
CONNECTING THE DOTS
ISC
High Performance

EESP Workshop 2026 · ISC High Performance · Hamburg, June 26, 2026

CONNECTING THE DOTS
ISC High Performance

SIS

A Sustainability Index for Evaluating

Energy-Efficient LLM Inference

Across CPU and GPU Deployments

Urooj Asgher

School of Informatics and Cybersecurity

HPC NEXUS LABORATORY
TU DUBLIN
IRISH RESEARCH COUNCIL
An Chomhairle um Thaighde in Éirinn

A question I asked myself...

Is LLM Inference Truly Sustainable?

Why LLM Inference Sustainability is Hard to Measure

1,287 MWh

GPT-3 Training Energy

- Training large models is highly energy-intensive, resulting in significant baseline carbon emissions.
- Equivalent to powering 120 homes for a year.

[1-2]

Recurring

Inference vs. Training

- Inference runs millions of times in production; cumulative cost is massive.

[3-4]

Fragmented

How Metrics Are Reported

- Metrics are often analysed independently, making it difficult to assess overall sustainability.

[5-6]

No unified framework exists to compare sustainability across models and hardware platforms

Research Gap & Proposed Solution

Current State of LLM Sustainability Evaluation

Metrics are analysed independently. Energy, accuracy and runtime are never combined into one measure [6].

Incompatible scales, units and measurement conditions make cross-study comparison unreliable [7].

No unified framework exists for cross-hardware CPU and GPU comparison [8].

No reproducible benchmark exists for heterogeneous hardware environments [9].

SIS Framework

To address these challenges, we propose the Sustainability Index Score (SIS): a unified framework for evaluating LLM inference sustainability across models and hardware platforms.

SIS Framework: Design & Methodology (1/5)

SIS is proposed as a unified framework for evaluating the sustainability of LLM inference across heterogeneous hardware platforms.

1

Metrics & SIS Formulation

1.1 Metric calculation
1.2 Accuracy in SIS

2

Normalization & Final SIS Score

3

Models and Datasets

4

Experimental Setup

SIS Framework: Design & Methodology (2/5)

SUSTAINABILITY METRICS

Metric	Unit	Goal
Energy Consumption	*Joules / Query, Task, Epoch*	↓ Lower
Carbon Emissions	*gCO_2eq / Query, Task, Epoch*	↓ Lower
Model Efficiency	*Accuracy ÷ Energy (Acc/J)*	↑ Higher
Token Energy Eff.	*Tokens / J*	↑ Higher
Execution Time	*Seconds / Query or Epoch*	↓ Lower
Hardware Efficiency	*Accuracy ÷ hardware-hours (%/h)*	↑ Higher
Throughput	*Tokens / Sec*	↑ Higher
Reuse / Deployment	*Quantized / Reusable*	✓ Yes preferred
Model Size	*Parameters / GB*	↓ Smaller
Computation Usage	*FLOPs per inference*	↓ Lower
Memory Usage	*GB*	↓ Lower
Carbon Intensity	*g CO_2 / kWh*	~ Context dep.

Energy | Performance | Model | Environment

GOVERNING EQUATIONS

(1) Energy / Query

$$E_{query} = \frac{E_{total}}{N_{queries}}$$

(2) Energy in kWh

$$E_{kWh} = \frac{E_{total}}{3.6 \times 10^6}$$

(3) Carbon Emissions

$$CO_{2_{eq/q}} = \frac{E_{kWh} \times CI}{N_{queries}}$$

(4) Inference Latency

$$T_{inf} = \frac{T_{total}}{N_{queries}}$$

(5) Energy Efficiency

$$\eta_{energy} = \frac{Acc}{E_{total}}$$

(6) Model Size

$$Size_{model} = \frac{Params \times Prec}{8 \times 10^9}$$

(7) FLOPs / Inference

$$FLOPs_{inf} = FLOPs_{layer} \times N_{layers} \times N_{tok}$$

(8) Memory Usage

$$Mem_{usage} = \frac{Bytes_{alloc}}{10^9}$$

(9) Hardware Efficiency

$$\eta_{hw} = \frac{Acc}{T_{hw}}$$

(10) Throughput

$$R = \frac{N_{tokens}}{T_{total}}$$

(11) Token Energy Eff.

$$\eta_{tokenE} = \frac{N_{tokens}}{E_{total}}$$

(12) Accuracy in SIS

$$Acc = \frac{N_{correct}}{N_{total}}$$

SIS Framework: Design & Methodology (3/5)

Normalisation

All quantitative metrics are normalised to [0, 1] before aggregation, since they differ in scale and units.

↓ Lower is better (13)

$$Norm_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Energy · Carbon · Runtime · Model Size · FLOPs · Memory

↑ Higher is better (14)

$$Norm_i = 1 - \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Accuracy · Efficiency · Throughput · Token Energy Eff.

x_i = observed value for metric i
x_{min} = minimum value · x_{max} = maximum value across models

Final SIS Score

The final SIS is computed as a weighted sum of all normalised metrics. Equal weights avoid bias across all metrics.

(15)

$$SIS = \sum_{i=1}^{n} w_i \cdot Norm_i$$

w_i = weight of metric i (equal: $w_i = 1/n$)

$Norm_i$ = normalised value of metric i

n = total number of metrics

SIS Classification Thresholds

SIS Score Range	Sustainability Level
0.0 – 0.3	Low Impact
0.3 – 0.7	Medium Impact
0.7 – 1.0	High Impact

SIS Framework: Design & Methodology (4/5)

Four open-weight, instruction-tuned, decoder-only transformer LLMs evaluated. Selected for strong adoption, optimised availability and suitability for controlled inference. Similarly sized models (7B to 8B) and one smaller model (Phi-3.5-mini) included to examine the impact of model size on sustainability.

LLMs Evaluated

Qwen2.5-7B-Instruct — 7B
Alibaba — Similarly sized (7B to 8B range)

Mistral-7B-Instruct-v0.3 — 7B
Mistral AI — Similarly sized (7B to 8B range)

Llama-3.1-8B-Instruct — 8B
Meta — Similarly sized (7B to 8B range)

Phi-3.5-mini-Instruct — ~3B
Microsoft — Smaller model to examine model size impact on sustainability

Deployed via llama.cpp using GGUF Q4_K_M quantisation, reflecting realistic deployment in resource-constrained environments.

Evaluation Dataset

1,500 Total prompts

500 Per benchmark

GSM8K — 500 samples

Mathematical Reasoning
Tests mathematical reasoning
Evaluation rule: Numerical answer matching

MMLU — 500 samples

Multi-domain Knowledge
Covers multi-domain knowledge
Evaluation rule: Option matching for A B C D choices

Truthful QA — 500 samples

Factual Correctness
Evaluates factual correctness
Evaluation rule: Predefined answer matching

Overall Accuracy — balanced performance measure across all three benchmarks (16)

$$Acc_{overall} = \frac{Acc_{reason} + Acc_{mcq} + Acc_{truth}}{3}$$

SIS Framework: Design & Methodology (5/5)

Hardware and Experimental Setup

HPC Nexus Server Infrastructure

CPUs	2x Intel Xeon Gold 6430
Cores	64 cores 128 threads
GPUs	NVIDIA L40S 48 GB each
Modes	CPU-only and GPU-accelerated

Execution Settings

- Same prompt dataset, decoding parameters, token limits and temperature across all runs
- CPU runs used fixed multi-threading for consistent parallelism
- GPU runs offloaded model layers with all other settings unchanged
- Runtime metrics recorded: execution time, token count and output length
- Energy consumption measured with idle baseline correction applied

Physical Power Meter

Used the physical external power meter to measure energy consumption directly during inference runs.

Specifications

Voltage	230 V at 50 Hz nominal
Accuracy	0.5% voltage 1% current 1.5% power
Storage	Internal 96 GB memory
Connect	USB 2.0 (921600 baud) and Wi-Fi
Export	CSV format for post-processing

Energy Measurement Framework Design

I have designed the Energy Measurement Framework and published ***[10]***. Used in this work to capture fine-grained hardware energy metrics during inference with idle baseline correction, ensuring repeatable and comparable results.

Energy measurement design repository

github.com/urooj88/OmegawattFramework-HPC

Results: Accuracy Analysis (1/5)

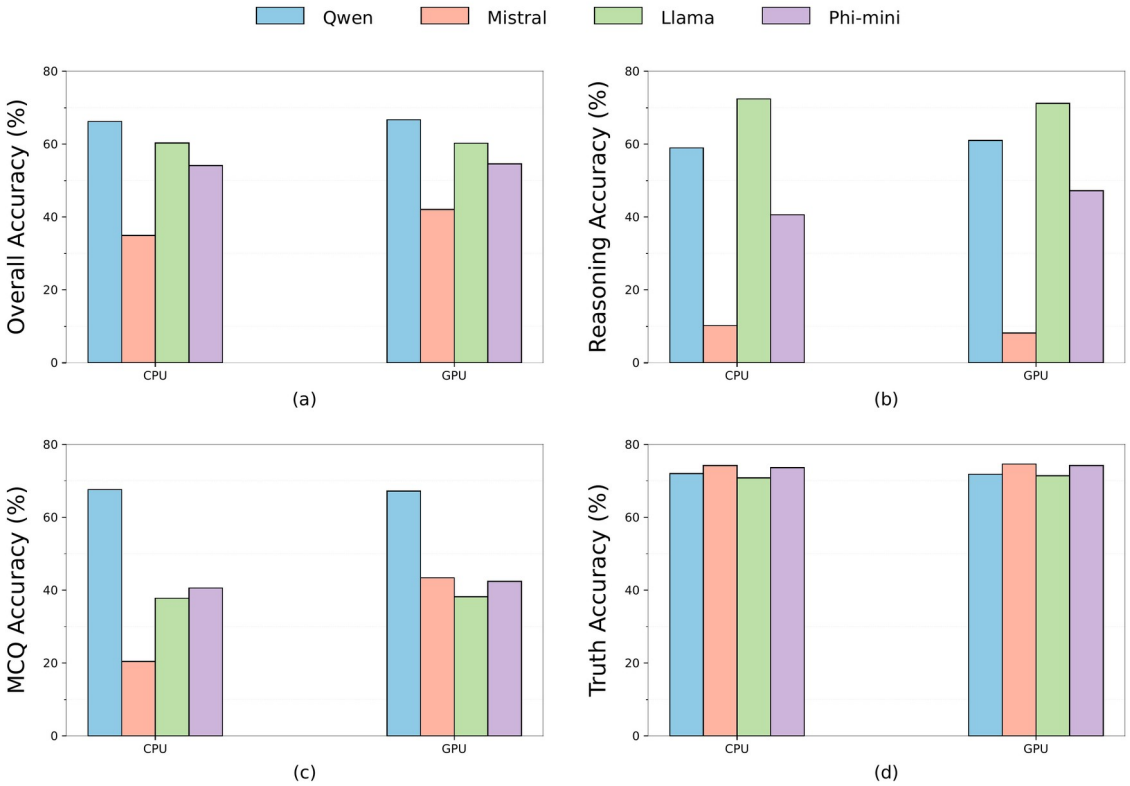

Fig. 1. Comparison of model performance across CPU and GPU in terms of (a) overall accuracy (b) reasoning accuracy (c) MCQ accuracy and (d) truthfulness accuracy.

Qwen2.5 achieves the highest overall accuracy, indicating strong generalisation across reasoning, MCQ, and truth-based tasks.

- LLaMA achieves strong reasoning accuracy and maintains stable performance across both hardware settings.
- Mistral shows lower reasoning accuracy but improves on MCQ tasks under GPU execution.

- Truth accuracy remains relatively consistent across all models with only minor variations between CPU and GPU.
- Factual knowledge lives in model weights, not the hardware.

Model accuracy is task-dependent. No single model uniformly outperforms others across all evaluation categories.

Results: Throughput & Runtime (2/5)

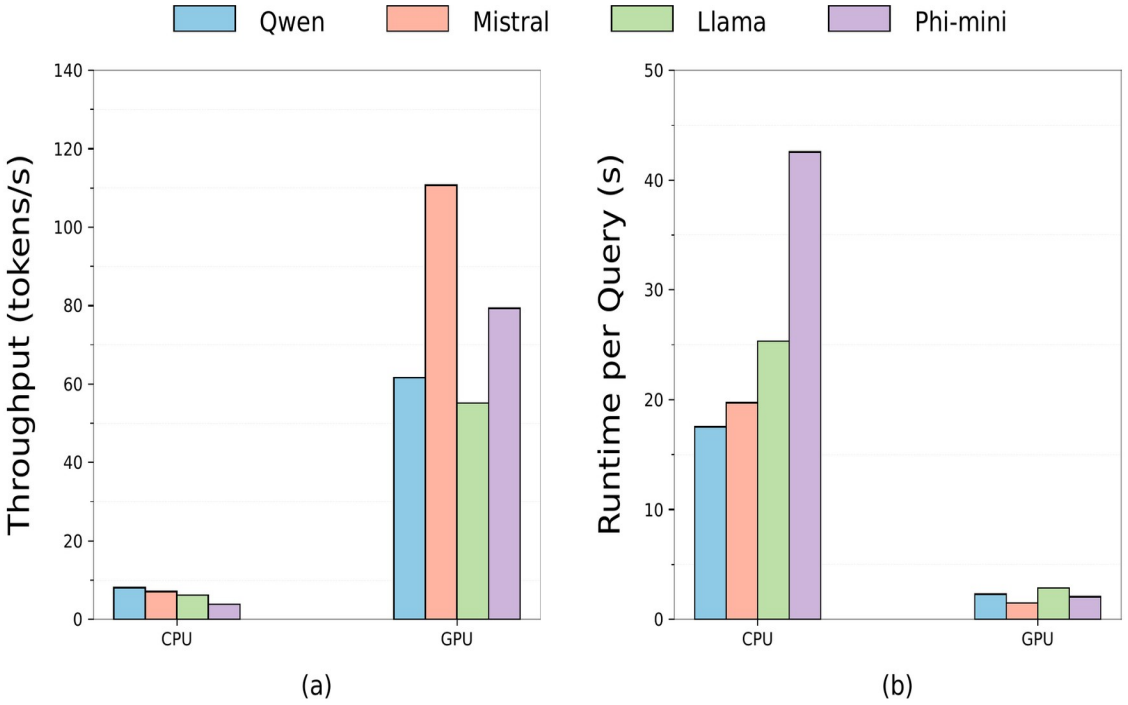

Fig. 2. Performance comparison between CPU and GPU execution for (a) throughput (tokens per second) and (b) runtime per query.

All models show a substantial increase in throughput when executed on GPU.

- Mistral achieves the highest throughput and demonstrates the largest improvement compared to its CPU performance.
- Qwen2.5 maintains consistently low runtime on both CPU and GPU.
- LLaMA shows moderate improvement.

- Phi-mini, which has the highest runtime on CPU, experiences the most noticeable reduction on GPU.
- Both throughput and runtime feed directly into the SIS score as two of the nine components.

Results: Energy, Carbon & Token Efficiency (3/5)

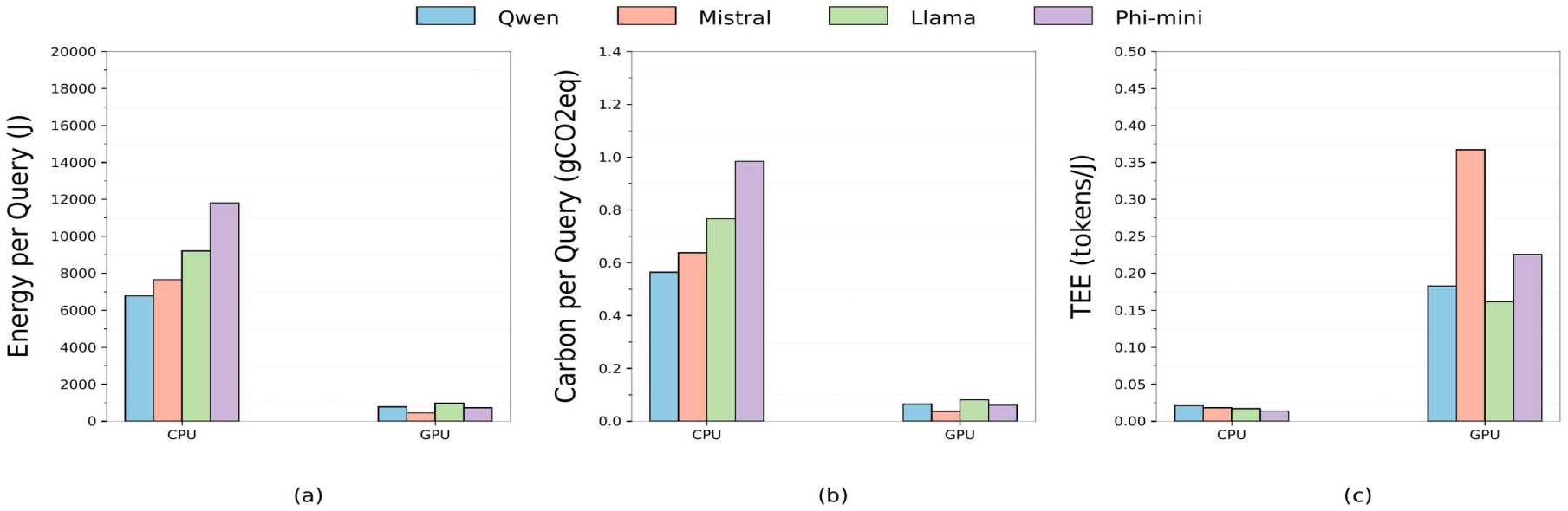

Fig. 3. Sustainability-related metrics across models: (a) energy consumption per query (b) carbon emissions per query and (c) token efficiency (tokens per joule).

- Phi-mini exhibits the most substantial decrease in energy when moving to GPU.
- Qwen2.5 and Mistral also show notable energy reduction; all four LLMs improve without exception.

- Carbon emissions per query follow the same trend as energy consumption across all models.
- proportional because emissions are calculated directly from energy.

Mistral achieves the highest token efficiency, followed by Phi-mini.

Results: Model Complexity & Resource Usage (4/5)

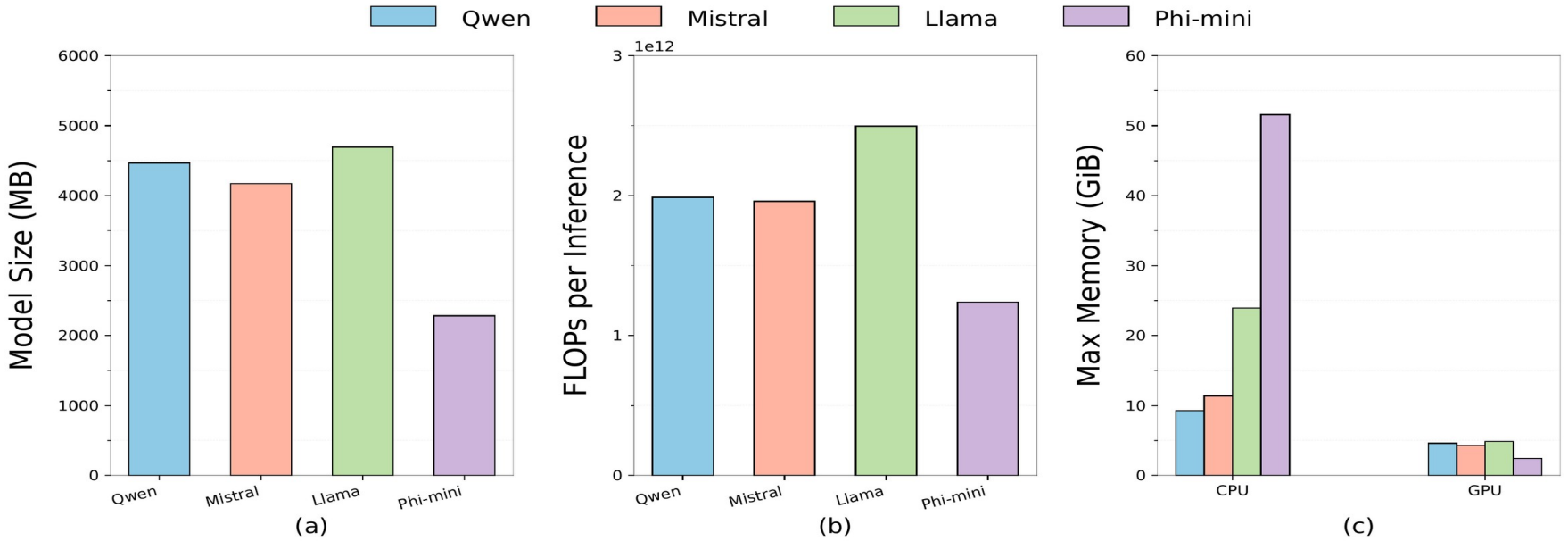

Fig. 4. Resource usage comparison across models, including (a) model size, (b) computational cost (FLOPs per inference), and (c) memory consumption.

LLaMA has the largest size, followed by Qwen2.5 and Mistral, while Phi-mini is significantly smaller.

- LLaMA requires the highest number of operations. Phi-mini has the lowest computational demand.
- Qwen2.5 and Mistral sit in the middle; moderate model size and moderate computational demand.

- Lower memory in GPU runs reflects redistribution from RAM to VRAM, not a reduction in total memory.
- Model size, FLOPs, and memory are three of the nine SIS components; smaller and lower means better sustainability score.

The SIS Score & Key Finding (5/5)

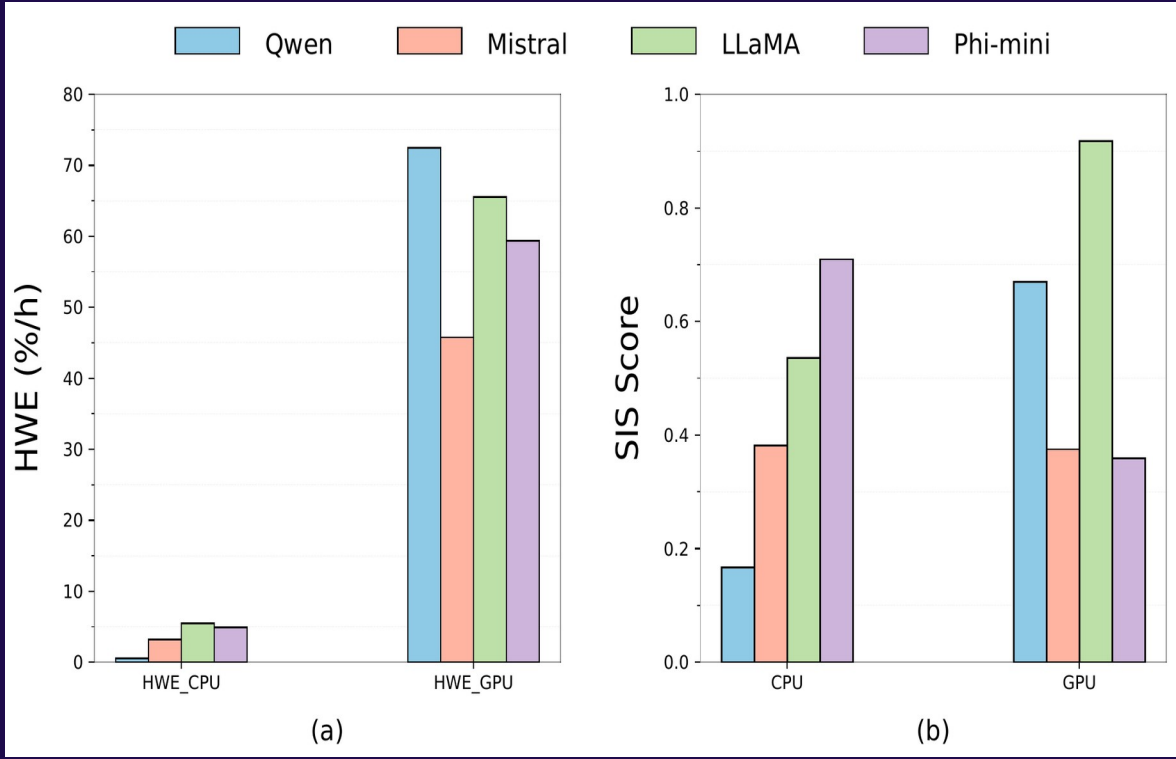

Fig. 5. (a) hardware efficiency (HWE) and (b) SIS score.

Best SIS on GPU

Phi-mini

Key Insights

- Phi-mini achieves the lowest SIS under GPU execution, followed by Mistral.
- Qwen2.5 attains the highest hardware efficiency but does NOT achieve the best SIS score.
- Overall sustainability is governed by the combined effect of all metrics, not any single dimension.

Future Work

Distributed HPC Inference

Extension of SIS to distributed inference in heterogeneous HPC environments with multiple nodes.

Communication & Partitioning

Analysing communication overhead, workload partitioning, and hardware heterogeneity on energy and performance.

Energy-Efficient LLM Infrastructure

How HPC strategies can support energy-efficient LLM infrastructure amid growing HPC–LLM convergence.

References

[1] L. Mei and M. Stamp, “Energy considerations for large pretrained neural networks,” arXiv preprint arXiv:2506.01311, 2025.

[2] B. Tomlinson, R. W. Black, D. J. Patterson, and A. W. Torrance, “The carbon emissions of writing and illustrating are lower for AI than for humans,” Scientific Reports, vol. 14, no. 1, p. 3732, 2024. doi: 10.1038/s41598-024-54271-x.

[3] S. A. Khowaja, P. Khuwaja, K. Dev, W. Wang, and L. Nkenyereye, “ChatGPT needs SPADE (Sustainability, Privacy, Digital Divide, and Ethics) evaluation: A review,” Cognitive Computation, vol. 16, no. 5, pp. 2528–2550, 2024.

[4] J. Fernandez, C. Na, V. Tiwari, Y. Bisk, S. Luccioni, and E. Strubell, “Energy considerations of large language model inference and efficiency optimizations,” in Proceedings of the 63rd Annual Meeting of the Assoc iation for Computational Lin-guistics (Volume 1: Long Papers), pp. 32556–32569, 2025.

[5] Y. Jin, G.-Y. Wei, and D. Brooks, “The Energy Cost of Reasoning: Analyzing Energy Usage in LLMs with Test-time Compute,” arXiv preprint arXiv:2505.14733, 2025.

[6] J. Haase, F. Klessascheck, J. Mendling, and S. Pokutta, “Sustainability via LLM Right-sizing,” arXiv preprint arXiv:2504.13217, 2025.

[7] R. Rubei, A. Moussaid, C. Di Sipio, and D. Di Ruscio, “Prompt engineering and its implications on the energy consumption of Large Language Models,” in 2025 IEEE/ACM 9th International Workshop on Green and Sustainable Software (GREENS), 2025, pp. 60–67. doi: 10.1109/GREENS66463.2025.00014.

[8] S. Samsi, D. Zhao, J. McDonald, B. Li, A. Michaleas, M. Jones, W. Bergeron, J.Kepner, D. Tiwari, and V. Gadepally, “From words to watts: Benchmarking the energy costs of large language model inference,” in 2023 IEEE High Performance Extreme Computing Conference (HPEC), pp. 1–9, 2023

[9] M. F. Argerich and M. Patiño-Martínez, “Measuring and improving the energy efficiency of large language models inference,” IEEE Access, vol. 12, pp. 80194–80207, 2024.

[10] U. Asgher et al., "Evaluating Hardware and Software Power Measurement Tools: Assessing Accuracy in Measuring Application Energy Consumption for Data-Parallel Workloads," Springer LNCS, DOI: 10.1007/978-3-031-95652-2_39, June 2025.

[11]

Thank You

Questions

B00168170@mytudublin.ie

Technological University Dublin

***Acknowledgment:** *This research has been conducted with the support of the Irish Research Council under Grant No. GOIPG/2022/147. All experiments have been conducted using the HPC-Nexus Testbed at Technological University Dublin, Ireland.*

ENJOYED THIS WORKSHOP?

Don’t forget to rate it!

Thank you!

CONNECTING THE DOTS ↘