGPT	5.47	7.54	661.65	929.39
OpenLLaMA	5.67	7.70	506.73	662.73
LLMPruner	6.01	6.91	622.92	1023.51
ShortGPT	5.67	8.56	555.14	698.02
FlexGen	21.12	26.04	3166.27	4674.42
CLONE^{-HW}	4.81	5.56	462.72	552.18
CLONE	**3.46**	**3.54**	**322.76**	**392.15**

Table: System efficiency (latency and energy) comparison of various schemes to inference WikiText2 dataset.

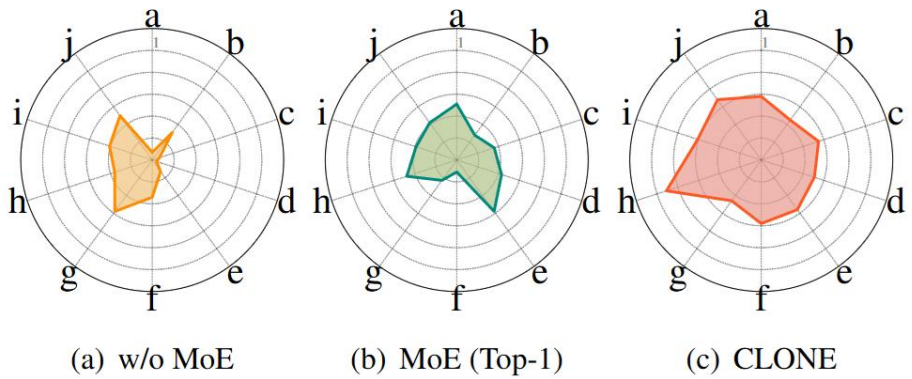
CLONE effectively speeds up the inference process and achieves significant energy saving.

- CLONE jointly considers the latency and energy budgets as metrics to generate optimal pruning configuration;
- CLONE decouples prefill and decoding phases, applying request-level DVFS to boost effectiveness;
- Special accelerator boosts efficiency by exploiting dedicated hardware resources.

CLONE Evaluation: Ablation Study



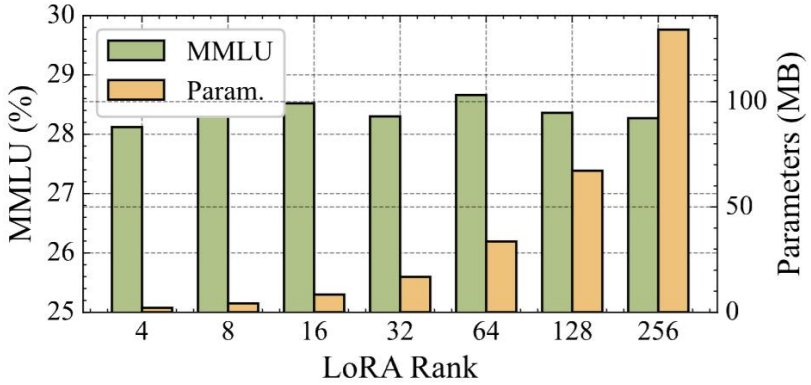
Study of the Generative Pruning Configuration.



Effectiveness of Request-wise MoE Router.

Method	Energy (Wh)		Latency (s)	
	Jetson NX	Jetson Nano	Jetson NX	Jetson Nano
CLONE^{-HW}	4.81	5.56	462.72	552.18
CLONE	**3.46**	**3.54**	**322.76**	**392.15**

Effectiveness of the Hardware Accelerator.



Impact of the Rank of LoRA Adapters.

Agenda

01 Background & Motivation

02 Design

03 Evaluation



Conclusion

CLONE Conclusion

- Proposes a **software-hardware co-design** for efficient LLM inference on edge devices, balancing speed, energy, and performance.

- Uses **offline/online optimization** to adapt to static weights, hardware, I/O patterns, and runtime variability, supported by a 28nm specialized accelerator.

- Evaluated through extensive experiments on commercial edge devices using real-world LLMs and benchmarks.

CLONE: Customizing LLMs for Efficient Latency-Aware Inference at the Edge

Chunlin Tian, Xinpeng Qin, Kahou Tam, Li Li, Zijian Wang, Yuanzhe Zhao, Minglei Zhang, Chengzhong Xu

tianclin0212@gmail.com

Thanks !

Boston, July 8, 2025