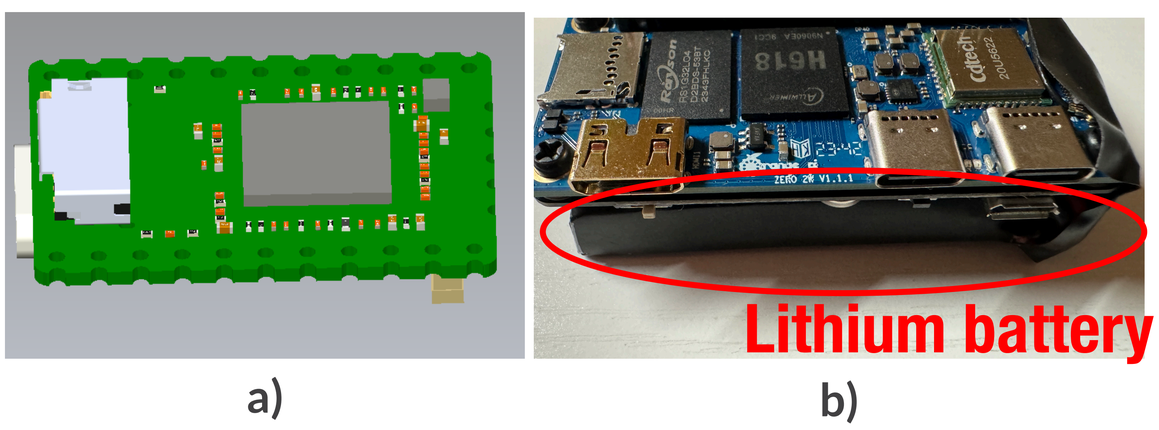
Tiny LLM Device: Enhancing Efficiency with On-Device Multi-Modal Inference on Low-Cost SoCs

Yilong Li[1], Jingyu Liu[1], Shuai Zhang[2], Hao Zhang[1], Jayaram Raghuram[1], Suman Banerjee[1]

University of Wisconsin−Madison[1], Amazon Web Sevice[2]

Abstract

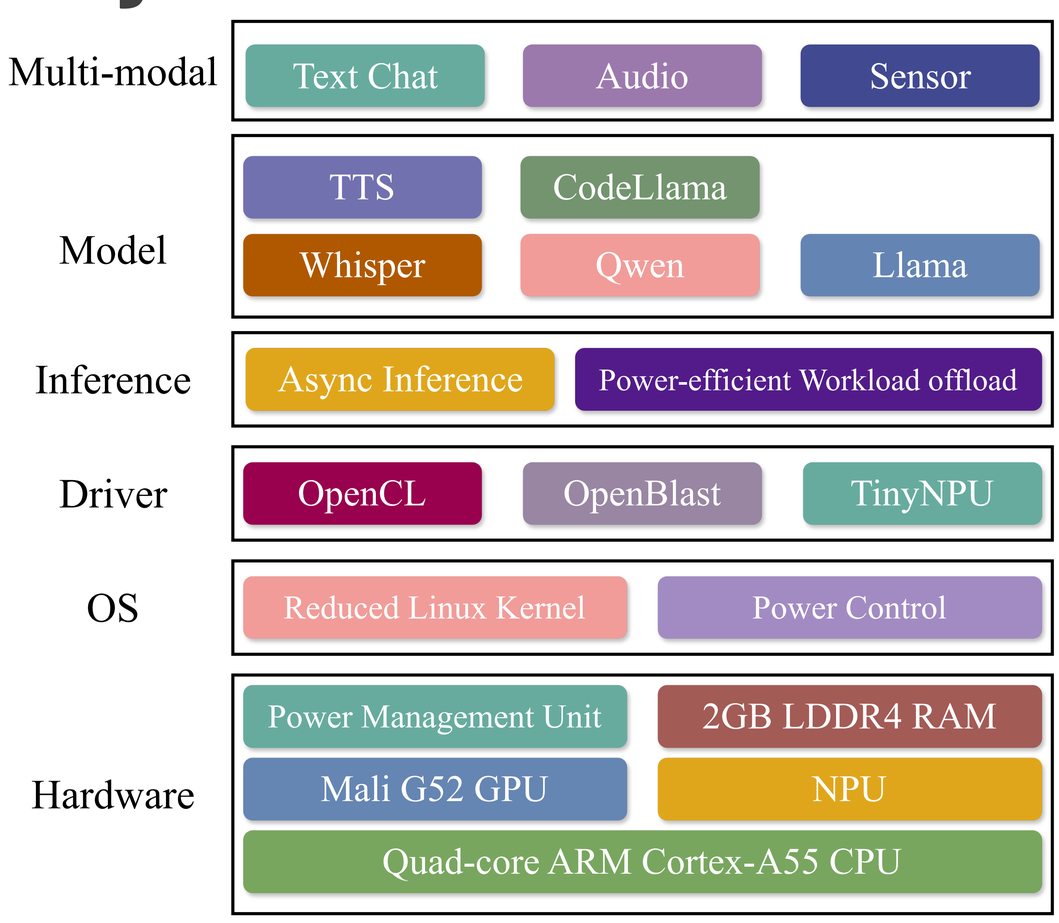
- Our on-going project develops a tiny device for efficient large language models (LLMs) inference optimized for low-power, embedded SoCs.
- It operates independently on minimal battery power for up to two days, providing intelligent, interactive LLM inference without internet connectivity. Enhanced by NPU offloading, GPU scheduling, and advanced optimization techniques, our device integrates language, voice, and sensor data for seamless human-device interaction.



a) b)

Figure 1. a) Custom-designed low power hardware with battery power. b) Prototype on development board with a Lithium battery

Design



System Architecture of TinyLLM Device

Challenges:

- Hardware Constraints and Power Management: limited CPU performance, small memory, lack of powerful GPU and short battery life.
- Model Efficiency and Size.
- Multi-modality Processing: Integrating and processing multiple types of inputs: text, audio, code and sensor data.

Performance and Functionalities

```
> I want to travel in Madison Wisconsin,from Washington DC
 Absolutely! Here are some suggestions for trips you could take in Madison:

1. Visit the State Capitol Building and grounds
2. Explore the University of Wisconsin-Madison campus
3. Take a tour of the historic downtown area
4. Check out the Madison Museum of Contemporary Art
5. Take a scenic drive along the Lake Monona Trail
6. Visit the Wisconsin Historical Society and explore its collections
7. Explore the Madison Public Market, which features local food vendors
8. Take a guided walking tour of downtown Madison
9. Attend a concert or performance at one of the many venues in the area
```

Figure 3. LLM chatbot (Llama).

```
[Start speaking]
 So what gave you that historic moment for closing?
 these teams.
 Oh, look how they're doing is rough some twists, they're pushing out, they're ready to fight, they're a
they're not ready for it!
 Get caught to the bottom of the stairs! Imagine making that call in the thirtieth round. How's your han
 hand not shaking when you make that push to take that fight out in the open green and carry the only tw
people.
 people on site they're going to be coming for it only 20 seconds left now being you don't have enough t
s a huge defense will
 all of a sudden. The windows froze. Rook is here to shut it down. He wants the overtime
 down the 10 second three and he's got the crossfire as well there's absolutely no way phase the
 It made their way back and it's over time 15-15. There was five or ten seconds there.
 that. When now the more instructing distance of the people, I don't want to be trying to defend again.
 [BLANK_AUDIO]
 [ Silence ]
```

Figure 4. Whisper real-time speech to text.

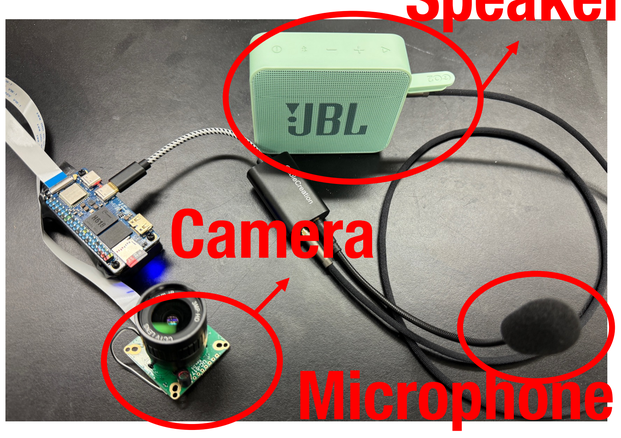


Figure 5. Tiny LLM Device with multi-modality (Voice, Vision).

```
USER: How is the picture showing?
ASSISTANT: The picture is showing in black and white.
USER: What is it for?
ASSISTANT:  The pier is for walking across the water, providing a bridge-like structure for p
eople to cross the lake or river.
USER: Are you an on-device model?
ASSISTANT: Yes, I am an on-device model.
USER: Can you describe the picture in detail?
ASSISTANT:  The image features a pier extending out over a body of water, such as a lake or r
iver. The pier is made of wood and provides a bridge-like structure for people to cross the w
ater. The scene is captured in black and white, adding a timeless and classic feel to the ima
ge. The pier appears to be empty, with no people visible in the picture. The focus is on the
pier itself and the water beneath it, creating a serene and peaceful atmosphere.
USER: hi
ASSISTANT:  Hello! How can I help you today?
USER: What is the photo for? What is it like?
ASSISTANT:  The photo is a black and white image of a pier extending over a body of water, su
ch as a lake or river. The pier is made of wood and provides a bridge-like structure for peop
le to cross the water. The scene is peaceful and serene, with no people visible in the pictur
e. The focus is on the pier itself and the water beneath it, creating a classic and timeless
atmosphere.
USER: bye
ASSISTANT:  Goodbye! Have a great day.
```

Figure 6. Vision Multi-Modal LLM to understand the photos.

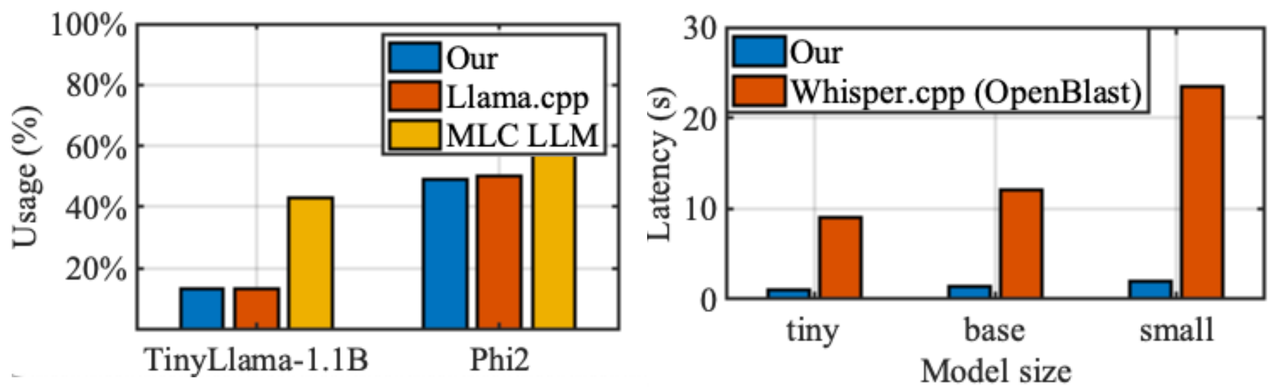


Figure 7. Performance: a) Memory usage of LLM. our system, Llama.cpp, and LLM on the same SoC, with TinyLlama and Phi2 utilizing 4-bit quantization. b) Latency in Speech to Text based on OpenAI Whisper on device.

Our tiny system outperforms existing design on tiny devices in both speed and efficiency

Technical Focus

Resource efficiency in LLM on edge server and tiny devices. Asynchronous inference for LLMs, NPU offloading, and optimized GPU inference

Contributions

We proposed an efficient LLM inference design for tiny, low-power devices, implemented through a software-hardware co-design on a small board powered by low-cost CPUs and GPUs, using battery power. The battery can support at least 2 days' usage.

References

[1]Github Repo. https://github.com/JimmyLi-Network/tiny_LLM_device.git.

CONTACT US yli758@wisc.edu info-athena@duke.edu @TheAthenaInst @The Athena AI Institute NSF AI Institute - Athena