Offline Energy-Optimal LLM Serving: Workload-Based Energy Models for LLM Inference on Heterogeneous Systems

Grant Wilkins, Srinivasan Keshav, Richard Mortier

Department of Computer Science, University of Cambridge

HotCarbon 2024

Energy Demand of AI: In the News

THE MIT PRESS READER

The Staggering Ecological Impacts of Computation and the Cloud

nature

Generative AI's environmental costs are soaring – and mostly secret

YaleEnvironment360

As Use of A.I. Soars, So Does the Energy and Water It Requires

The Register®

Singapore to offer enterprises incentives to buy greener hardware

Tropical nation ends DC build hiatus and calls for energy optimization everywhere – even software

Laura Dobberstein and Simon Sharwood Fri 31 May 2024 // 01:31 UTC

The New York Times

A New Surge in Power Use Is Threatening U.S. Climate Goals

A boom in data centers and factories is straining electric grids and propping up fossil fuels.

By Brad Plumer and Nadja Popovich March 14, 2024

Our Solution: Workload-Based Models

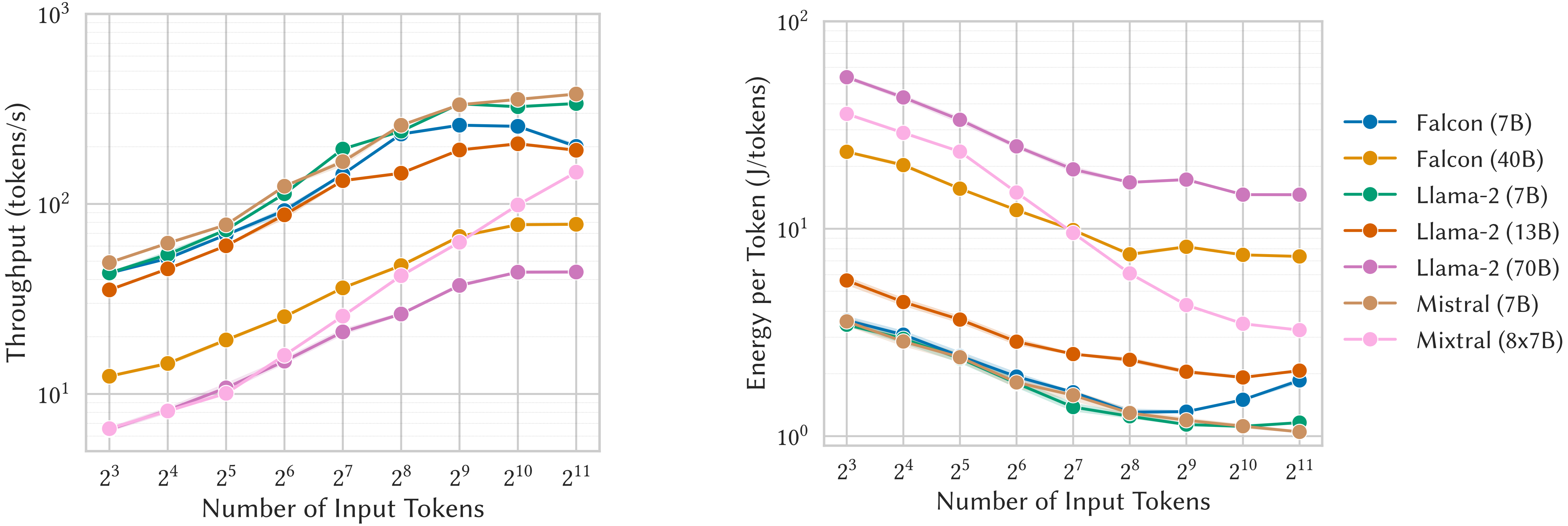
Don’t look in aggregate, look at the node/device-level energy usage.

(Image Credits: NVIDIA)

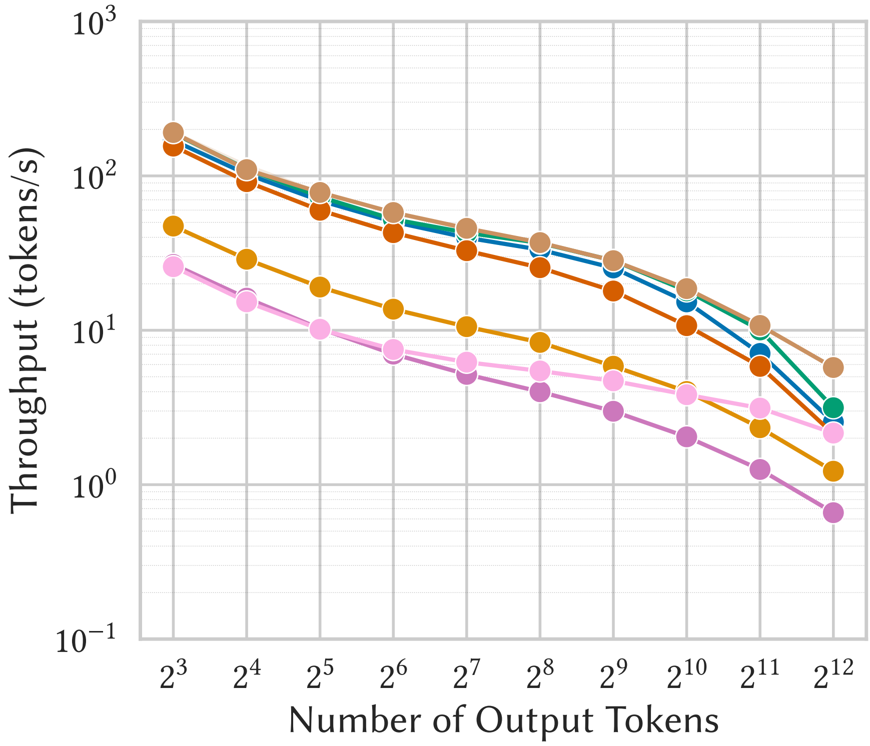
Contributions

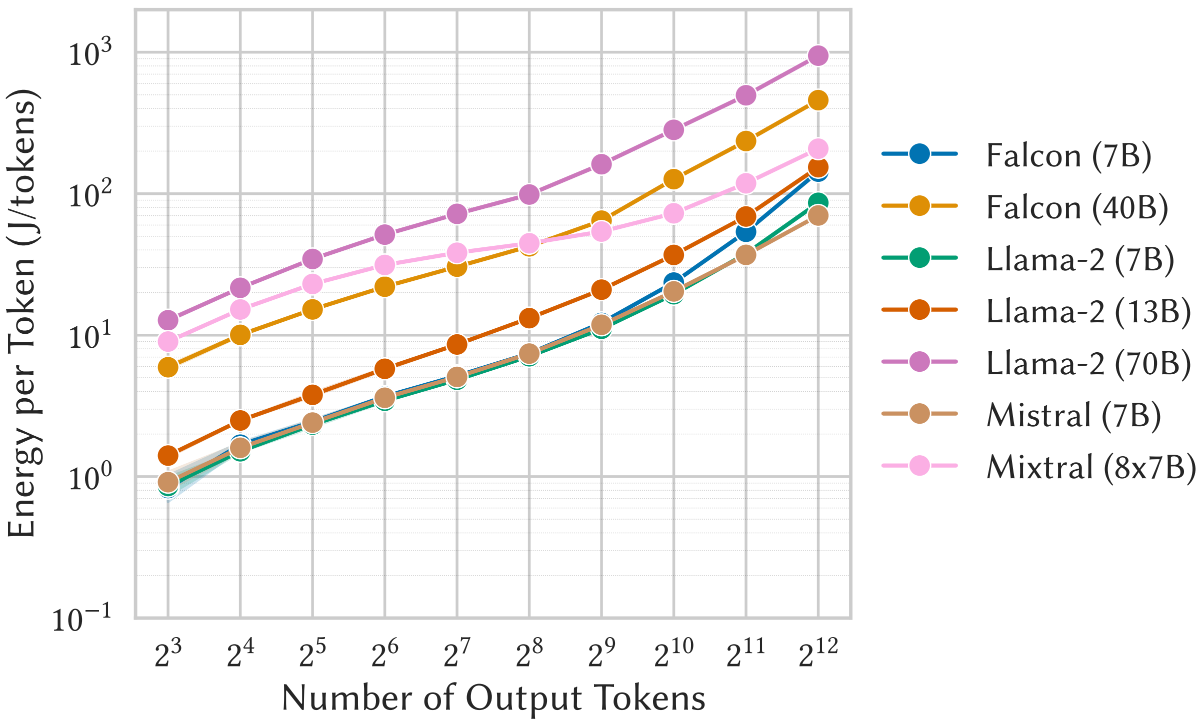
1. Quantified energy usage of multiple open-source LLMs
2. Developed multiple workload-based models for energy consumption / runtime
3. Demonstrated offline routing with operational tuning
4. Open-source energy profiling tools

Energy Efficiency and Throughput: Input Tokens

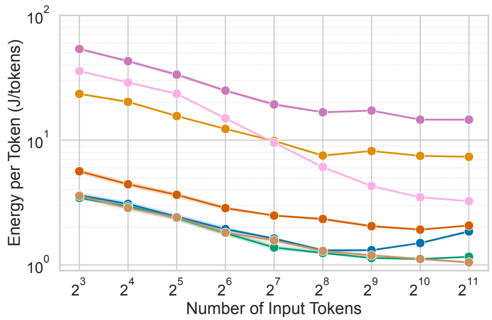


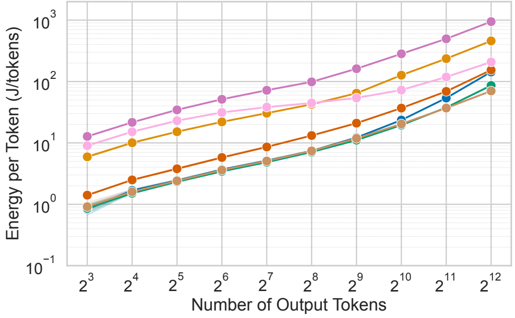
Energy Efficiency and Throughput: Output Tokens





Workload-Based Models





$$e_K(\tau_{in}, \tau_{out}) = \alpha_{K,0}\tau_{in} + \alpha_{K,1}\tau_{out} + \alpha_{K,2}\tau_{in}\tau_{out}$$

LLM (# Params)	Energy Model (e_K)		
	R^2	F-statistic	p-value
Falcon (7B)	0.964	681.2	2.53e-55
Falcon (40B)	0.972	904.5	1.78e-60
Llama-2 (7B)	0.973	942.3	3.76e-61
Llama-2 (13B)	0.972	887.8	3.60e-60
Llama-2 (70B)	0.976	1022	6.66e-62
Mistral (7B)	0.975	997.0	1.70e-61
Mixtral (8x7B)	0.980	1238	4.97e-65

$$\min_{Q_K \in Q} \sum_{K \in \mathcal{K}} \sum_{(\tau_{in}, \tau_{out}) \in Q_K} \zeta \widehat{e_K}(\tau_{in}, \tau_{out}) - (1-\zeta)\widehat{a_K}(\tau_{in}, \tau_{out}) \quad (2)$$

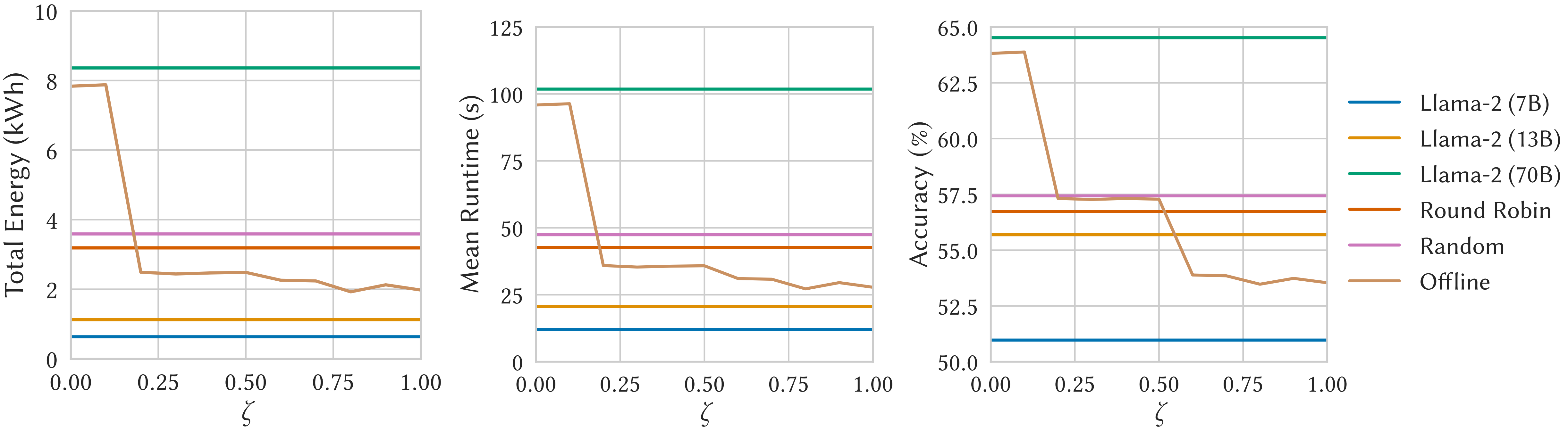
$$\text{s.t., } 0 < \frac{|Q_K|}{|Q|} < 1 \quad (3)$$

$$Q = \bigcup_{K \in \mathcal{K}} Q_K \quad (4)$$

$$Q_I \cap Q_J = \emptyset, I \neq J, \forall I, J \in \mathcal{K}, \quad (5)$$

Data **Models** **Optimization**

Offline Routing



Limitations

1. No optimizations for inference (e.g. DeepSpeed or vLLM)
2. Zeta parameter tuning is not automated
3. No current extension to online setting

Takeaway

We can profile a system and rapidly create accurate models of energy consumption that can be used to make informed scheduling decisions.

Questions: gfw27@cam.ac.uk