Towards Sustainable Large Language Model Serving

Sophia Nguyen*, Beihao Zhou*, Yi Ding, and **Sihang Liu**

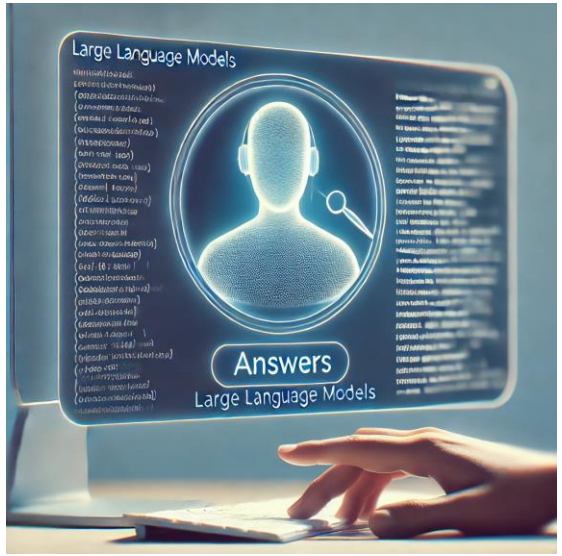
UNIVERSITY OF WATERLOO

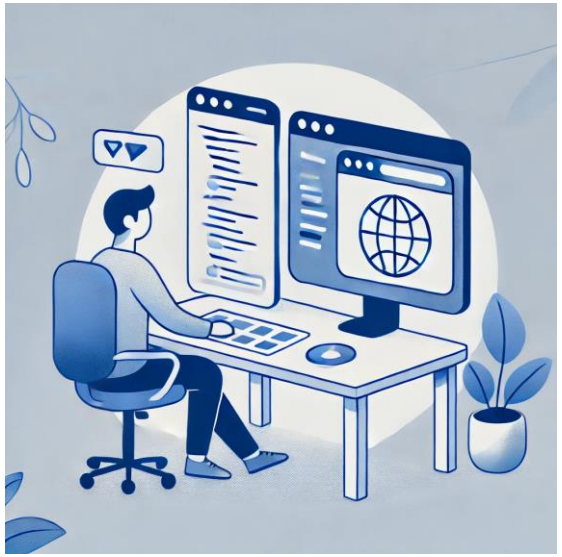
PURDUE UNIVERSITY

* Equal contribution

Large Language Models (LLMs)

- Large language models (LLMs) are widely used

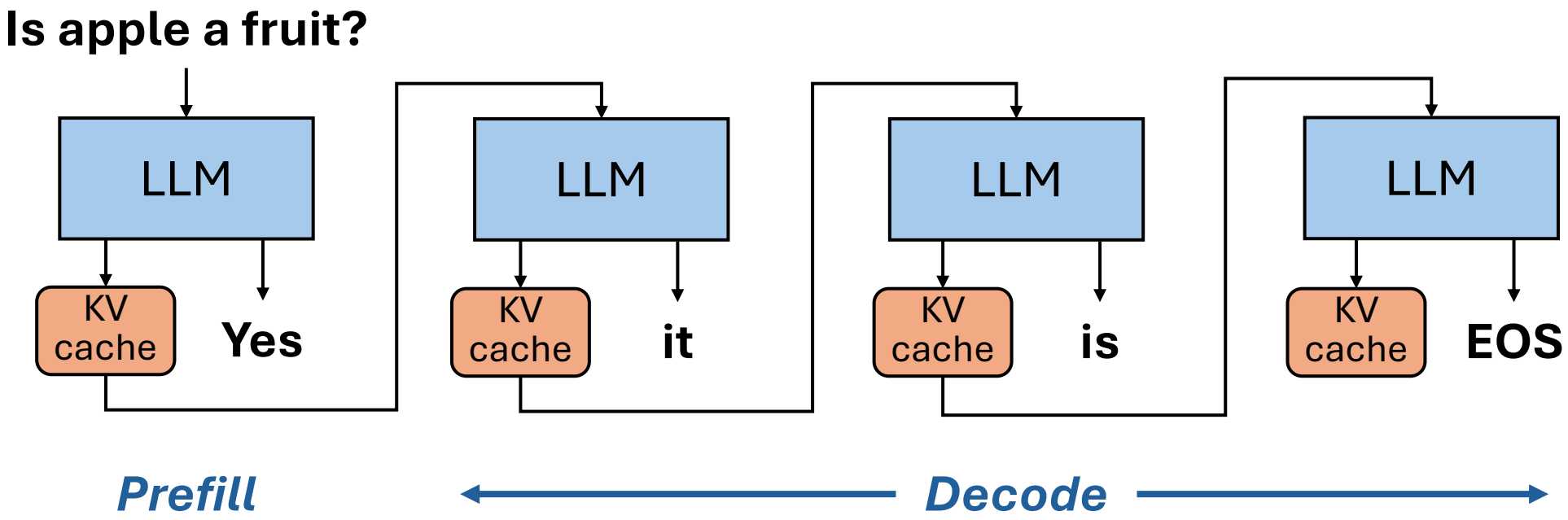




* Images are generated by GPT-4o

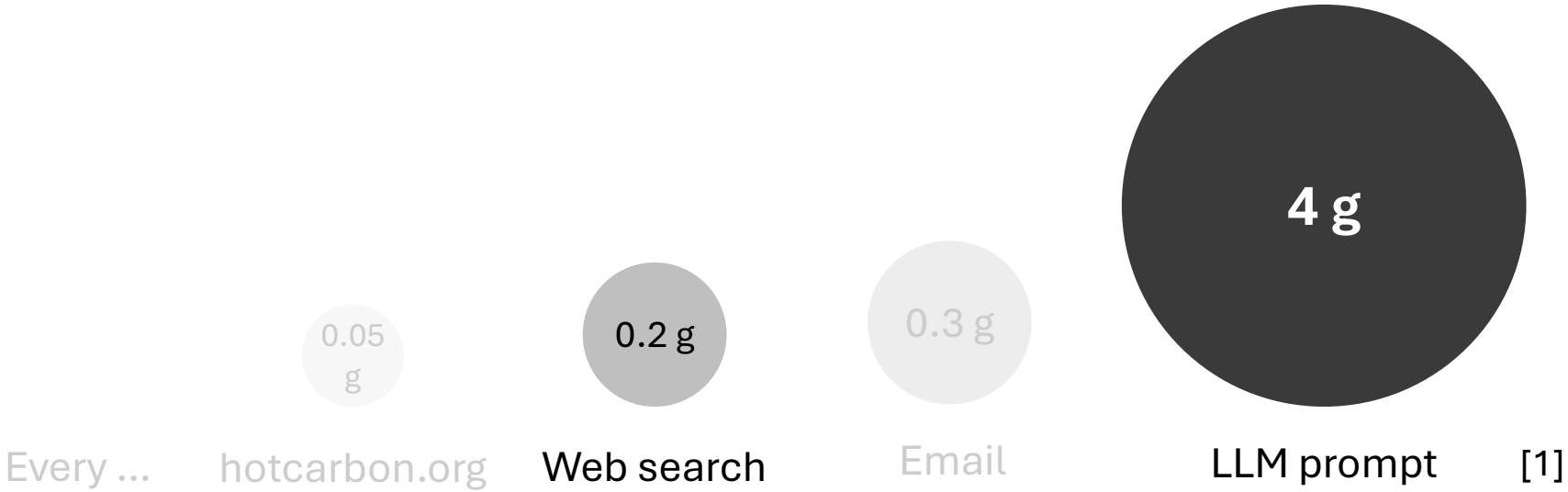
Large Language Models (LLMs)

- Large language models (LLMs) are widely used
- Inference of LLMs follow an auto-regressive pattern



LLM Environmental Impact – Operational

- LLMs are compute-intensive
- Serving LLMs causes **high operational carbon emissions**

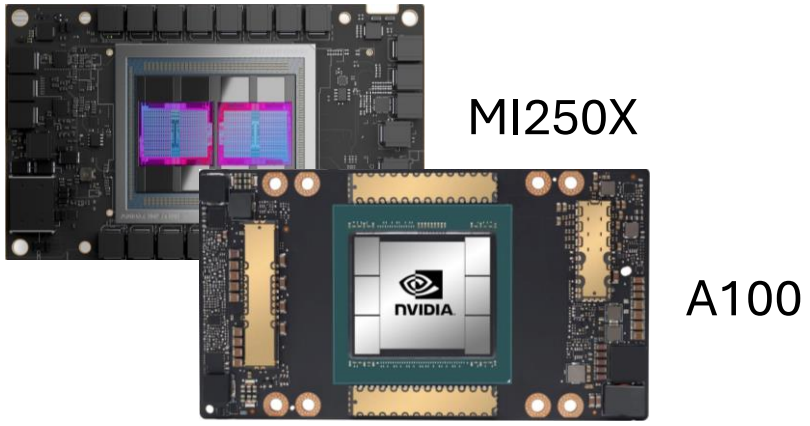


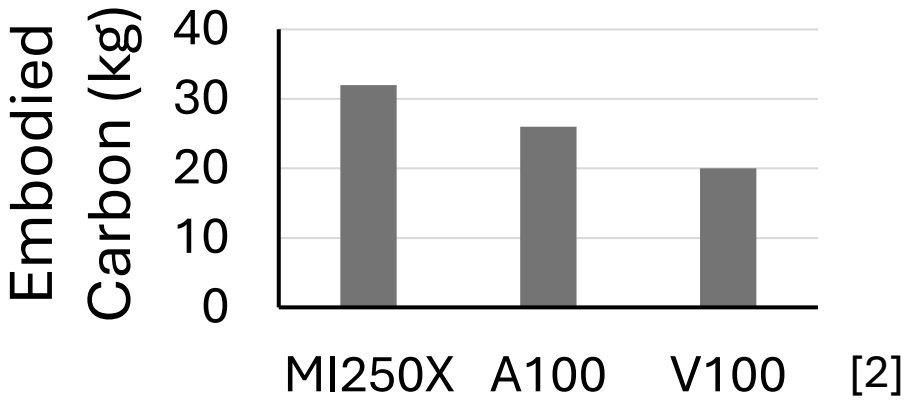
20x more carbon emissions

[1] *Wong. Gen AI's Environmental Ledger: A Closer Look at the Carbon Footprint of ChatGPT. 2024*

LLM Environmental Impact – Embodied

- LLMs require high-end GPUs or ML accelerators
- Manufacturing these devices causes **high embodied carbon emissions**





Serving LLM has serious environmental impact

[2] *Li et al. Toward Sustainable HPC: Carbon Footprint Estimation and Environmental Implications of HPC Systems. SC 2023.*

Overview

- Goal
  Understand the environmental impact of LLM serving by analyzing its **performance** and **carbon emissions**
- Approach
  - Characterize LLMs through **low-level monitoring** and profiling
  - Model both **operational** and **embodied** carbon emissions of LLMs

Carbon Modeling

- Carbon emissions of an LLM prompt C_{prompt} consists of **operational** $C_{prompt,op}$ and **embodied** $C_{prompt,em}$ carbon emissions

Total carbon emission of a prompt: $C_{prompt} = C_{prompt,op} + C_{prompt,em}$

Carbon Modeling – Operational

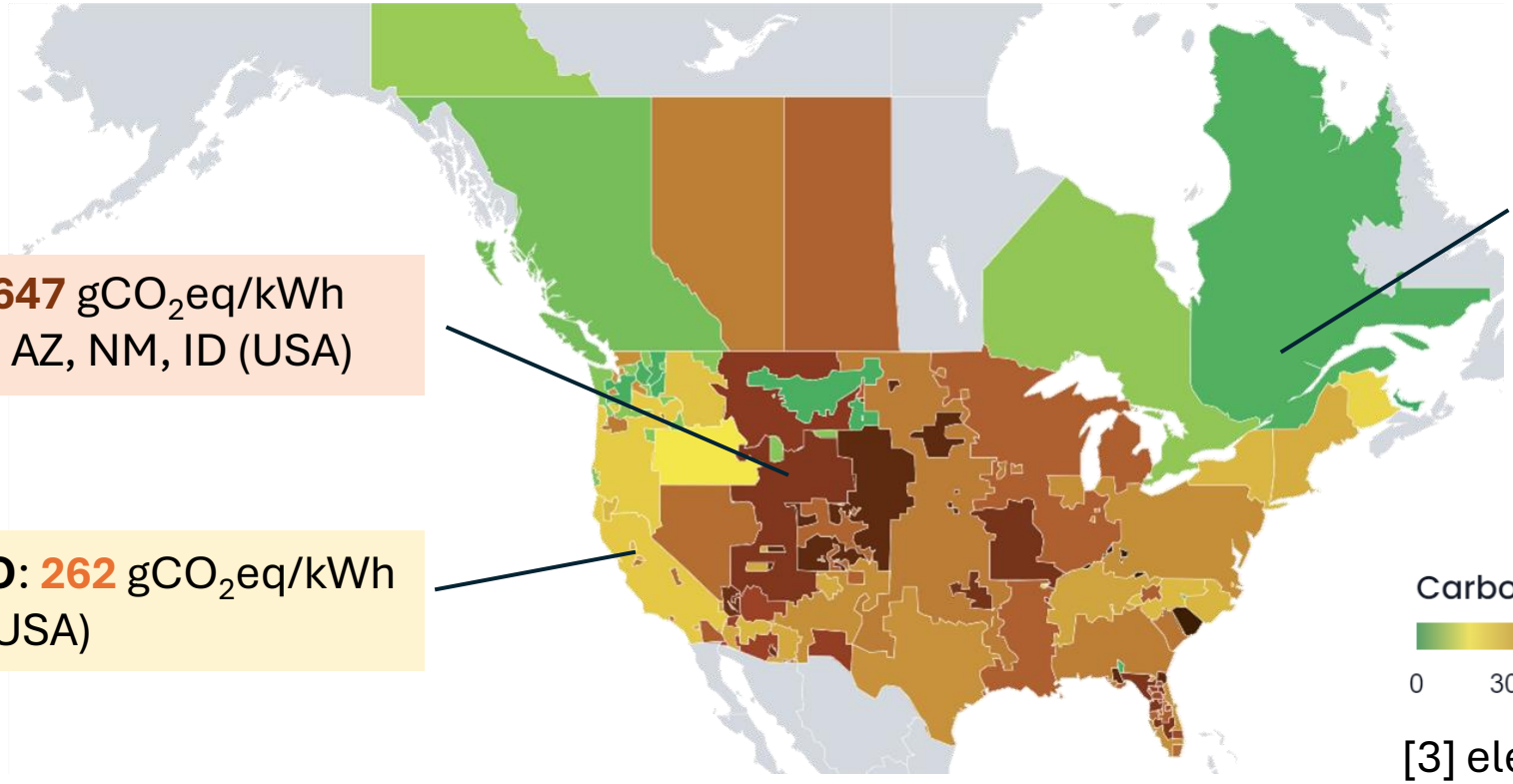
Total carbon emission of a prompt: $C_{prompt} = C_{prompt,op} + C_{prompt,em}$

- **Operational carbon** of a prompt $C_{prompt,op}$ depends on
  - Energy consumption E_{prompt} of the prompt
  - Carbon intensity CI of the area where the GPU is running

$$C_{prompt,op} = E_{prompt} \cdot CI$$

Carbon Intensities in This Study

Carbon intensities (CIs) are based on the average value in 2023



PACE: **647** gCO_2eq/kWh
WY, UT, AZ, NM, ID (USA)

CISO: **262** gCO_2eq/kWh
CA (USA)

QC: **31** gCO_2eq/kWh
Québec (Canada)

Carbon intensity (gCO_2eq/kWh)

0 300 600 900 1200 1500

[3] electricitymaps.com

Carbon intensities differ across regions due to their energy sources

Carbon Modeling – Embodied

Total carbon emission of a prompt: $C_{prompt} = C_{prompt,op} + C_{prompt,em}$

- **Operational carbon** of a prompt $C_{prompt,op}$ depends on
  - Energy consumption E_{prompt} of the prompt
  - Carbon intensity CI of the area where the GPU is running

$$C_{prompt,op} = E_{prompt} \cdot CI$$

- **Embodied carbon** of a prompt depends on [4]
  - Embodied carbon of the GPU C_{em}
  - Prompt execution time t_{prompt}
  - GPU lifetime LT

$$C_{prompt,em} = t_{prompt} / LT \cdot C_{em}$$

[4] Gupta et al. ACT: Designing Sustainable Computer Systems With An Architectural Carbon Modeling Tool. ISCA 2022.

Embodied Carbon Modeling

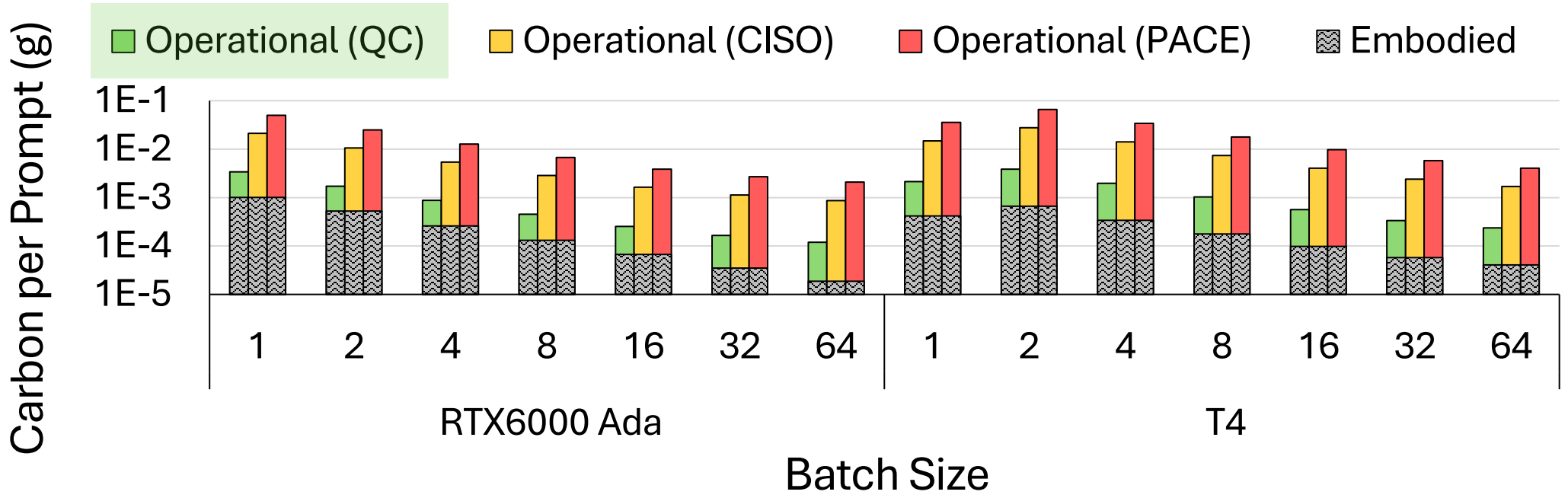
- Model embodied carbon C_{em} based on chip area and memory size [4]

GPUs in this study	**RTX 6000 Ada**	**T4**
Size	608.4 mm^2	545 mm^2
Technology Node	5 nm	12 nm
Memory Capacity	48 GB	16 GB
Thermal Design Power (TDP)	300 W	70 W
Year	2023	2018
Embodied Carbon	**26.6 kg**	**10.3 kg**

[4] Gupta et al. ACT: Designing Sustainable Computer Systems With An Architectural Carbon Modeling Tool. ISCA 2022.

Carbon Emissions in Different Regions

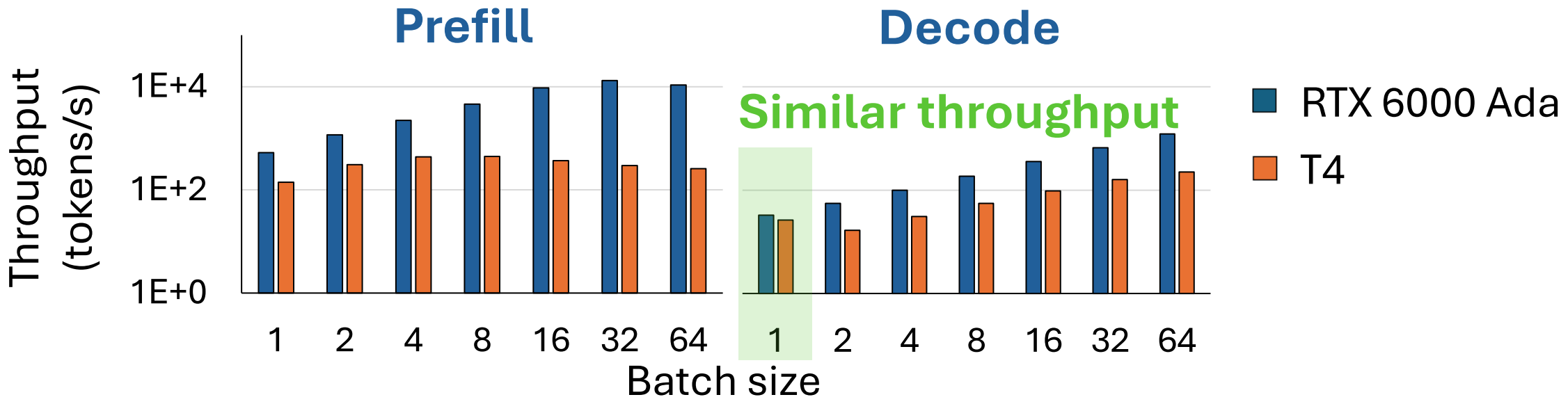
- Evaluate *per-prompt* carbon emissions of 1B-parameter LLaMA with prompts from Alpaca dataset



In regions with lower CI, embodied carbon is more significant, making older GPUs more beneficial

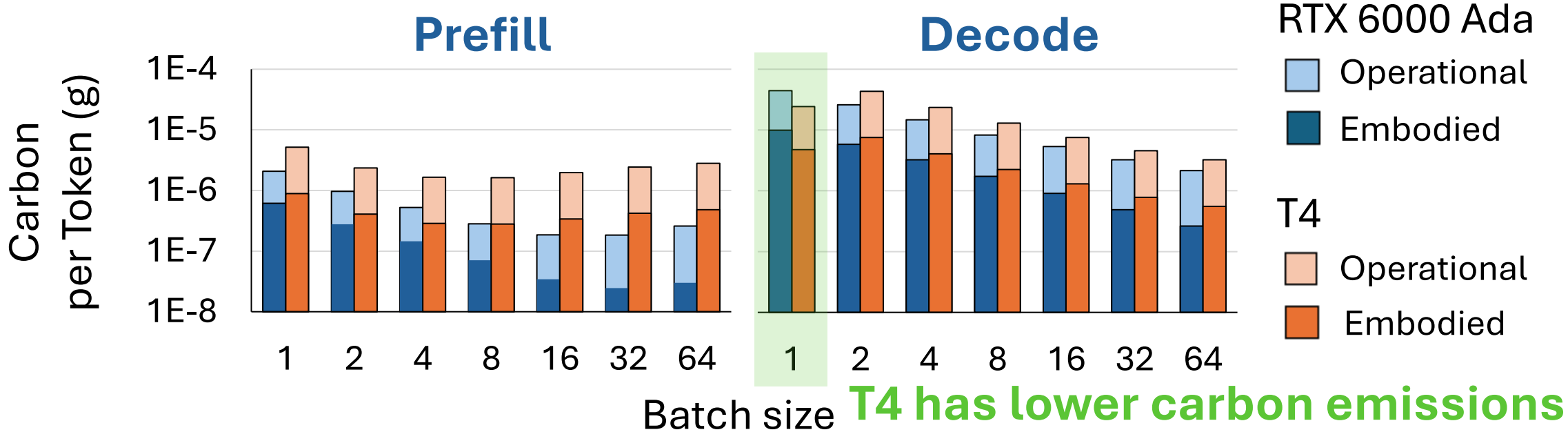
Performance vs. Carbon Emission

- Evaluate **prefill** and **decode** stages of 1B-parameter LLaMA



Performance vs. Carbon Emission

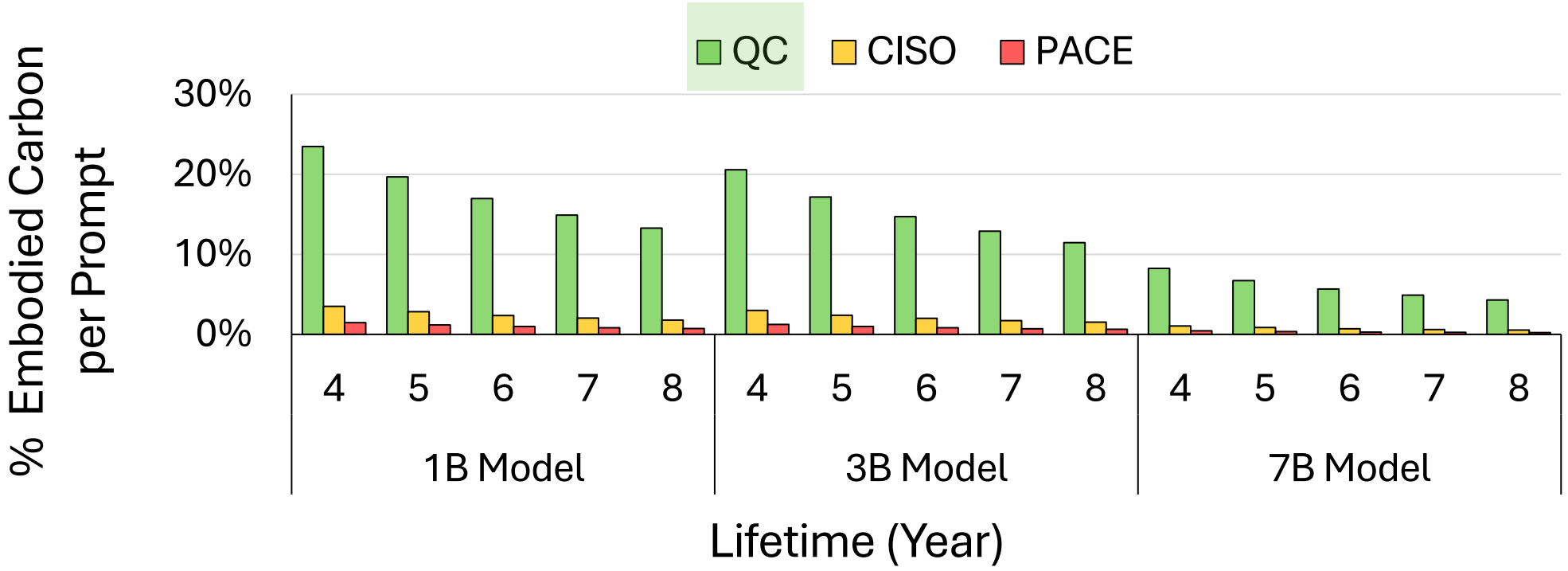
- Evaluate **prefill** and **decode** stages of 1B-parameter LLaMA
- Calculate carbon emission based on carbon intensity of **QC**



RTX 6000 Ada is faster and has lower carbon emissions than T4, except when batch size is 1

Impact of Extending GPU Lifetime

High impact on low-CI regions



Extending GPU lifetime lowers embodied carbon emissions – particularly prominent in regions with lower carbon intensities

Takeaways

- **Region matters:**
  Older GPUs are overall less efficient but more beneficial in regions with lower carbon intensities
- **Workload matters:**
  Old, lower-tier GPUs may have lower carbon emissions in less compute-intensive scenarios
- **Lifetime matters:**
  Exploiting use cases of old, lower-tier GPUs can extend their lifetime, effectively reducing their embodied carbon emissions

Towards Sustainable Large Language Model Serving

Sophia Nguyen*, Beihao Zhou*, Yi Ding, and **Sihang Liu**

UNIVERSITY OF WATERLOO

PURDUE UNIVERSITY

* Equal contribution