Estimate Inference Energy of Neural Network

Adrien Sardi

GreenDays 2026

As part of Nokia & Inria Common Research Lab

Adrien Sardi[1,2,3], Marie Line Alberi-Morel[1], Sara Alouf[2], Frédéric Giroire[2,3], Joanna Moulierac[2,3]

[1] Bell Labs, Nokia Networks France
[2] Inria Université Côte d'Azur
[3] Université Côte d'Azur, CNRS

Introduction

Methodology

Results

Next Steps

Introduction

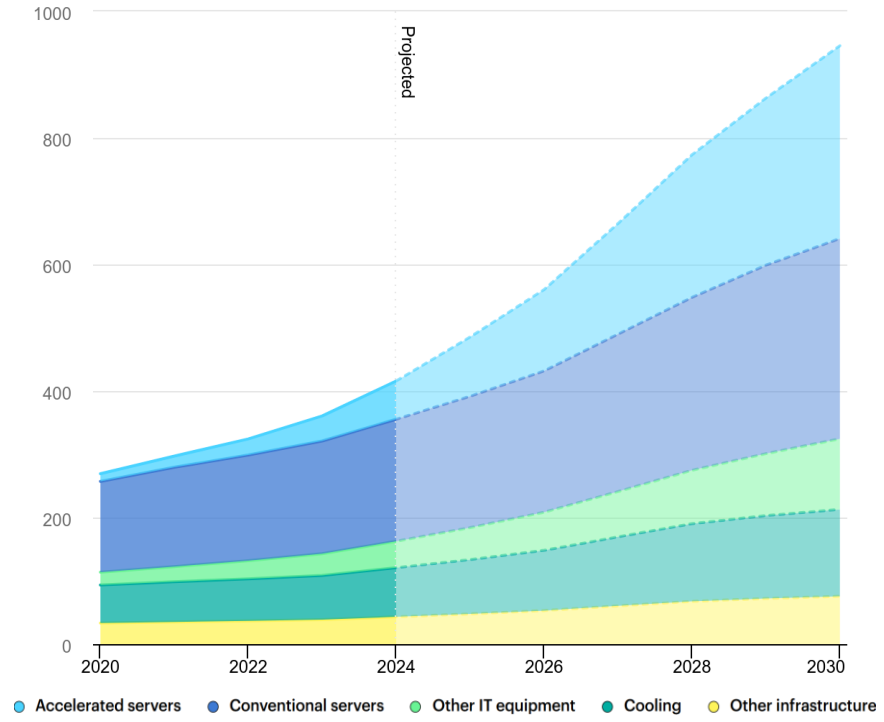
Methodology

Results

Next Steps

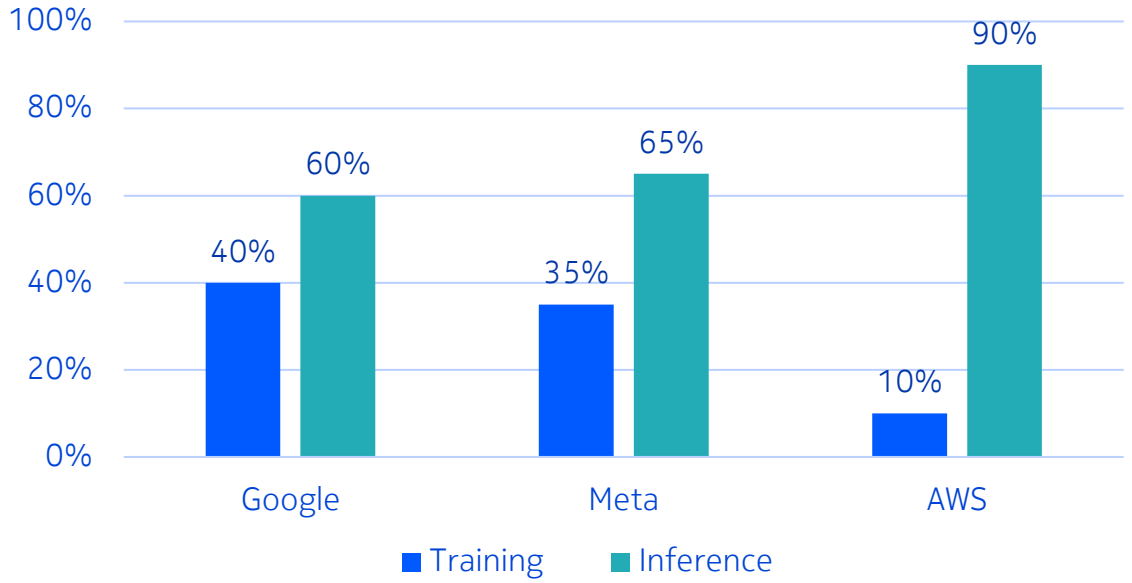
AI Energy Demand Increases

Global DC electricity consumption will increase, mainly driven by AI



Source: Energy and AI report 2025 International Energy Agency (IAE)

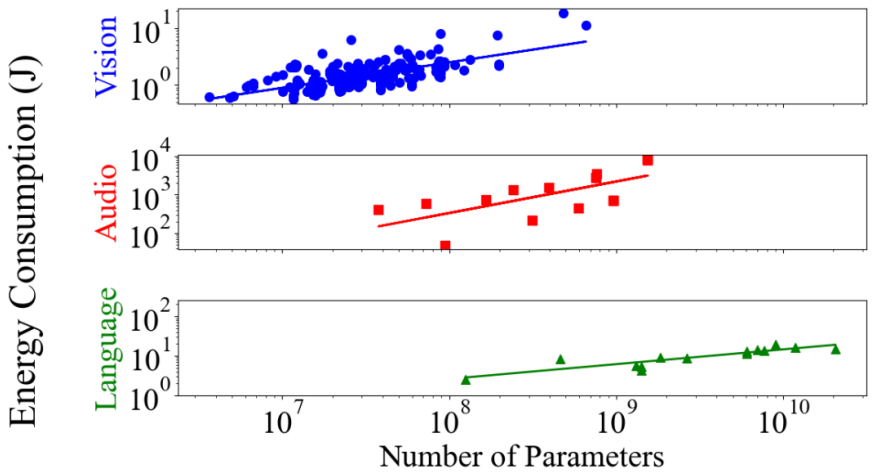
Inference phase often surpasses training in energy consumption



Source: (Patterson et al., 2022; Wu et al., 2022; De Chateauvieux et al., 2022)

Estimating the Energy Consumption AI Algorithms

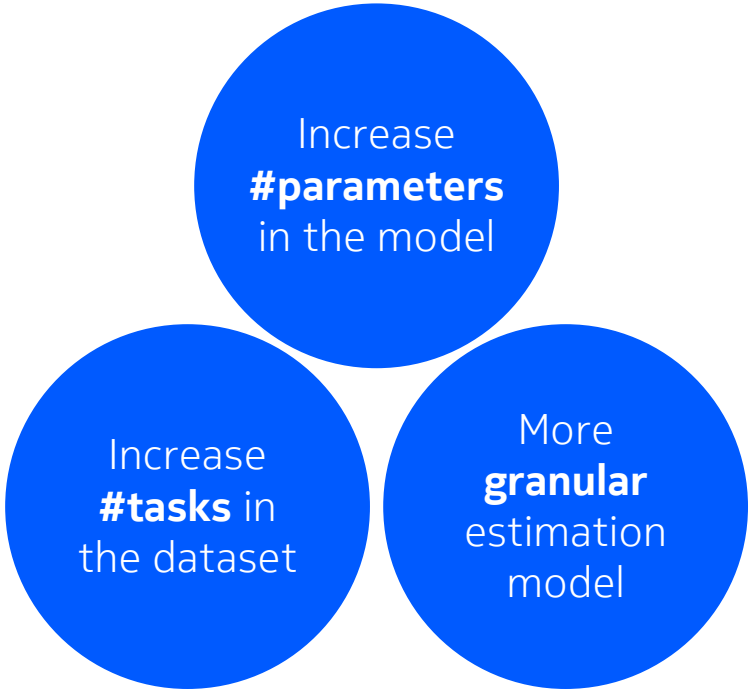
A simple model…



Source: Barros et al., "Small is Sufficient: Reducing the World AI Energy Consumption Through Model Selection" (https://doi.org/10.48550/arXiv.2510.01889)

- Linear model with one feature : $y = a \cdot \text{Feature} + b$
- One parameter fitting provides poor results across **tasks** or **scenarios**

… that needs to be improved



State of the Art (SOTA)

Model	Features					Task			Model Type
	MACs/FLOPs	Activation	Parameters	Input Size	Batch Size	Vision	NLP	Audio	Layer-wise
(Cai et al. 2017)	X					X			X
(Getzner et al. 2023)	X					X			X
MAC[1] (Desislavov et al. 2023)	X					X			
HJ[2] (Yang & Armour 2025)	X	X	X			X			
WattLayer	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**

[1] *Linear model with one feature being number of Multiple accumulated operations (MACs):* $y = a \cdot MAC + b$

[2] *Log-Linear Model with 3 features:* $y = e^{const} \times MAC^{A} \times Activations^{B} \times Parameters^{C}$

➔ **We propose a task independent layer-wise methodology to estimate the energy consumption of AI algorithms**

Introduction

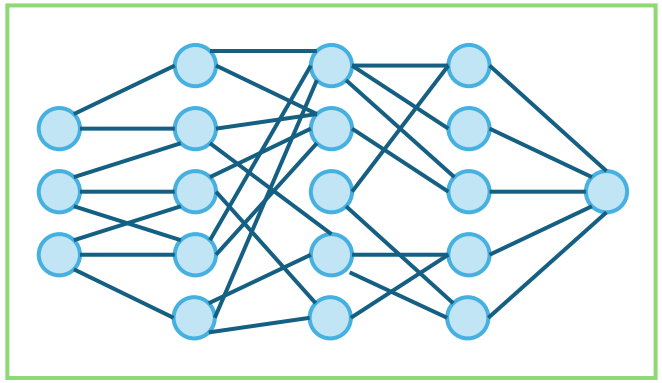
Methodology

Results

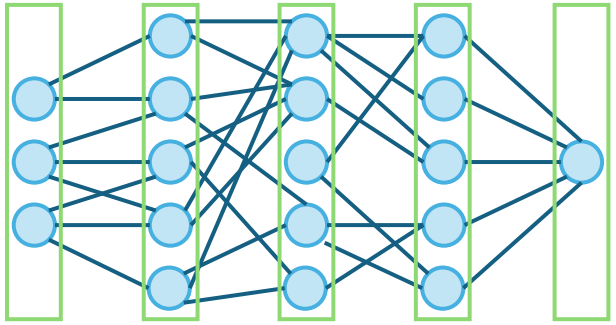
Next Steps

Methodology

Layer-wise Estimation



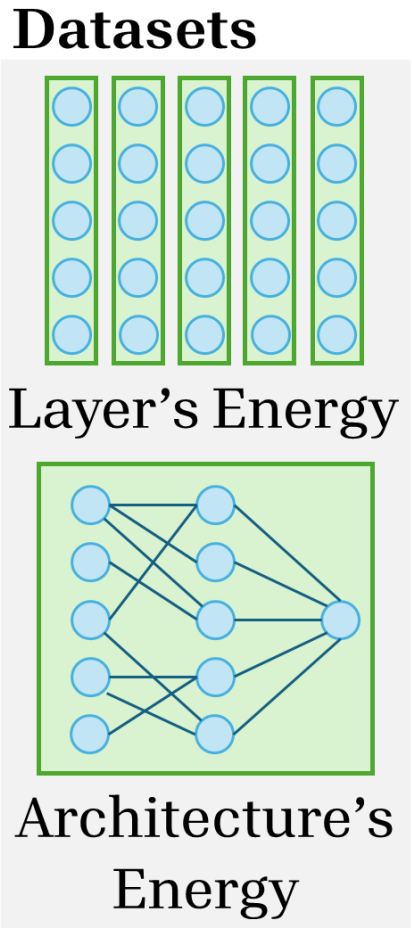
$$= E_{network}$$



$$= \sum_{i=0}^{n_{layer}} E_{layer_i}$$

Methodology

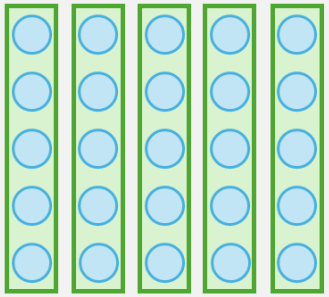
1



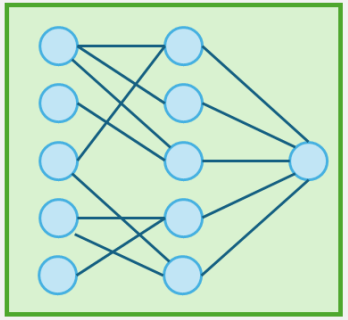
Methodology

1

Datasets



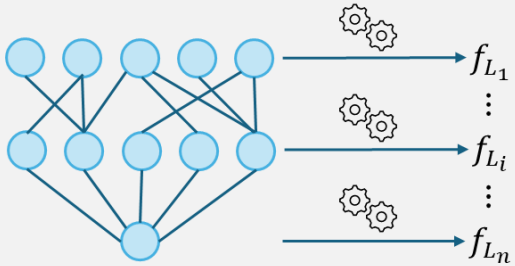
Layer's Energy



Architecture's Energy

2

Train

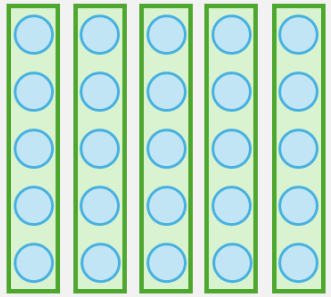


Fitting layer-wise estimation functions

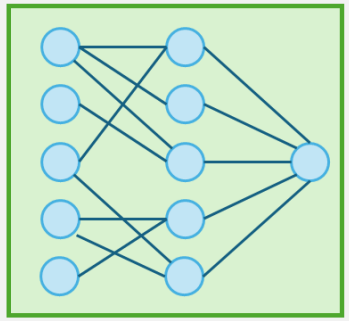
Methodology

1

Datasets



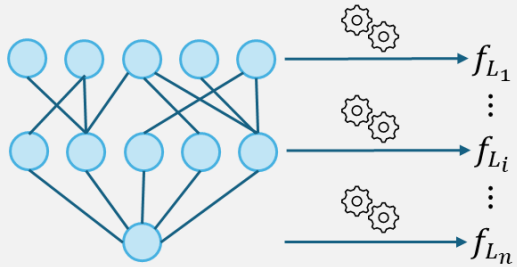
Layer's Energy



Architecture's Energy

2

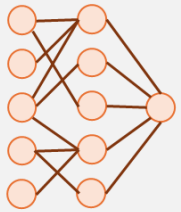
Train



Fitting layer-wise estimation functions

Layer-wise functions

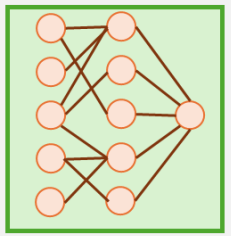
Test



$$E = \alpha \sum_{i}^{N_{layers}} f_i(X_i)$$

Neural Networks

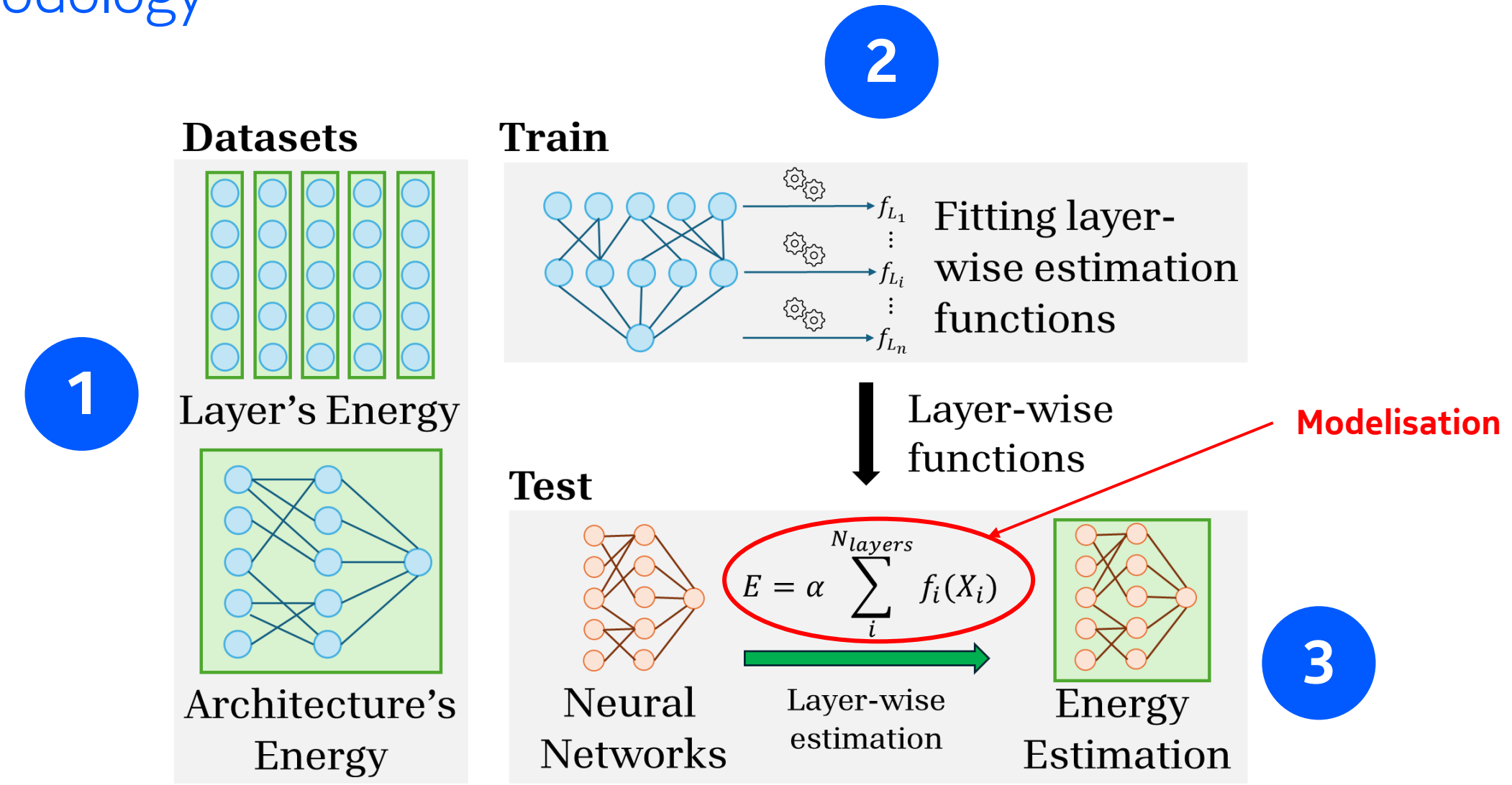
Layer-wise estimation



Energy Estimation

3

Methodology



Introduction

Methodology

Results

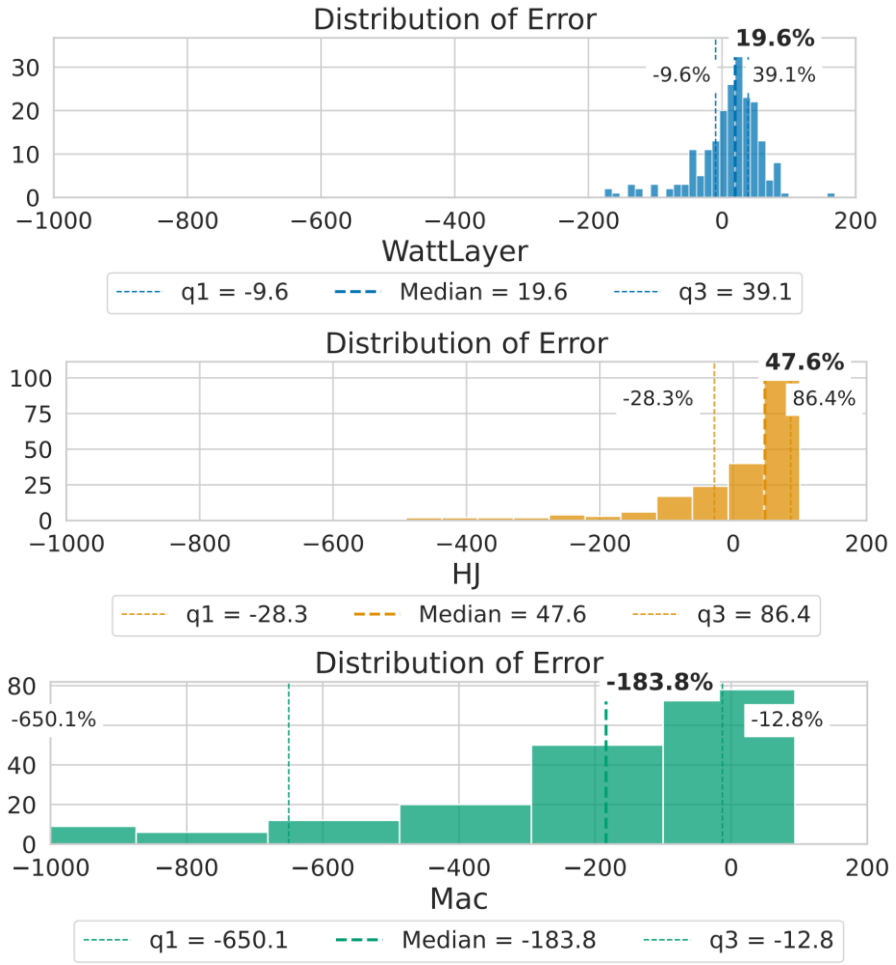
Next Steps

Distribution of error across all architectures

- **Train Datasets**: Vision, NLP and Audio architectures
- **Test Datasets**: one task or balanced between the 3 tasks.
- **Batch sizes:** 1, 32 and 128
- **SOTA 1:** Log-Linear Regression from (Yang & Armour 2025)

  $y = e^{cst} \times MAC^{A} \times Activations^{B} \times Parameters^{C}$
- **SOTA 2:** Linear Regression $y = a \cdot MAC + b$

➔ Our methodology outperforms SOTA models with a **Median Error of 19.6%**



Performance Metrics compared to SOTA models (Task by Task)

- **Train Datasets**: Vision, NLP and Audio architectures
- **Test Datasets**: one task or balanced between the 3 tasks.
- **Batch sizes:** 1, 32 and 128
- **SOTA 1:** Log-Linear Regression from (Yang & Armour 2025)

  $y = e^{cst} \times MAC^{A} \times Activations^{B} \times Parameters^{C}$
- **SOTA 2:** Linear Regression $y = a \cdot MAC + b$

➔ Our methodology outperforms SOTA models for every **task-by-task** evaluation

Vision

METRIC	WATTLAYER	SOTA 1 Vision Datasets	SOTA 1 Balanced Datasets	SOTA 2 Vision Datasets	SOTA 2 Balanced Datasets
MAPE	**40.10%**	63.90%	99.80%	132.60%	833.70%
MedAPE	**33.60%**	48.70%	86.80%	52.90%	229.00%
MaxAPE	**168.90%**	259.20%	870.10%	3073.30%	18784.30%
MinAPE	0.40%	1.90%	**0.02%**	0.70%	8.20%

NLP

METRIC	WATTLAYER	SOTA 1 Balanced Datasets	SOTA 1 Vision Datasets	SOTA 2 Balanced Datasets
MAPE	**45.60%**	152.90%	341.40%	3091.20%
MedAPE	**46.20%**	68.10%	265.40%	2526.30%
MaxAPE	**175.80%**	1510.50%	1158.10%	5713.40%
MinAPE	**4.70%**	26.20%	105.50%	518.60%

Audio

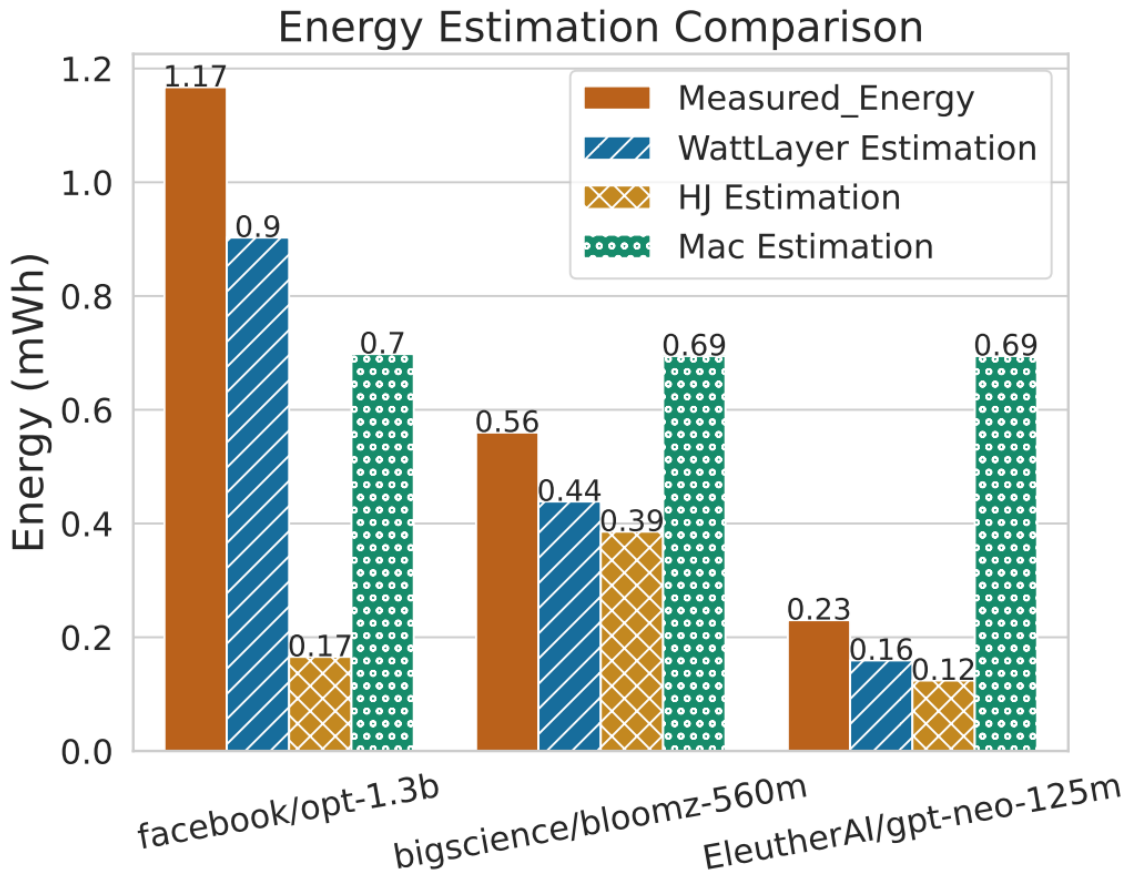
METRIC	WATTLAYER	SOTA 1 Balanced Datasets	SOTA 1 Vision Datasets	SOTA 2 Balanced Datasets
MAPE	**26.60%**	67.70%	74.60%	58.30%
MedAPE	**22.20%**	67.50%	68.60%	53.80%
MaxAPE	**61.10%**	75.10%	100%	93.60%
MinAPE	15.60%	63.40%	64.10%	**1.72%**

Evaluating Zero-Shot Generalization to Large Language Models

- Generalization on task not seen during the training phase without fine tuning

- The three LLMs have respectively
  - 1.3 billion parameters
  - 560 million parameters
  - 125 million parameters

➔ Our model accurately predicts the energy consumption of **unseen tasks**



Introduction

Methodology

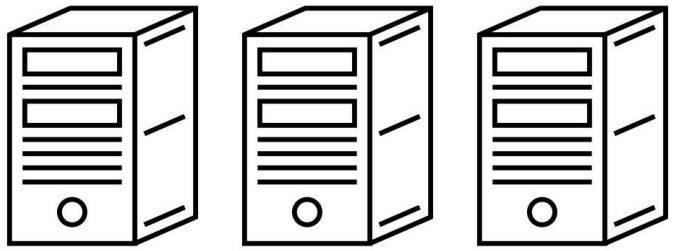
Results

Next Steps

Conclusion and Next Steps

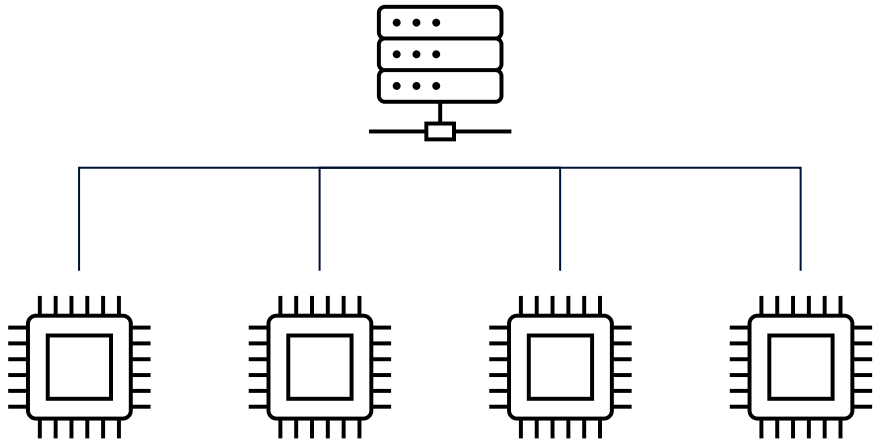
One model for all hardware

- Hardware-specific features to calibrate the model



Distributed inference

- Energy estimation for an execution on several machines
- Cloud/Edge Optimization



Thank you!

Bibliography (I)

Black, S., Leo, G., Wang, P., Leahy, C., and Biderman, S. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL https: //doi.org/10.5281/zenodo.5297715.

Cai, E., Juan, D.-C., Stamoulis, D., and Marculescu, D. NeuralPower: Predict and deploy energy-efficient convolutional neural networks, 2017. URL http://arxiv. org/abs/1710.05420

Courty, B., Schmidt, V., Luccioni, S., Goyal-Kamal, MarionCoutarel, Feld, B., Lecourt, J., LiamConnell, Saboni, A., Inimaz, supatomic, Leval, M., Blanche, L., Cru- ´ veiller, A., ouminasara, Zhao, F., Joshi, A., Bogroff, A., de Lavoreille, H., Laskaris, N., Abati, E., Blank, D., Wang, Z., Catovic, A., Alencon, M., Stechly, M., Bauer, C., de Araujo, L. O. N., JPW, and MinervaBooks. ´ mlco2/codecarbon: v2.4.1, May 2024. URL https: //doi.org/10.5281/zenodo.11171501

De Chateauvieux, B., Pick, E., Ferguson, D., and Sisson, B. Optimize AI/ML workloads for sustainability: Part 3, deployment and monitoring, 2022. URL https://aws.amazon.com/blogs/architecture/optimize-aiml-workloads-for-sustainability-part-3-deployment-andmonitoring/

Desislavov, R., Mart´ınez-Plumed, F., and Hernandez-Orallo, ´ J. Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. Sustainable Computing: Informatics and Systems, 38:100857, 2023. doi: 10.1016/j.suscom.2023.100857.

Getzner, J., Charpentier, B., and Gunnemann, S. Accuracy ¨ is not the only metric that matters: Estimating the energy consumption of deep learning models, 2023. URL http: //arxiv.org/abs/2304.00897.

Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., Walsh, T., Hamrah, A., Santarlasci, L., Lotufo, J. B., Rome, A., Shi, A., and Oak, S. Artificial intelligence index report 2025, 2025. URL https://arxiv.org/abs/2504.07139.

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z.-X., Schoelkopf, H., Tang, X., Radev, D., Aji, A. F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., and Raffel, C. Crosslingual generalization through multitask finetuning, 2022. URL http:// arxiv.org/abs/2211.01786.

Patterson, D., Gonzalez, J., Holzle, U., Le, Q., Liang, ¨ C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., and Dean, J. The carbon footprint of machine learning training will plateau, then shrink, 2022. URL http: //arxiv.org/abs/2204.05149.

TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. https://github. com/pytorch/vision, 2016.

Wightman, R. Pytorch image models. https://github. com/rwightman/pytorch-image-models, 2019.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45, Online, October 2020. doi: 10.18653/v1/2020.emnlp-demos.6.

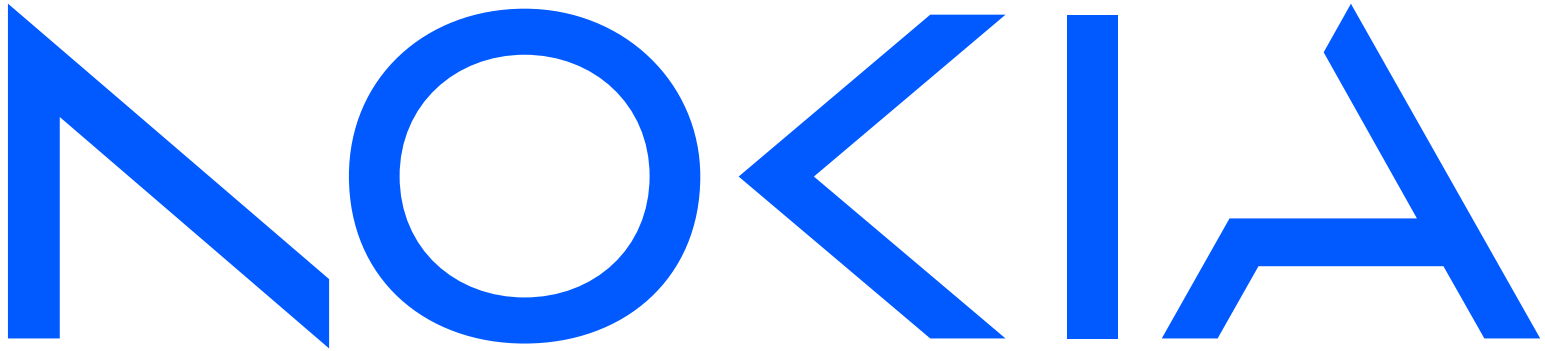
Bibliography (II)

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Behram, F. A., Huang, J., Bai, C., Gschwind, M., Gupta, A., Ott, M., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee, H.-H. S., Akyildiz, B., Balandat, M., Spisak, J., Jain, R., Rabbat, M., and Hazelwood, K. Sustainable AI: Environmental implications, challenges and opportunities, 2022. URL http://arxiv.org/abs/2111.00364.

Yang, Z. and Armour, W. The hidden Joules: Evaluating the energy consumption of vision backbones for progress towards more efficient model inference. In ICML 2025 - 42nd International Conference on Machine Learning, 2025. URL https://bytez.com/docs/ icml/45063/paper.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. OPT: Open pre-trained transformer language models, 2022. URL http://arxiv.org/abs/2205. 01068.

Questions?


NOKIA

Methodology

Experimental Protocol

- Experiments are conducted on NVIDIA GPUs: H100, A100 and GTX TITAN X
- Data collection tool: CodeCarbon
- Number of executions of one process: $N_{mes} \geq 4{,}000$
- Frequency of measurement: $f_{mes} = 10\ Hz$