Energy per answer, in Malaysia's reporting terms.



A measured benchmark of AI inference energy, scaled to a reference Malaysian data hall and expressed in the indicators of the MITI data-centre guidelines and the Energy Efficiency and Conservation Act 2024.

Summary

Indicator	Value	Basis
Facility energy per AI answer	0.17 Wh	Measured accelerator energy, scaled to the reference hall
Carbon per 1,000 answers	129 g CO_2	Peninsular grid factor, reference PUE
Avoidable accelerator energy	4.7% to 15.7%	Measured; range depends on how often traffic repeats itself
Annual saving, reference hall	20,300 to 67,900 GJ	Modelled from the measured shares
Capacity freed, reference hall	0.7 to 2.2 MW	Accelerator load that no longer has to be built

What was measured

A controlled benchmark on a production-class inference stack, across four workload types common in enterprise AI: support chat, questions over long documents, multi-step agents and long-form summaries. Load moved through idle, low, medium, peak and cool-down phases. Accelerator energy was read from the hardware counter continuously and joined to engine and request logs. Every request carried a label recording what it really was, so each finding could be checked.

Finding	Result
Energy follows busy time	Idle draw, plus a fixed draw while the accelerator is busy, plus a small cost per token explained 99.8% of energy across every measurement window. Counting tokens alone explained 58%.
Utilisation decides efficiency	The same hardware used 3.3 times less energy per request at peak load than at low load.
Memory, not compute, was the limit	The working memory for active requests was above 90% full for a third of the time and reached 100%. Requests were paused and recomputed, and response times rose sharply as memory filled while compute still had room.
Where energy went	Long summaries were 7% of requests and 37% of energy. Agent tasks were 64% of requests and 38% of energy, 15 points of it in parallel sub-calls.
Shared documents placed mid-prompt	0% reuse of earlier work, against 55% when the same document came first.

Method, raw data and scripts are available to operators and regulators under NDA.

Reference facility

To put the measurements in national terms, they are applied to a reference AI inference hall. The figures are assumptions, chosen to be typical, and every result below scales in proportion to them.

Assumption	Value
IT capacity	20 MW
Average IT load	70%
Accelerators' share of IT energy	70%
Power usage effectiveness (PUE)	1.4
Water usage effectiveness (WUE)	2.2 m^3 per MWh of IT energy
Grid emission factor, Peninsular Malaysia	0.758 t CO_2 per MWh
Annual IT energy	**122,640 MWh (441,504 GJ)**
Annual facility energy	**171,696 MWh**

MITI guideline indicators

The guidelines measure efficiency as ratios. Software that removes avoidable work lowers the energy, water and carbon behind every ratio, but leaves the ratios themselves unchanged, because cooling and power losses fall in proportion. The saving shows in absolute terms, and in energy per answer.

Indicator	Baseline	Conservative	Benchmark mix
PUE	1.40	1.40	1.40
WUE (m^3/MWh)	2.20	2.20	2.20
CUE (t CO_2/MWh IT)	1.06	1.06	1.06
Facility energy (MWh/yr)	171,696	166,047	152,827
Carbon (t CO_2/yr)	130,146	125,864	115,843
Water (m^3/yr)	269,808	260,931	240,156
Facility energy per answer (Wh)	0.170	0.164	0.151

CUE is the grid factor multiplied by PUE. Conservative counts only retried calls and mid-prompt documents. Benchmark mix adds repeated questions at the rate seen in the test.

EECA 2024 energy saving measures

At 441,504 GJ of IT energy a year, the reference hall is well above the Act's 21,600 GJ threshold, so it needs a registered energy manager, an energy management system and a yearly Energy Efficiency and Conservation report. The measures below are written as they would appear in that report's energy saving measures.

Measure	Facility energy (GJ/yr)	Carbon (t/yr)	Water (m^3/yr)
ESM 1. Remove identical retried calls	9,086	1,913	3,966
ESM 2. Place shared documents first, so earlier work is reused	11,250	2,369	4,911
Conservative total	**20,336**	**4,282**	**8,877**

Measure	Facility energy (GJ/yr)	Carbon (t/yr)	Water (m^3/yr)
ESM 3. Answer repeated questions from a cache	up to 47,594	up to 10,021	up to 20,775
Benchmark mix total	**67,930**	**14,303**	**29,652**

Each measure is verified the same way before it is counted: real requests are replayed with the change, and the saving stands only if answer quality holds.

Capacity

Energy is one view of the same saving. Removing 4.7% to 15.7% of accelerator work in the reference hall frees 0.7 to 2.2 MW of accelerator load. Because the benchmark showed memory, not compute, as the limit, that capacity serves more answers on the hardware already installed, before any new hall is built or any new grid connection is requested.

Limits

- Repeat rates vary widely between services, so ESM 3 is shown as an upper bound. Each operator's figures come from its own traffic.
- Energy was measured at the accelerator. Server, cooling and conversion losses are added through the stated share and PUE, not metered.
- The reference hall is a model. Its results scale in proportion to its assumptions.
- Per-workload energy is attributed by a statistical model; the totals it splits are measured.

Run it on your own hall

Croutons reads the exported request logs, engine metrics and accelerator telemetry an operator already has, on the operator's own machine, and produces these tables for that hall. Nothing needs to be installed in production and no prompts leave the site.

NRLC.ai · croutons.ai