KENNESAW STATE UNIVERSITY

Characterizing and Understanding Energy Footprint and Efficiency of Small Language Model on Edges

Supervised By

Kun Suo & Bobin Deng

Presented by

Md Romyull Islam

Outlines

Background & Motivation

Methodology & Performance Metrics

Results, Observations, & Insights

Conclusion

Outlines

Background & Motivation

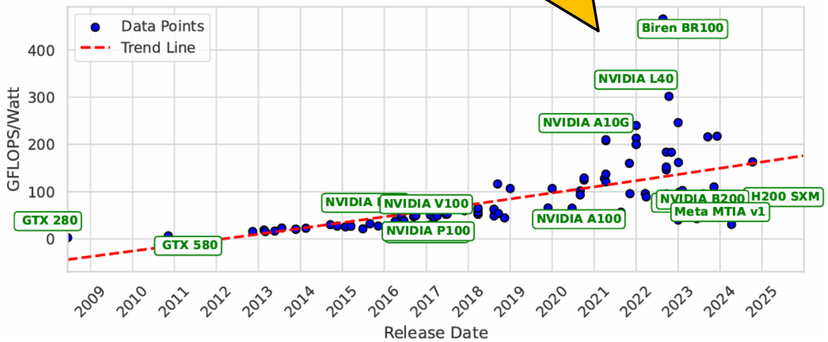
Methodology & Performance Metrics

Results, Observations, & Insights

Conclusion

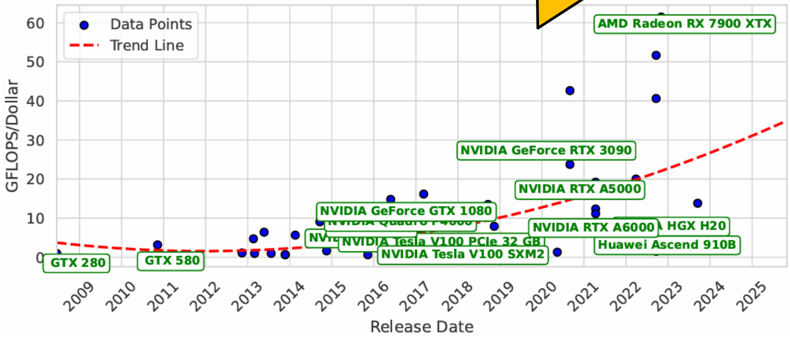
Why this Research for?

GPU energy usage is growing linearly with the increase of server GPU computational capacity



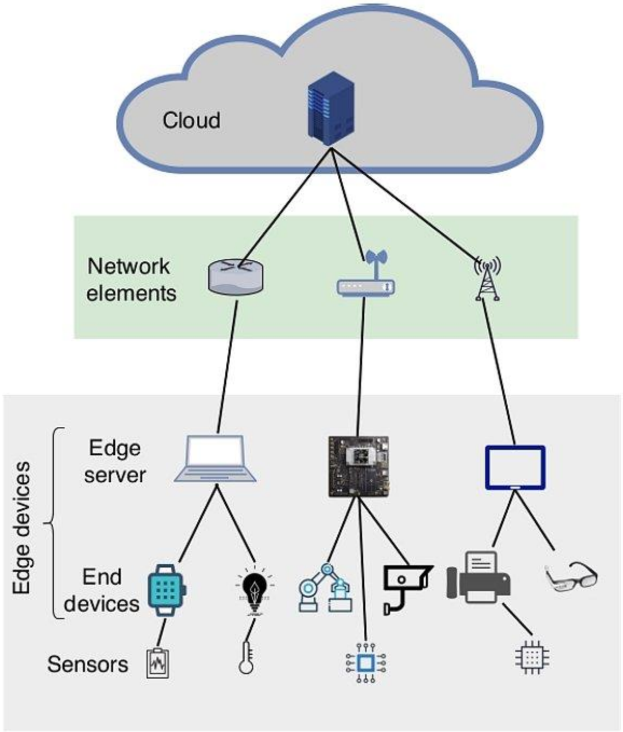
GPU Performance vs Energy Usage

GPU price is increasing exponentially over the improvement of GPU computational capacity



GPU Performance vs Hardware Price

When Edge Meets SLM



Related Works

- **Energy Utilization on Edge Devices:**

  Studies have evaluated energy demands for edge AI applications, focusing on energy and latency constraints

- **Optimizing Language Models for Edge:**

  Techniques like model pruning, quantization, and knowledge distillation have been developed to reduce model size and enhance efficiency

- **Performance in Custom Edge-AI Systems:**

  Research on specialized edge AI systems has highlighted the need for low-latency and energy-efficient models for IoT and mobile applications

These works here has focused on finding out balance between large models or specific tasks, few studies provide a comparative analysis of small language models across various edge devices.

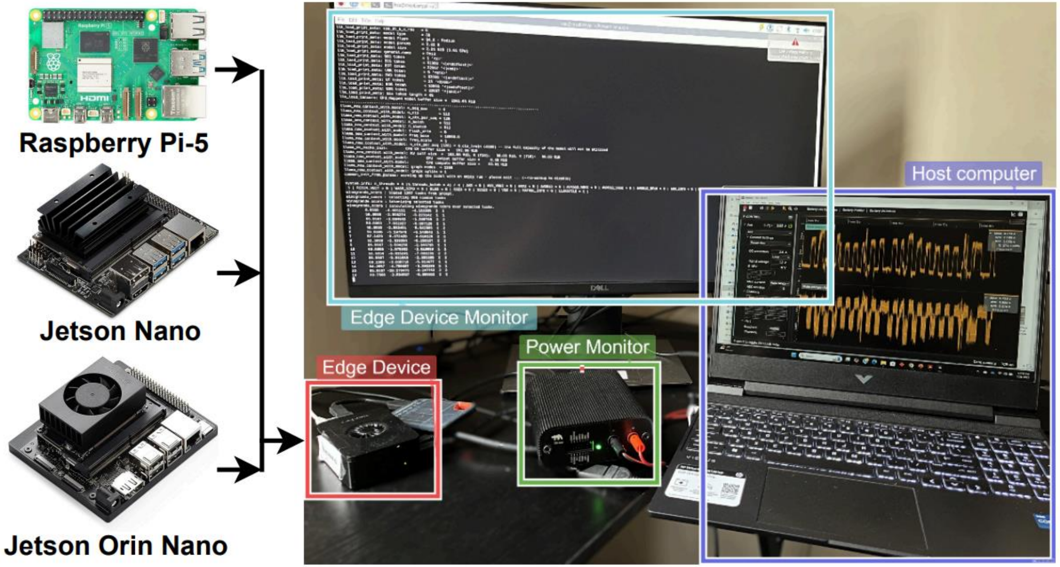
Our work aims to fill this gap.

What this Research Is All About?

- **In-Depth Studies:** This research delves into balancing energy efficiency and performance in Small Language Models (SLMs).
- **Energy Footprint Evaluation:** Systematic collection and analysis of data to evaluate SLMs' real-world performance on edge devices.
- **Insights and Recommendations:** Findings provide practical recommendations for hardware and model selection in edge computing tailored to the specific requirements of SLMs.

Outlines

Used Hardware Setup



Device Name	Memory	Memory Freq	Memory Band	Memory Type	CPU Freq.	GPU Freq.	CPU Core	GPU Core	Disk Size
Raspberry Pi 5B	4GB	4267MHz	17GB/s	LPDDR4X	2.4GHz	0.0	4	0.0	128GB
Jetson Nano	4GB	3200MHz	25.6GB/s	LPDDR4	1.43GHz	640MHz	4	128	64GB
Jetson Orin Nano	8GB	6375MHz	102GB/s	LPDDR5	1.7GHz	1020MHz	6	1024	128 GB

Language Models and Their Parameters

Model Name	Model Size	Tokens Trained on
TinyLlama	1.1B	3T
Phi-3 mini	3.8B	3.3T
Gemma 2	2B	2T
Llama 3.2-1B	1.24B	9T

Benchmarks

- **Massive Multitask Language Understanding (MMLU):** This comprehensive benchmark uses 57 multiple-choice tasks spanning various domains (e.g., Abstract Algebra, Clinical Knowledge) to assess the models' broad domain knowledge and reasoning ability.
- **HellaSwag**: A multiple-choice dataset that evaluates commonsense reasoning and contextual understanding by requiring models to select the most plausible continuation of a given textual narrative.
- **Winogrande**: This dataset focuses on pronoun resolution and contextual reasoning by presenting sentences with linguistic ambiguity, building upon the Winograd Schema Challenge to assess deep language understanding.

Performance Metrics

Higher **EDP** reflects worst performance

Metric	Description	Formula
Accuracy	Percentage of correct predictions.	(Total Predictions / Correct Predictions) ×100
Latency	Average time taken for one inference.	Total Inferences / Total Latency
Throughput	Number of inferences completed per second.	Total Time (s) / Total Inferences
Energy per Inference (Wh)	Total energy consumed per inference.	Total Inferences / Energy (Wh)
Energy-Delay Product (EDP)	Holistic efficiency combining energy and delay.	Energy (J) × Delay (s)
Energy-Delay Product per Billion Parameters (EDP/B)	Energy-delay efficiency normalized by model size.	EDP (J·s) / Model Size (Billion Parameters)
Watt-hours per Billion Parameters (Wh/B)	Energy usage normalized by model size.	Energy (Wh) / Model Size (Billion Parameters)

Outlines

Background & Motivation

Methodology & Performance Metrics

Results, Observations, & Insights

Conclusion

Consolidated Performance Across Benchmarks And Devices

Phi-3 Mini (3.8B) achieved **64.8% MMLU**, but required up to **274.62 Wh** (Jetson Nano)

Device	Model	Acc. (%)			Energy (Wh)								
		M	H	W	M	H	W	M	H	W	M	H	W
Raspberry Pi 5	Llama 3.2	39.4	59.0	64.2	9.07	10.34	2.05	4.06	13.46	1.80	1.33e8	2.00e8	6.63e6
	Phi-3 mini	62.3	76.5	69.6	65.64	36.47	9.81	34.36	48.00	9.18	8.22e9	2.52e9	1.62e8
	TinyLlama	19.0	42.5	61.6	10.9	9.43	2.46	5.00	12.53	2.24	1.96e8	1.70e8	9.94e6
	Gemma 2	30.4	68.25	69.0	32.77	23.15	5.36	17.17	30.55	5.05	2.03e9	1.02e9	4.87e7
Jetson Nano	Llama 3.2	39.3	58.5	63.0	31.54	29.64	5.66	19.20	38.46	4.96	2.18e9	4.56e8	5.05e7
	Phi-3 mini	**64.8**	76.0	69.2	274.62	101.73	22.57	184.63	112.10	19.60	**1.83e10**	**4.56e9**	**7.97e8**
	TinyLlama	19.2	42.0	61.6	38.71	30.66	6.09	23.55	43.16	5.34	3.27e9	5.31e8	1.63e8
	Gemma 2	33.8	67.5	69.0	67.76	67.65	12.14	41.23	94.23	10.80	6.78e9	2.55e9	2.36e8
Jetson Orin Nano	Llama 3.2	39.8	58.5	63.4	9.71	5.83	1.24	3.92	5.47	0.92	1.37e8	1.28e7	5.67e5
	Phi-3 mini	63.4	76.25	69.6	45.18	20.69	4.75	18.16	19.33	3.52	2.94e9	1.60e8	8.32e6
	TinyLlama	18.0	41.75	62.4	12.04	5.59	1.28	4.90	5.24	0.94	2.12e8	1.17e7	6.17e5
	Gemma 2	33.6	67.75	68.8	20.75	12.88	2.75	8.39	12.52	2.05	4.99e8	6.44e7	2.82e6
Jetson Orin Nano (GPU)	Llama 3.2	39.4	58.0	65.2	1.71	0.434	0.102	0.57	0.33	0.05	3.50e6	**5.75e4**	2.74e4
	Phi-3 mini	64.3	76.25	70.2	6.56	1.05	0.305	2.43	1.04	0.18	**5.74e7**	4.36e4	**2.79e5**
	TinyLlama	17.4	42.0	61.0	2.06	0.427	0.110	0.78	0.34	0.06	5.81e6	5.88e4	3.22e4
	Gemma 2	33.6	68.0	68.4	2.88	0.862	0.186	1.06	0.66	0.10	1.10e7	2.27e4	9.66e4

Note: Accuracy (Acc.), Energy, Latency, and EDP are listed for MMLU (M), HellaSwag (H), and Winogrande (W). Latency is per inference.

Throughput And Efficiency Metrics

Device	Model	Ops/s & Tokens/s			Tokens/Wh			Energy/Sec (W)		
		M	H	W	M	H	W	M	H	W
Raspberry Pi 5	Llama 3.2	0.25	12.21	16.46	110.25	6317.60	7187.32	0.00223	0.00192	0.00228
	Phi-3 mini	0.03	3.88	3.72	15.23	2041.87	1737.51	0.00189	0.00190	0.00214
	TinyLlama	0.20	14.99	15.30	91.74	7949.42	6969.51	0.00218	0.00188	0.00219
	Gemma 2	0.06	5.34	5.75	30.52	2802.55	2701.68	0.00191	0.00190	0.00212
Jetson Nano	Llama 3.2	0.05	4.26	5.95	31.71	2204.96	2603.18	0.00164	0.00193	0.00228
	Phi-3 mini	0.01	1.66	1.74	3.64	7322.72	755.44	0.00149	0.00227	0.00230
	TinyLlama	0.04	4.35	6.42	25.83	2445.04	2814.62	0.00164	0.00178	0.00228
	Gemma 2	0.02	1.73	2.68	14.76	960.02	1193.77	0.00164	0.00180	0.00225
Jetson Orin Nano	Llama 3.2	0.26	30.05	32.24	102.98	11210.12	11801.61	0.00248	0.00266	0.00270
	Phi-3 mini	0.06	9.64	9.70	22.13	3601.02	3588.42	0.00249	0.00268	0.00270
	TinyLlama	0.20	35.84	36.38	83.02	13417.71	13316.41	0.00246	0.00267	0.00271
	Gemma 2	0.12	13.03	14.17	48.19	5042.06	5265.09	0.00247	0.00257	0.00268
Jetson Orin Nano (GPU)	Llama 3.2	1.76	492.17	578.90	**584.80**	150647.00	144454.90	0.00301	0.00327	0.00379
	Phi-3 mini	0.41	179.23	188.50	152.44	70954.29	55869.86	0.00269	0.00253	0.00333
	TinyLlama	1.28	545.81	601.49	485.44	**175690.16**	**155863.64**	0.00263	0.00311	0.00377
	Gemma 2	0.94	246.81	296.46	347.22	75286.29	77854.84	0.00271	0.00328	0.00358

Note: Ops/s(for MMLU) & Tokens/s, Tokens/Wh, and Energy/Sec (W) are listed for MMLU (M), HellaSwag (H), and Winogrande (W).

Model Efficiency

Metric	Best Model	Value (Benchmark/Device)	Implication
Tokens/Wh (Efficiency)	**TinyLlama** (1.1B)	**175,690**	Ultra-low power champion (HellaSwag / Orin Nano GPU)
Normalized EDP/B	**Llama 3.2** (1.24B)	$2.42 * 10^4\ J.s\ /B$	Lowest energy-delay cost per parameter (Avg. / Orin Nano GPU)
Worst EDP	**Phi-3 Mini** (3.8B)	$1.83 * 10^{10}\ J.s\ /B$	Least efficient overall (MMLU / Jetson Nano)

Hardware Impact: GPU vs. CPU

GPU acceleration is the dominant factor, reducing latency and energy consumption by orders of magnitude compared to CPU-only setups.

Metric	Best Hardware	Value (Llama 3.2)	Comparison
Latency (MMLU)	**Orin Nano (GPU)**	**0.57 seconds**	~33x faster than Jetson Nano CPU (19.2s)
Energy/Inference (MMLU)	**Orin Nano (GPU)**	**0.00171 Wh**	Highly optimized for single tasks
Max Throughput	**Orin Nano (GPU)**	**578.9 Tokens/s**	Drastically improved performance

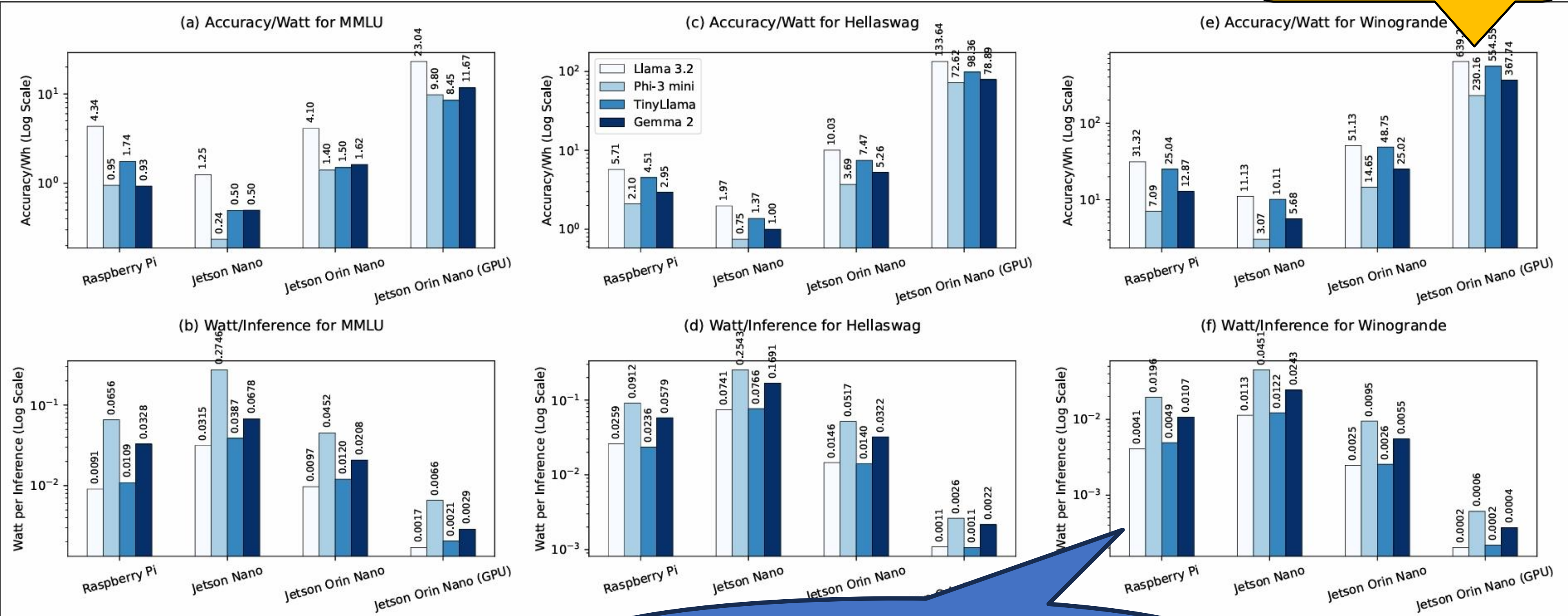
Normalized Energy And Efficiency Metrics Per Billion Parameters

Model	Size (B)	Raspberry Pi 5		Jetson Nano		Jetson Orin Nano		Jetson Orin Nano (GPU)	
		Wh/B	EDP/B (J·s/B)	Wh/B	EDP/B (J·s/B)	Wh/B	EDP/B (J·s/B)	Wh/B	EDP/B (J·s/B)
TinyLlama	1.1	6.91	1.46e8	22.86	4.96e8	5.73	1.15e7	0.79	2.94e4
Llama 3.2	1.24	5.77	1.16e8	17.97	3.04e8	4.51	9.58e6	**0.60**	**2.42e4**
Gemma 2	2.0	10.22	2.15e8	24.59	**4.83e8**	6.06	1.17e7	0.66	4.43e4
Phi-3 mini	3.8	9.82	2.16e8	**35.00**	4.82e8	6.20	1.64e7	0.69	7.34e4

Note: Wh/B = Watt-hours per Billion Parameters, EDP/B = Energy-Delay Product per Billion Parameters. Values are averaged across the benchmarks.

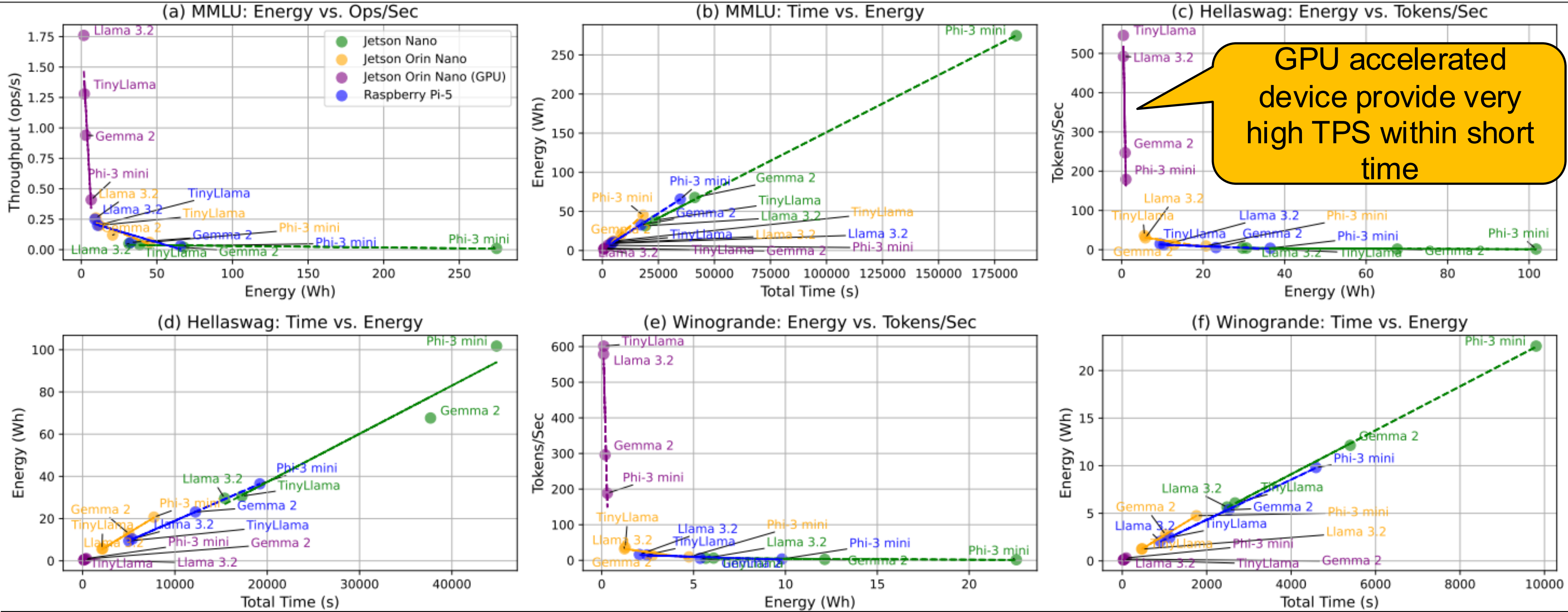
Prediction Accuracy Per Watt-Hour and Energy Consumption Per Inference

Lamma 3.2 Provide maximum accuracy compared to energy usage



Phi-2 Mini Consumes the most energy for each inference

Comparison of Total Time, Energy Consumption, and Tokens Per Second



Outlines

Background & Motivation

Methodology & Performance Metrics

Results, Observations, & Insights

Conclusion

Conclusion and Takeaways

- On GPU-accelerated devices (Jetson Orin Nano) offers the optimal balance for edge AI, providing strong accuracy with superior energy efficiency (lowest Energy-Delay Product and Watt-hours per Billion Parameters), while the higher accuracy of Phi-3 Mini comes at an impractical cost in energy consumption and latency.

- Hardware choice is critical for sustainable edge deployment, as GPU acceleration and high memory bandwidth (as found in the Jetson Orin Nano) are essential factors that drastically minimize inference latency and power draw across all models compared to CPU-only setups.


Q
&
A