Measuring AI Work in Malaysian Data Centres

An open method for verifying AI use and measuring output per allocated megawatt

Croutons Research, NRLC.ai · Version 0.1, draft for comment · 9 November 2026 ·
Licence: CC BY 4.0

Summary

Malaysia now approves only AI data centres and rations the power and water they receive. Both policies depend on facts that today's measurements cannot show. Building meters and PUE report how much energy a facility uses and how efficiently it delivers that energy to IT equipment. They cannot show whether the equipment is doing AI work, or how much useful work each megawatt and cubic metre produces.

This paper proposes an open, privacy-preserving method to measure both. It defines five metrics, built on counters that AI accelerators already report and on request metadata that AI services already log:

1. **AI workload share:** the share of a facility's IT energy used by AI training and inference.
2. **Accelerator utilisation:** how much of the installed AI hardware is doing work rather than idling.
3. **Answers per megawatt-hour:** useful AI output per unit of facility energy.
4. **Energy, water and carbon per answer,** in the units of the MITI sustainable data-centre guideline.
5. **Allocation utilisation:** useful output against the power and water a facility has been allocated.

The method never reads prompts, responses or customer data. It aligns with ISO/IEC 30134, the MITI guideline and the EECA 2024 reporting format, so operators can use one set of measurements for incentive applications, energy-manager reports and sustainability disclosure.

It is offered free to Malaysian agencies, standards bodies, universities and operators. We propose a first pilot on a public university AI cluster, with the university as independent verifier.

---

1. Two questions Malaysia now has to answer

Is an approved "AI data centre" doing AI work? On 24 February 2026 the government stopped accepting new non-AI data-centre applications. No written definition of an AI data centre had been published as of November 2026, and nothing in current approvals checks how a facility is used once it operates. A utility meter sees megawatts; it cannot tell AI inference from storage, general cloud hosting or cryptocurrency mining. Only measurement at the level of the workload can.

Is allocated power and water producing useful work? Both are now rationed:

- Peninsular data centres drew 1,374 MW of the 2,346 MW they had contracted (58.6%), and the government has said operators will have to pay for unused capacity.
- In 2024 Johor granted data centres 21% of the water they requested, and Selangor 28%.
- The Digital Infrastructure (Data Centre) Bill, due at Cabinet in 2027, is expected to require staggered 10-year load plans and site-by-site water assessment.

Allocating a scarce resource well needs a measure of what each unit produces. PUE, WUE and CUE describe how efficiently a building supports its IT equipment. None describes the output of that equipment.

What this method adds: a common, verifiable way to answer both questions, using data that AI facilities already generate, without exposing anything their customers send.

---

2. Principles

1. **Privacy by design.** The method uses hardware energy counters and request metadata only: timestamps, token counts, completion status and workload labels. It never reads prompts, responses, model weights or customer data. Raw data stays with the operator; only aggregated metrics leave the site.
2. **Standards-aligned.** Facility and IT energy follow ISO/IEC 30134-2 (PUE). Water and carbon follow ISO/IEC 30134-9 (WUE) and 30134-8 (CUE), as the MITI guideline requires. Reporting periods match EECA 2024 (12 consecutive months).
3. **Reproducible.** Every metric can be recomputed from logged data by a third party, such as a Registered Energy Manager, a Registered Energy Auditor or a university.
4. **Vendor-neutral.** Any accelerator that reports its own energy can be measured: NVIDIA GPUs through NVML and DCGM counters, and equivalents on other hardware. No proprietary tool is required.
5. **Open.** The method is published under CC BY 4.0. Thresholds, such as the AI workload share at which a facility counts as an AI data centre, are policy decisions for government; the method only measures.

---

3. Metrics and definitions

Terms used throughout:

- **Facility energy (E_fac)** and **IT energy (E_IT)**: as defined in ISO/IEC 30134-2, over the reporting period. PUE = E_fac / E_IT.
- **Accelerator energy (E_acc)**: energy reported by the accelerators' own counters (for NVIDIA hardware, the NVML or DCGM total-energy counter), summed over the period.
- **AI workload**: a job the operator labels as model training, fine-tuning or inference serving, through its scheduler or serving platform.
- **Answer**: one inference request that completed and returned output. Failed, cancelled and rejected requests are counted separately and are not answers.

- **Tokens**: input tokens processed and output tokens generated, as logged by the serving software. Reported alongside answers, because answers vary in size.

#	Metric	Definition	Unit	What it answers
1	**AI workload share (AIWS)**	IT energy used by AI workloads ÷ total IT energy. AI IT energy = metered energy of racks running AI workloads, split by each workload's share of accelerator energy where racks are shared	%	Is this facility doing AI work, and how much of it?
2	**Accelerator utilisation**	Accelerator-hours actively running AI work ÷ installed accelerator-hours; plus idle accelerator energy as a share of accelerator energy	%	Is the AI hardware working or waiting?
3	**Answers per megawatt-hour (APM)**	Answers ÷ facility energy attributable to AI (AI IT energy × PUE), in MWh. Also reported as tokens per MWh; training reported as training tokens per MWh	answers/MWh	How much useful AI output does each unit of energy produce?
4a	**Energy per answer**	Facility energy attributable to inference ÷ answers	Wh/answer	The basic efficiency figure
4b	**Water per answer**	Inference IT energy per answer (kWh) × site WUE (L/kWh, ISO/IEC 30134-9; numerically equal to MITI's m^3/MWh)	mL/answer	Cooling water behind each answer
4c	**Carbon per answer**	Energy per answer × grid emission factor. Location-based uses the Energy Commission factor for the region (2024 provisional: Peninsular 0.740, Sabah 0.539, Sarawak 0.199 kg CO2e/kWh); market-based uses contracted renewable supply where evidenced	g CO2e/answer	Carbon behind each answer, comparable to CUE
5	**Allocation utilisation**	Average power drawn ÷ contracted power; annual answers ÷ contracted MW; annual answers ÷ allocated water (m^3)	%, answers/MW, answers/m^3	Is allocated power and water producing output?

How the metrics fit together: AI workload share answers the first policy question. Allocation utilisation and answers per megawatt-hour answer the second. Energy, water and carbon per answer connect both to the MITI guideline's PUE, WUE and CUE, so nothing new replaces existing reporting; it extends it from the building to the work.

4. Measurement protocol

Data sources (all already present in a typical AI facility):

Source	What is read	Used for
Facility and IT meters (ISO/IEC 30134-2 measurement points)	Total facility energy; IT energy at UPS or PDU output	PUE; facility energy attributable to AI

Source	What is read	Used for
Rack or PDU meters	Energy per rack running accelerator nodes	AI IT energy, including servers, networking and storage in those racks
Accelerator counters	Cumulative energy, activity state, per device	Accelerator energy, utilisation, idle energy
Scheduler or serving logs	Per job or request: start and end time, workload label, tokens in and out, completion status	Answers, tokens, workload classification
Utility contract and water records	Contracted MW; water allocation and metered consumption; site WUE	Allocation utilisation; water per answer

Steps:

1. **Classify workloads.** Each job or serving endpoint is labelled AI or non-AI from scheduler metadata. Labels are checked against hardware behaviour: an AI label on hardware that shows no accelerator activity is flagged.
2. **Measure energy at three levels.** Facility, AI racks and accelerators, over the same intervals. Cumulative counters are read at least once a minute and at each request boundary where the serving software allows.
3. **Attribute energy to work.** Within each interval, accelerator energy is split across the requests or jobs running, in proportion to the work each did (tokens processed and generated, weighted by the time each was active). Energy while no work runs is recorded as idle energy, not attributed to answers. In our benchmark, an energy model of this form (idle draw, active time and tokens processed) explained 99.8% of the measured variation in accelerator energy.
4. **Scale to the facility.** Rack energy not captured by accelerator counters (CPUs, memory, networking) is apportioned in the same proportions; facility overhead is applied through measured PUE for the same period.
5. **Reconcile.** Attributed energy plus idle energy must equal measured energy at each level, within the stated meter accuracy. Any remainder is reported, not hidden.
6. **State uncertainty.** Each figure is reported with the accuracy class of the meters and counters used, and the share of energy that could not be attributed.

Reporting periods: monthly figures, plus a rolling 12-month figure to match EECA 2024.
A facility new to the method reports a 3-month baseline first.

5. Reporting template and how it maps to existing rules

One-page facility statement (per reporting period):

- Facility, operator, location, reporting period, contracted MW, allocated water
- Facility energy, IT energy, PUE, WUE, CUE (as already declared under the MITI guideline)
- AI workload share, with the split between training, fine-tuning and inference

- Accelerator utilisation and idle accelerator energy
- Answers, tokens, answers per MWh
- Energy, water and carbon per answer (location-based and market-based carbon)
- Allocation utilisation for power and water
- Attribution coverage, meter accuracy classes, and who verified the figures

Mapping to existing obligations:

Rule	What it already asks	What this method supplies
MITI sustainable data-centre guideline (DESAC applicants)	Declared PUE, CUE and WUE under ISO/IEC 30134; actual WUE disclosed yearly	Measured operating values, split by workload and expressed per answer
EECA 2024 annual report (Registered Energy Manager)	12 months of energy data, variable data, energy intensity, and savings measures in GJ and RM	Answers or tokens as the production variable; energy intensity as GJ per million answers; verified savings from efficiency measures
EECA energy audit (Registered Energy Auditor)	Audit of energy use and saving opportunities every 5 years	Workload-level evidence of where AI energy is used and wasted
GBI Data Centre Tool 2.0	Performance-based PUE, WUE, carbon and renewables	Operational evidence for those criteria (credit treatment to be confirmed with GBI)
Bursa / Securities Commission sustainability reporting (IFRS S2)	Scope 1 and 2 emissions; Scope 3 after transition relief; assurance from 2027 for the largest issuers	Auditable kWh and tCO2e, and per-service figures that tenants need for their own Scope 3
Digital Infrastructure (Data Centre) Bill (2027)	Not yet published; reported to include load plans and site water assessment	A ready, tested way to verify AI use and output per allocated MW and m^3

The method adds one production variable (answers or tokens) to measurements operators already make. It does not ask any facility to report twice.

6. Illustrative reference values

To show the scale of the numbers, here are values from Croutons Research's controlled benchmark of an open-weight model, scaled to a reference 20 MW AI hall at the MITI guideline's PUE of 1.4 and WUE of 2.2, on the Peninsular grid. They illustrate the method; they are not measurements of any Malaysian facility.

Metric	Reference value
Energy per answer (facility)	about 0.17 Wh
Answers per MWh	about 5.9 million
Water per answer	about 0.27 mL
Carbon per answer (location-based, 0.740 kg CO2e/kWh)	about 0.13 g, or 126 g per 1,000 answers

For comparison, Google has reported 0.24 Wh, 0.03 g CO2e and 0.26 mL of water for its median Gemini text prompt, measured across its own fleet on a lower-carbon grid.

What the benchmark shows about variation, which is what makes the metrics useful for policy:

- Energy per answer at low load was about 3.3 times higher than at high load. A hall that is allocated power but runs half-empty produces far fewer answers per MWh.
- One workload type (long summaries) made up 7% of requests but 37% of energy. Output per MWh depends heavily on what a facility runs, not only on how it is built.
- Between 5% and 16% of energy went to work that produced no new output: repeated requests, retries and documents placed where cached work could not be reused.

None of these differences shows up in PUE.

---

7. Governance, verification and a proposed pilot

Who measures and who verifies. Operators measure, using their own meters, counters and logs. Figures are verified by a party with no commercial interest in the result: a Registered Energy Manager or Registered Energy Auditor under EECA 2024, or a university. Croutons Research supplies the method and support free of charge, with a reference implementation available on request, and does not act as verifier.

Privacy. Only aggregated metrics leave the facility. Request-level metadata is processed on site and can be discarded after aggregation. No prompts, responses, model weights or customer identities are collected.

Openness. The method is published under CC BY 4.0 and versioned. Changes are made in the open, with comment from agencies, operators, universities and standards bodies. We would welcome its development through the Malaysian Technical Standards Forum's Green ICT, Environment & Climate Change Working Group, as an extension of the Specification for Green Data Centres (MCMC MTSFB TC G004:2024).

Proposed pilot. A first measurement on a public-university AI cluster in Malaysia:

1. Weeks 1–2: agree scope, data sources and privacy terms; install counter and log collection.
2. Weeks 3–6: measure all five metrics across normal research and teaching workloads.
3. Weeks 7–8: the university verifies the figures and publishes a first verified AI facility statement, with lessons for the method.

The pilot costs the host nothing beyond staff time and gives Malaysia a public, independently verified reference point before the data-centre bill is drafted.

8. Limitations and open questions

1. **Thresholds are policy, not measurement.** The method reports AI workload share; deciding what share qualifies a facility as an AI data centre is for government.
2. **Training has no "answer".** Training and fine-tuning are reported as energy, accelerator-hours and tokens processed. A useful output measure for training needs further work.
3. **Answers are not equal.** A short reply and a long reasoning answer differ greatly in work. Tokens per MWh should always be reported alongside answers per MWh, and comparisons between facilities should state their workload mix.
4. **Multi-tenant colocation.** Where the operator does not run the AI workloads, tenants must supply serving metadata. The method needs a standard tenant data exchange, ideally defined with operators.
5. **Hardware coverage.** Not every accelerator reports energy with the same accuracy, and some do not report it at all. Rack-level metering is the fallback, with wider uncertainty stated.
6. **Gaming.** Labels can be misapplied and output inflated, for example with trivial requests. Cross-checks between labels and hardware activity, and verification by an independent party, reduce but do not remove this risk.
7. **Water and carbon depend on site data.** Per-answer water uses site WUE and per-answer carbon uses published grid factors; both inherit the accuracy of those inputs.

Comments on any of these are welcome.

9. References

- MITI: Guidelines for Sustainable Development of Data Centres
- Energy Efficiency and Conservation Act 2024 (Act 861) · EEC Regulations 2024 · Energy Commission: who EECA covers
- Energy Commission: grid emission factors 2022–2024 (provisional)
- MCMC MTSFB TC G004:2024 Specification for Green Data Centres
- Malay Mail: new non-AI data centres frozen (24 Feb 2026)
- w.media: contracted versus used capacity · The Star: paying for unused allocations
- EdgeProp: water granted at 21–28% of requests
- w.media: national strategic direction and data-centre bill
- Securities Commission: National Sustainability Reporting Framework
- Google: environmental impact of Gemini prompts
- ISO/IEC 30134-2 (PUE), 30134-8 (CUE), 30134-9 (WUE)

Version history

v0.1, 9 November 2026: first public draft for comment.

This work is licensed under CC BY 4.0