THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

HOW HUNGRY IS AI?

Benchmarking Energy, Water, and Carbon Footprint of LLM Inference

Presented by: Abdeltawab Hendawi

Director of the AI-Lab for Research and Innovation

University of Rhode Island (hendawi@uri.edu)

Authors: Nidhal Jegham, Marwan Abdelatti, Chan Young Koh, Lasaad Moubarki, Abdeltawab Hendawi

THINK BIG WE DO

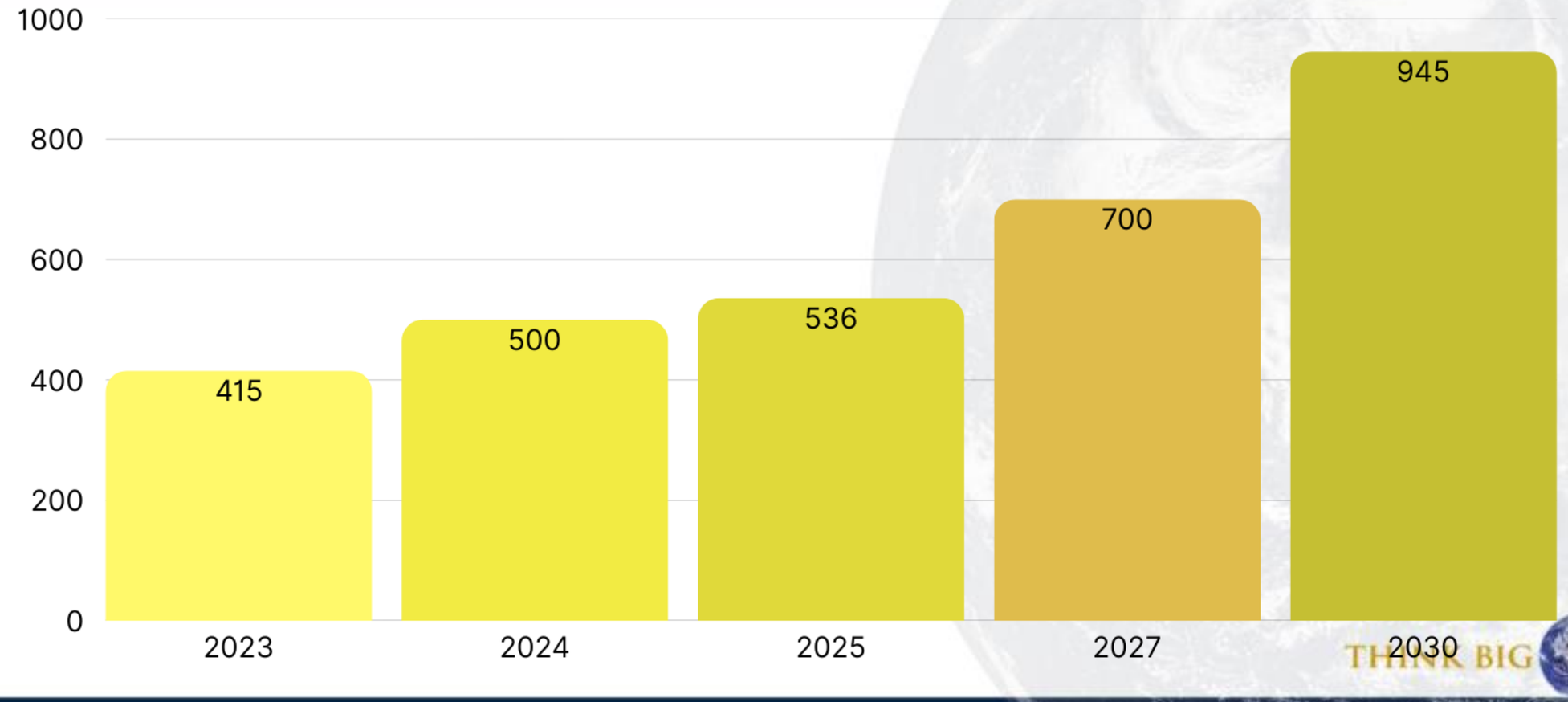
THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

WHAT IS THE ENVIRONMENTAL IMPACT OF AI?

THINK BIG WE DO

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

ANNUAL ENERGY CONSUMPTION OF DATA CENTERS (TWH)

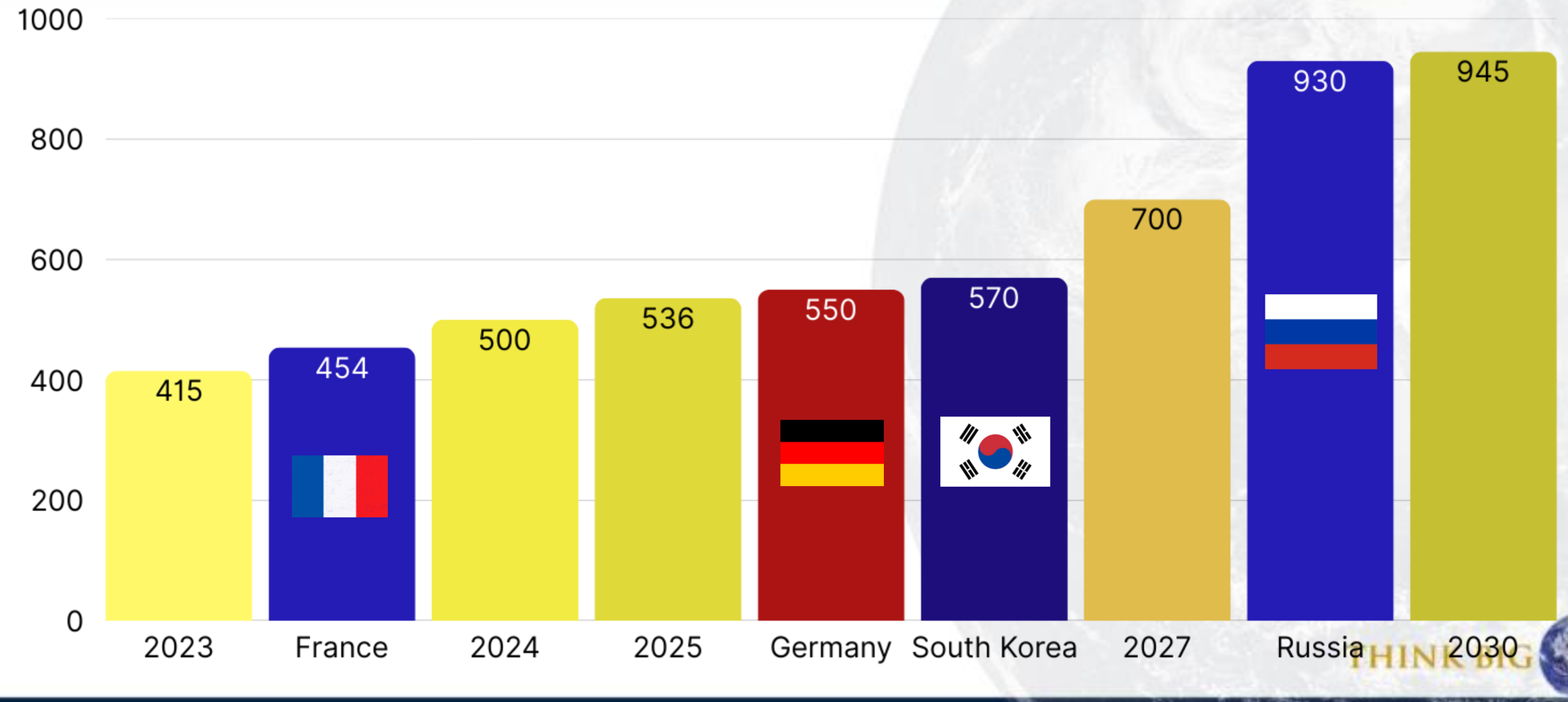


THINK BIG WE DO

1. HTTPS://WWW.DELOITTE.COM/US/EN/INSIGHTS/INDUSTRY/TECHNOLOGY/TECHNOLOGY-MEDIA-AND-TELECOM-PREDICTIONS/2025/GENAI-POWER-CONSUMPTION-CREATES-NEED-FOR-MORE-SUSTAINABLE-DATA-CENTERS.HTML

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

ANNUAL ENERGY CONSUMPTION OF DATA CENTERS (TWH)



THINK BIG WE DO

1. HTTPS://WWW.DELOITTE.COM/US/EN/INSIGHTS/INDUSTRY/TECHNOLOGY/TECHNOLOGY-MEDIA-AND-TELECOM-PREDICTIONS/2025/GENAI-POWER-CONSUMPTION-CREATES-NEED-FOR-MORE-SUSTAINABLE-DATA-CENTERS.HTML

2. HTTPS://WWW.WORLDOMETERS.INFO/ELECTRICITY/

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

20%

OF DATA CENTER POWER DEMAND IS ATTRIBUTED TO ARTIFICIAL INTELLIGENCE

THINK BIG WE DO™

1. HTTPS://WWW.SCIENCEDIRECT.COM/SCIENCE/ARTICLE/ABS/PII/S2542435125001424

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

90%

OF AI POWER DEMAND IS ATTRIBUTED TO THE INFERENCE PHASE

THINK BIG WE DO

1. HTTPS://WWW.SCIENCEDIRECT.COM/SCIENCE/ARTICLE/PII/S2210537923000124

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

WHY AI'S ENVIRONMENTAL IMPACT REMAINS A BLACK BOX

Sustainability Reports Without Substance

THINK BIG WE DO

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

MICROSOFT SUSTAINABILITY REPORT 2024

1.2 Energy

Table 6 – Energy consumption within the organization (MWh)

	FY20	FY21	FY22	FY23
Total energy consumption[1]	**11,283,502**	**14,133,987**	**18,644,872**	**24,007,868**
Non-renewable fuel consumed	**449,304**	**446,417**	**473,137**	**413,955**
Natural gas	218,557	249,443	273,964	**150,972**
Crude oil/diesel	147,297	143,370	117,195	**160,754**
LPG/propane/jet fuel	40,450	4,245	34,152	**54,239**
Gasoline	43,000	49,359	47,826	**47,990**
Electricity, heating, cooling, and steam	**10,834,198**	**13,687,570**	**18,171,735**	**23,593,913**
Electricity	10,770,714	13,621,517	18,153,454	**23,567,502**
Cooling (chilled water)	51,026	54,953	7,393	**12,090**
Hot water/steam	12,458	11,100	10,888	**14,321**
Total renewable electricity consumption[2]	**10,244,377**	**12,969,393**	**18,153,454**	**23,567,502**
Renewable energy credits and PPAs	10,244,059	12,969,246	18,153,218	**23,564,161**
Onsite renewable energy	318	147	236	**3,341**

1. Only reported categories and values are applicable to Microsoft's energy consumption. Renewable fuels, electricity sold, heating sold, cooling sold, and steam sold categories are currently not applicable. Reported values for FY23 expressed in gigajoules (GJ): total energy consumption equals 86,428,325 GJ, and total non-renewable fuel consumed equals 1,490,239 GJ.
2. Reported values represent Microsoft's total renewable energy consumption expressed in MWh from onsite, renewable energy credits, power purchase agreements (PPAs), and green power tariff programs. Values reflect Microsoft's renewable electricity consumption at the time of reporting.

Table 7 – Renewable energy metrics

	FY20	FY21	FY22	FY23
Percentage of renewable electricity[1]	100%	100%	100%	**100%**
Percentage of direct renewable electricity	–	–	62%	**59%**

1. Values reflect Microsoft's percentage of renewable electricity consumption at the time of reporting.

Table 8 – Energy intensity (MWh/revenue $M)

	FY20	FY21	FY22	FY23
Electricity consumed within the organization (MWh)	10,770,714	13,621,517	18,153,454	**23,567,502**
Revenue ($M)	143,015	168,088	198,270	**211,915**
Electricity consumption normalized by revenue (MWh/$M)	75	81	92	**111**

1.3 Water

Table 9 – Water and effluents (megaliters)[1]

	FY20	FY21	FY22	FY23
Total water withdrawals[2]	**7,936**	**8,068**	**10,706**	**12,951**
Third-party water	7,831	8,011	10,665	**12,926**
Surface water	89	41	39	**21**
Ground water	16	16	2	**4**
Total water discharges[2,3]	**3,740**	**3,295**	**4,307**	**5,107**
Third-party water	3,740	3,295	4,307	**5,107**
Total water consumption[2]	**4,196**	**4,773**	**6,399**	**7,844**

1. For FY23, total water withdrawal from areas with water stress was 5,326 megaliters (ML) (41%) and was primarily sourced from third-party water; total water discharge to areas with water stress was 2,045 ML (40%); and total water consumption from areas with water stress was 3,281 ML (42%). Water risk assessment was conducted using WRI's Aqueduct tool for areas in high or extremely high baseline water stress.
2. Brackish surface water/seawater and produced water categories are not relevant to Microsoft since there is no direct withdrawal or discharge of water from and to these sources. For withdrawals, data breakdown between "freshwater" and "other water" categories, and data for third-party withdrawal sources for areas with water stress is currently unavailable and will be part of data improvements going forward. For the periods presented we are not gathering data around water storage since it is not a significant portion of our water inventory.
3. Only discharges to third parties are relevant since water that is not consumed at Microsoft sites is discharged to local municipal treatment plants. Discharges to surface water, groundwater, and seawater, and volume sent for use to other organizations are not applicable. For discharges, data breakdown between "freshwater" and "other water" categories is currently unavailable and will be part of data improvements going forward. Primary treatment of water is not relevant because there are no onsite water treatment plants in Microsoft operations, as there is no requirement to conduct onsite primary treatment of discharge by any environmental regulation or standard.

1. HTTPS://WWW.MICROSOFT.COM/EN-US/CORPORATE-RESPONSIBILITY/SUSTAINABILITY/REPORT/

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

OBSTACLES & CHALLENGES

- Proprietary Hardware and Models
- Inference is a Much Bigger Problem than Training
- Hidden Data Centers Locations
- Lack of Model-specific or Region-specific Environmental Disclosures

THINK BIG WE DO™

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

OBSTACLES & CHALLENGES

Lack of a Standardized Methodology to Quantify The Environmental Footprint of LLM Models

THINK BIG WE DO

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

OUR CONTRIBUTION & NOVELTY

The first research to provide **prompt-level** estimates of **energy**, **water**, and **carbon** footprint across 30+ **open-source** and **proprietary** LLMs using public API performance data.

THINK BIG WE DO

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

PRELIMINARIES

Power Usage Effectiveness (PUE)

$$PUE = \frac{TotalFacilityEnergy}{ITEquipmentEnergy}$$

Water Usage Effectiveness (WUE)

$$WUE = \frac{AnnualWaterUsage(liters)}{ITEquipmentEnergy\,(kWh)}$$

Carbon Intensity Factor (CIF)

$$CIF = \frac{TotalCO_2eEmissions\,(Kg)}{TotalFacilityEnergy\,(kWh)}$$

THINK BIG WE DO™

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

METHODOLOGY

Model	Launch Date	Company	Host	Hardware	Critical Power (kW)	PUE	WUE (on-site, L/kWh)	WUE (off-site, L/kWh)	CIF ($kgCO_2e/kWh$)
GPT-4.1	Apr, 2025	OpenAI	Microsoft Azure	DGX H200/H100 [30, 31]	10.20 [32]	1.12 [33]	0.30 [34]	4.35 [35]	0.35 [36]
GPT-4.1 mini	Apr, 2025								
GPT-4.1 nano	Apr, 2025								
o4-mini (high)	Apr, 2025								
o3	Apr, 2025								
o3-mini (high)	Jan, 2025								
o3-mini	Jan, 2025								
o1	Dec, 2024								
o1-mini	Sep, 2024								
GPT-4o (Mar '25)	May, 2024								
GPT-4o mini	July, 2024	OpenAI	Microsoft Azure	DGX A100*	6.50[37]	1.12	0.30	4.35	0.35
GPT-4 Turbo	Nov, 2023								
GPT-4	Mar, 2023								
DeepSeek-R1	Jan, 2025	Deepseek	Deepseek	DGX H800 [8]	10.20 [38]	1.27 [39]	1.20 [39]	6.016 [35]	0.6 [40]
DeepSeek-V3	Dec, 2024								
DeepSeek-R1	Jan, 2025	Deepseek	Microsoft Azure	DGX H200/H100	10.20	1.12	0.30	4.35	0.35
DeepSeek-V3	Dec, 2024								
Claude-3.7 Sonnet	Feb, 2025	Anthropic	AWS	DGX H200/H100 [41, 42]	10.20	1.14 [43]	0.18 [43]	5.11 [35]	0.287 [44]
Claude-3.5 Sonnet	Jun, 2024								
Claude-3.5 Haiku	Nov, 2024								
LLaMA-3.3 70B	Dec, 2024	Meta	AWS	DGX H200/H100	10.20	1.14	0.18	5.11	0.287
LLaMA-3.2-vision 90B	Sep, 2024								
LLaMA-3.2-vision 11B	Sep, 2024								
LLaMA-3.2 3B	Sep, 2024								
LLaMA-3.2 1B	Sep, 2024								
LLaMA-3.1-405B	Jul, 2024								
LLaMA-3.1-70B	Jul, 2024								
LLaMA-3.1-8B	Jul, 2024								
LLaMA-3-70B	Apr, 2024								
LLaMA-3-8B	Apr, 2024								

THINK BIG WE DO

METHODOLOGY

Energy Consumption Per-Query Formula

$$\underbrace{\frac{\text{Output Length}}{\text{TPS}}}_{\text{Time to generate the response (s)}}$$

- ***Output Length:*** *Total number of tokens in the model's output*
- ***TPS (Tokens Per Second):*** *Output token generation speed*

METHODOLOGY

Energy Consumption Per-Query Formula

$$\underbrace{\frac{\text{Output Length}}{\text{TPS}}}_{\text{Time to generate the response (s)}} + \underbrace{\text{Latency}}_{\text{Time to process the query (s)}}$$

- ***Output Length:*** *Total number of tokens in the model's output*
- ***TPS (Tokens Per Second):*** *Output token generation speed*
- ***Latency:*** *Time (in seconds) from prompt submission to first token generated*

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

METHODOLOGY

Energy Consumption Per-Query Formula

$$\underbrace{\left(\frac{\frac{\text{Output Length}}{\text{TPS}} + \text{Latency}}{3600}\right)}_{\text{Total inference time (hours)}}$$

- ***Output Length:*** *Total number of tokens in the model's output*
- ***TPS (Tokens Per Second):*** *Output token generation speed*
- ***Latency:*** *Time (in seconds) from prompt submission to first token generated*

METHODOLOGY

Energy Consumption Per-Query Formula

$$\underbrace{P_{\text{GPU}} \times U_{\text{GPU}}}_{\text{GPU power (kW)}}$$

- P_{GPU}: *Maximum rated power (in kW) of the GPUs*
- U_{GPU}: *Proportion of GPU capacity actively used*

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

METHODOLOGY

Energy Consumption Per-Query Formula

$$\underbrace{P_{\text{GPU}} \times U_{\text{GPU}}}_{\text{GPU power (kW)}} + \underbrace{P_{\text{non-GPU}} \times U_{\text{non-GPU}}}_{\text{Non-GPU power (kW)}}$$

- P_{GPU} : *Maximum rated power (in kW) of the GPUs*
- U_{GPU} : *Proportion of GPU capacity actively used*
- $P_{\text{non-GPU}}$: *Maximum rated power (in kW) of the non-GPU components (e.g., CPU, RAM, SSD)*
- $U_{\text{non-GPU}}$: *Proportion of non-GPU system power actively used during inference*

THINK BIG WE DO

METHODOLOGY

Energy Consumption Per-Query Formula

$$\underbrace{\left(\underbrace{P_{\text{GPU}} \times U_{\text{GPU}}}_{\text{GPU power (kW)}} + \underbrace{P_{\text{non-GPU}} \times U_{\text{non-GPU}}}_{\text{Non-GPU power (kW)}}\right)}_{\text{Total system power draw (kW)}}$$

- P_{GPU} *: Maximum rated power (in kW) of the GPUs*
- U_{GPU} *: Proportion of GPU capacity actively used*
- $P_{\text{non-GPU}}$ *: Maximum rated power (in kW) of the non-GPU components (e.g., CPU, RAM, SSD)*
- $U_{\text{non-GPU}}$ *: Proportion of non-GPU system power actively used during inference*

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

METHODOLOGY

Energy Consumption Per-Query Formula

$$E_{\text{query (kWh)}} = \underbrace{\left(\frac{\frac{\text{Output Length}}{\text{TPS}} + \text{Latency}}{3600} \right)}_{\text{Total inference time (hours)}} \underbrace{\left(\underbrace{P_{\text{GPU}} \times U_{\text{GPU}}}_{\text{GPU power (kW)}} + \underbrace{P_{\text{non-GPU}} \times U_{\text{non-GPU}}}_{\text{Non-GPU power (kW)}} \right)}_{\text{Total system power draw (kW)}} \cdot \text{PUE}$$

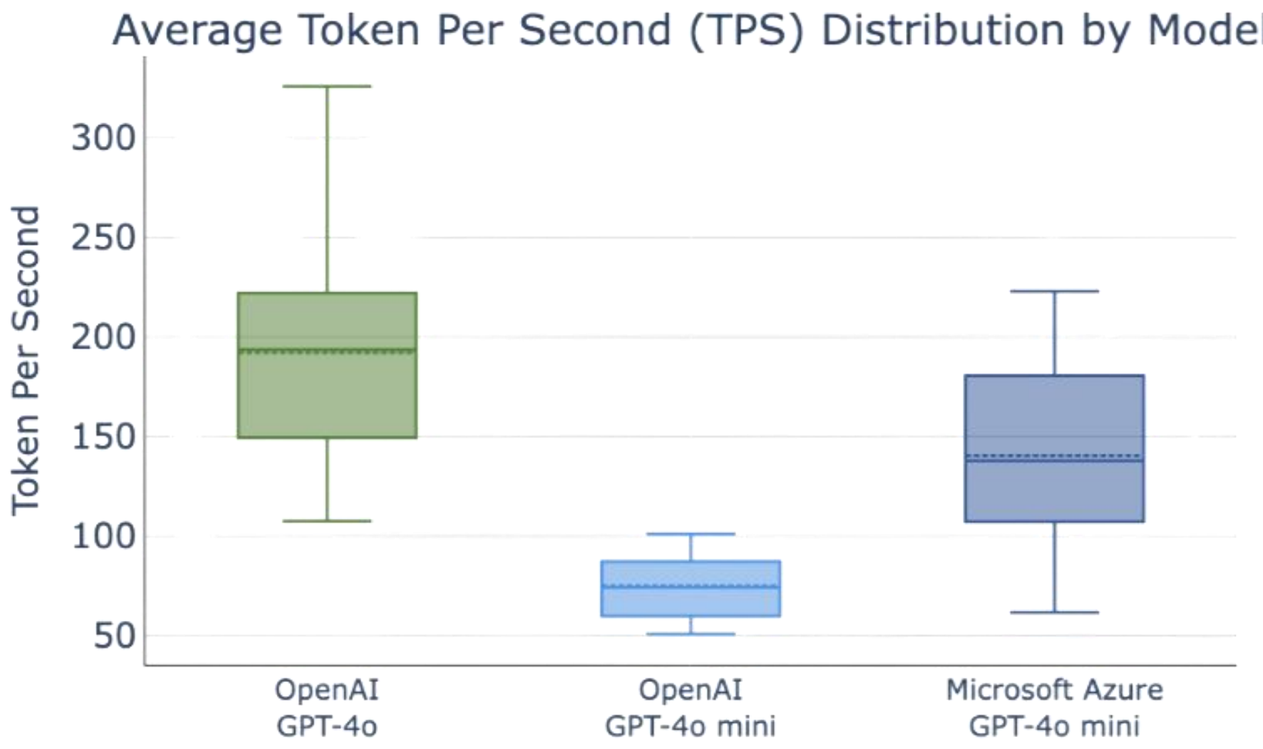
- ***Power Usage Effectiveness (PUE):*** *accounts for additional energy overhead like cooling, lighting, and power delivery systems in the data center.*

THINK BIG WE DO

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

METHODOLOGY

GPT-4o Mini Hardware Estimation



OpenAI GPT-4o mini is 60% slower than **OpenAI GPT-4o**

→**OpenAI *GPT-4o mini*** *is either deployed on* ***H100*** *or* ***A100****, compared to GPT-4o's* ***H200****.*

Azure GPT-4o mini is 27% slower than OpenAI GPT-4o

→***Microsoft Azure GPT-4o*** *mini is most likely deployed on* ***H100*** *or* ***A100*** *as well.*

OpenAI GPT-4o mini is 47% slower than **Azure GPT-4o mini**

→***OpenAI GPT-4o mini*** *is most likely deployed on* ***A100****, and* ***Microsoft Azure GPT-4o mini*** *is deployed on* ***H100***

THINK BIG WE DO

How Hungry is AI? | Dashboard Overview

Last Refresh: 30-Nov-2025 17:05:36

Dashboard created by Nidhal Jegham

Sustainability Metrics

- Average Carbon Emissions (gCO2e)
- Average Energy Consumption (Wh)
- Average Water Consumption (Site & Source, mL)
- Average Water Consumption (Site, mL)
- Average Water Consumption (Source, mL)

Company: All

Model Size: All

Query Length: Medium

Number of Queries: 100B

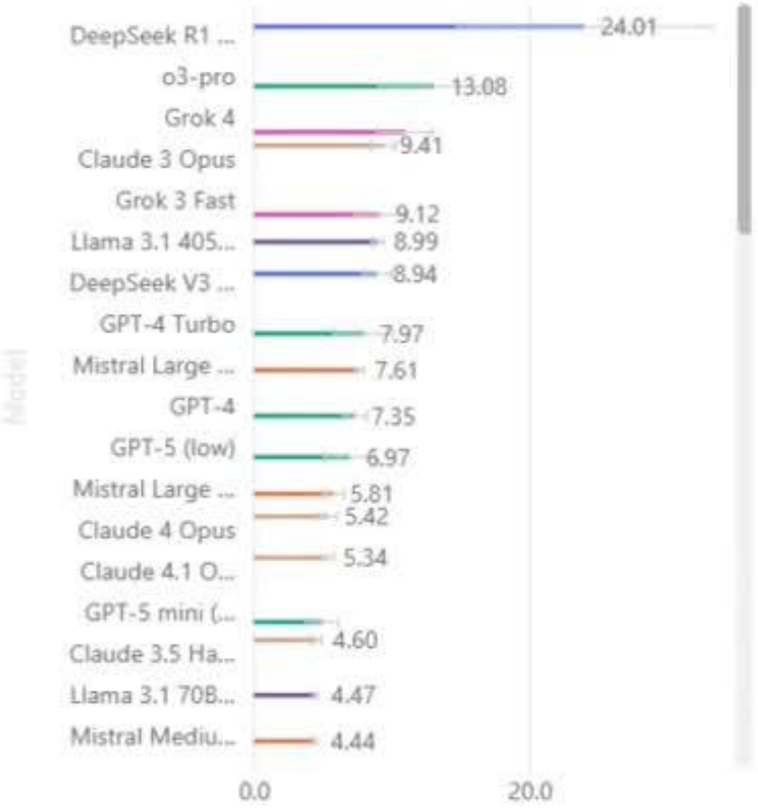
Reset Settings

Download Data (CSV)

Company	Model	Host	Hardware	Model Size	PUE	WUE (Site) (L/kWh)	WUE (Source) (L/kWh)	CIF (KgCO2/kWh)
Select a Model	Select a Model	Select a Model	Select a Model	Select a...	Select a...	Select a...	Select a...	Select a...

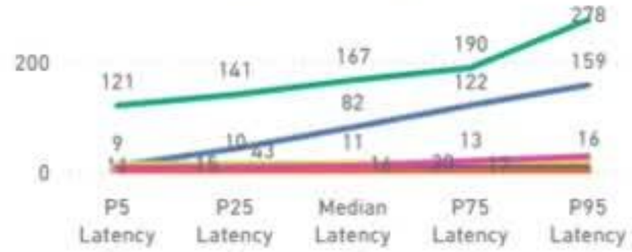
Average Energy Consumption (Wh) | Mult Companies Models | Multiple Sizes | Medium Prompts

Anthropic DeepSeek Google Meta Mistral AI OpenAI xAI



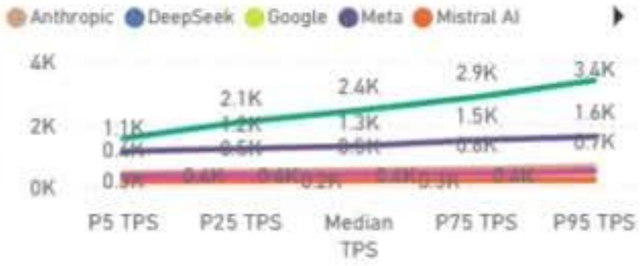
Latency (s) for Multiple Models

Anthropic DeepSeek Google Meta Mistral AI



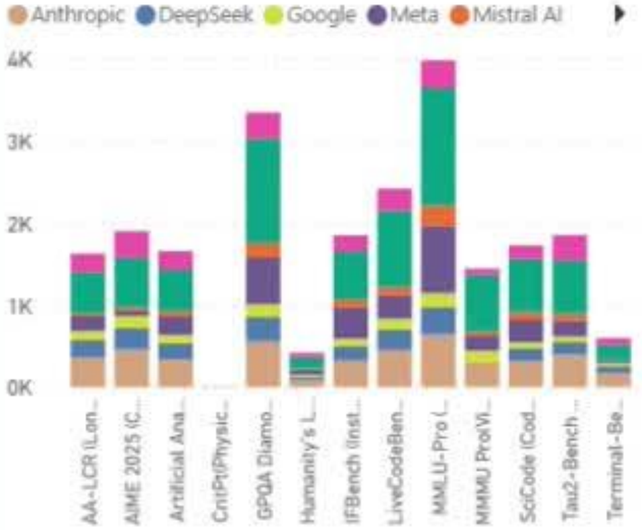
Tokens-per-Second (TPS) for Multiple Models

Anthropic DeepSeek Google Meta Mistral AI



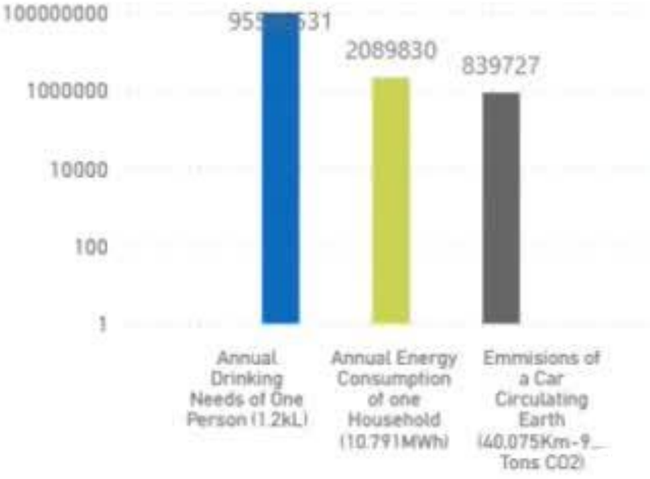
Performance Overview of Multiple Models

Anthropic DeepSeek Google Meta Mistral AI



Environmental Footprint Equivalent of Multiple Models for 100B Queries

Carbon Emissi... Energy Cons... Water Cons...



THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

INDUSTRY VALIDATION

Sam Altman (OpenAI CEO):

GPT-4o consumes 0.34 Wh Per Query

How Hungry is AI Estimates:

GPT-4o consumes 0.42±0.085 Wh Per Short Query

THINK BIG WE DO™

1. HTTPS://BLOG.SAMALTMAN.COM/THE-GENTLE-SINGULARITY

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

INDUSTRY VALIDATION

Mistral LCA Report:

Mistral Large 2 emits 1.14 gCO2e per 400-tokens query

How Hungry is AI Estimates:

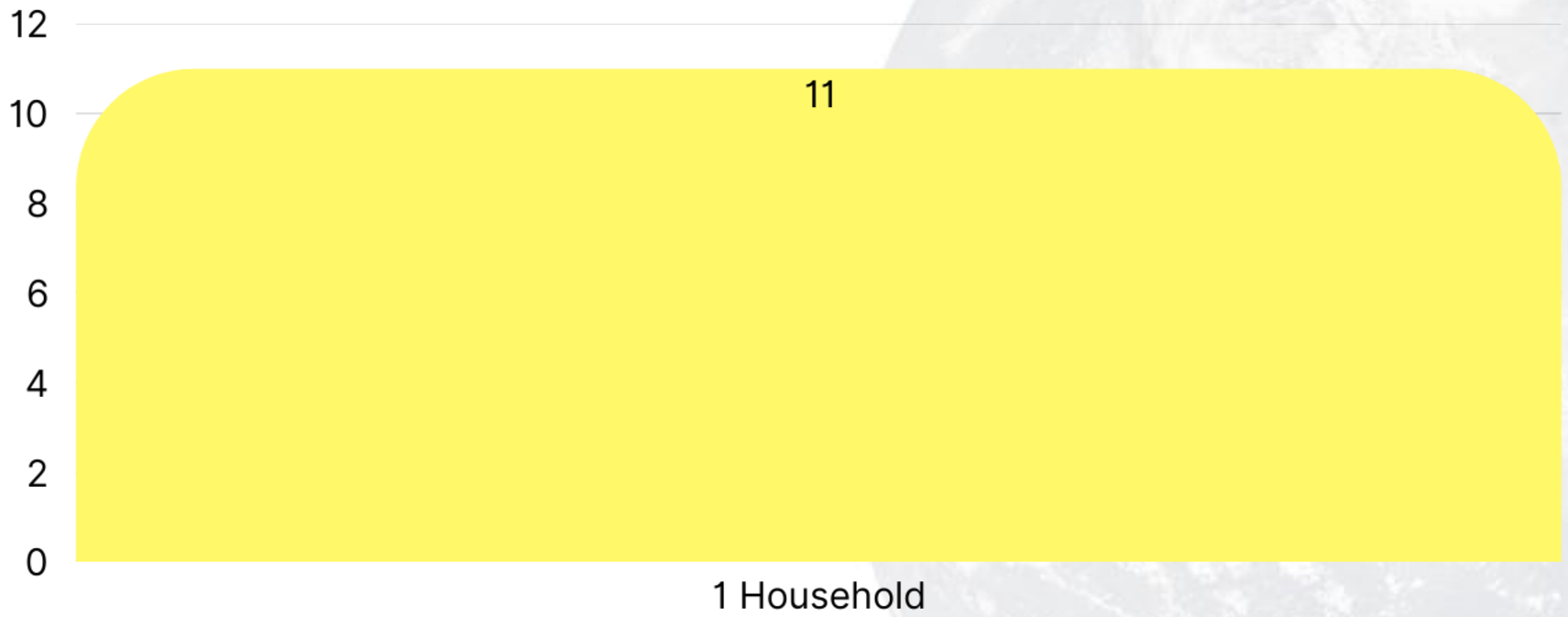
Mistral Large 2 emits 1.09±0.1 gCO2e per 400-tokens query

THINK BIG WE DO

1. HTTPS://MISTRAL.AI/NEWS/OUR-CONTRIBUTION-TO-A-GLOBAL-ENVIRONMENTAL-STANDARD-FOR-AI

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

GPT-4O ANNUAL ENERGY CONSUMPTION (MWH)

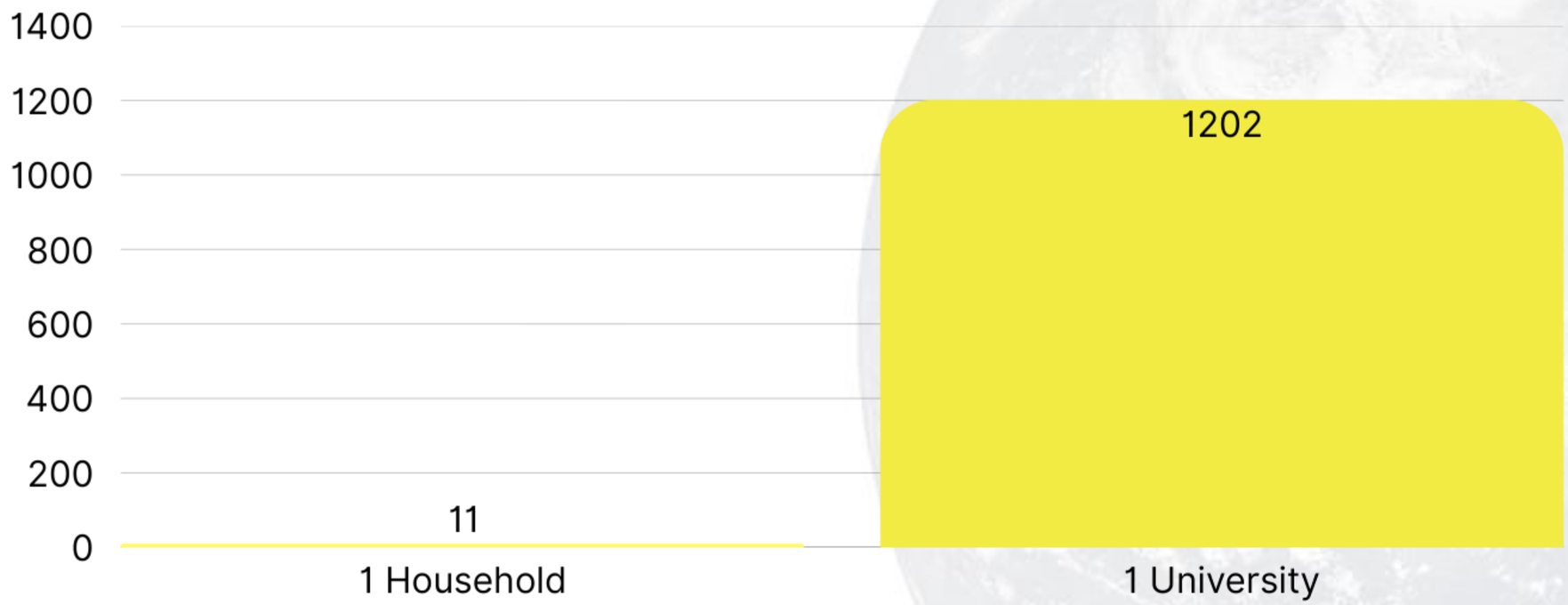


THINK BIG WE DO

1. HTTPS://WWW.EIA.GOV/ENERGYEXPLAINED/USE-OF-ENERGY/ELECTRICITY-USE-IN-HOMES.PHP

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

GPT-4O ANNUAL ENERGY CONSUMPTION (MWH)

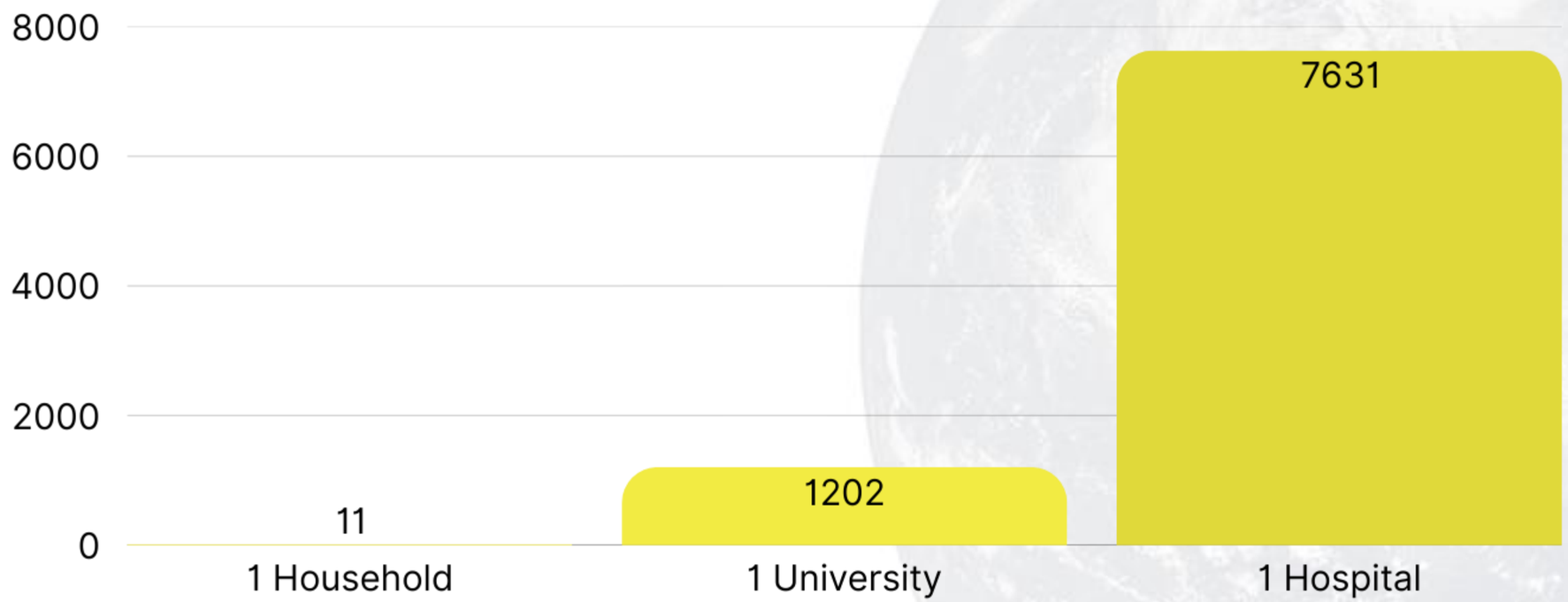


THINK BIG WE DO

1. HTTPS://WWW.EIA.GOV/ENERGYEXPLAINED/USE-OF-ENERGY/ELECTRICITY-USE-IN-HOMES.PHP
2. HTTPS://WWW.EIA.GOV/CONSUMPTION/COMMERCIAL/DATA/2018/BC/PDF/B1.PDF

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

GPT-4O ANNUAL ENERGY CONSUMPTION (MWH)

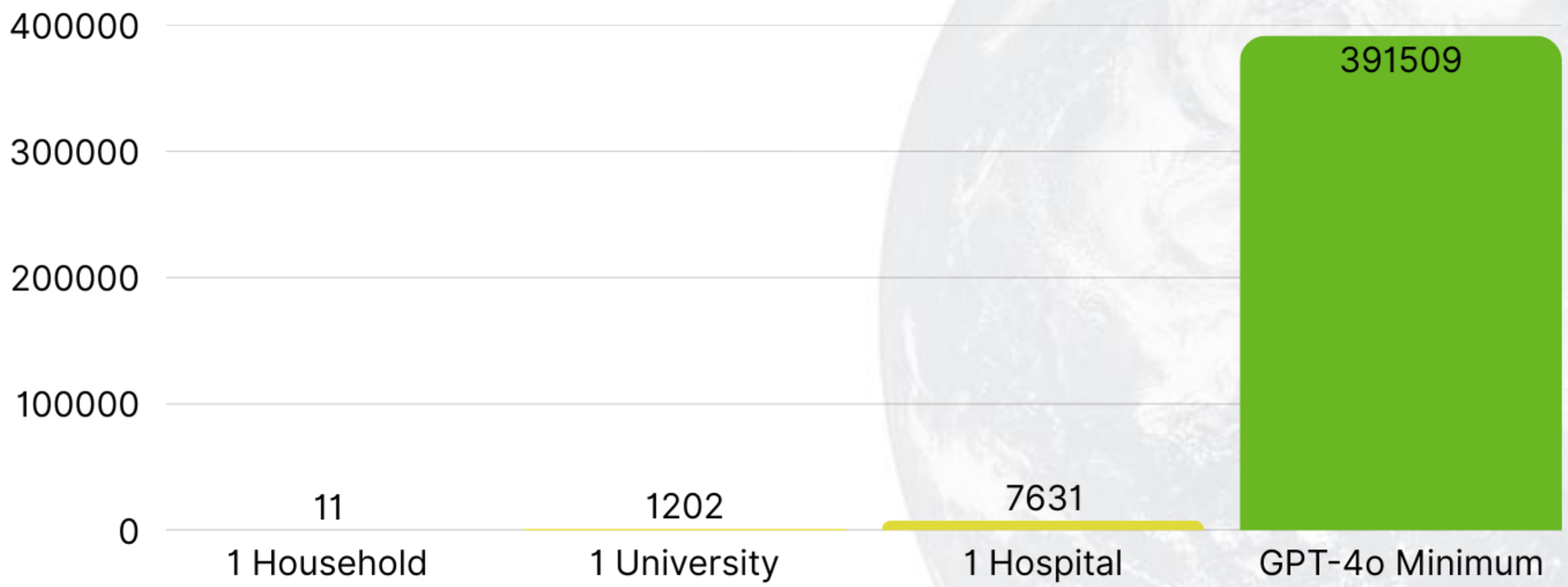


THINK BIG WE DO

1. HTTPS://WWW.EIA.GOV/ENERGYEXPLAINED/USE-OF-ENERGY/ELECTRICITY-USE-IN-HOMES.PHP
2. HTTPS://WWW.EIA.GOV/CONSUMPTION/COMMERCIAL/DATA/2018/BC/PDF/B1.PDF

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

GPT-4O ANNUAL ENERGY CONSUMPTION (MWH)

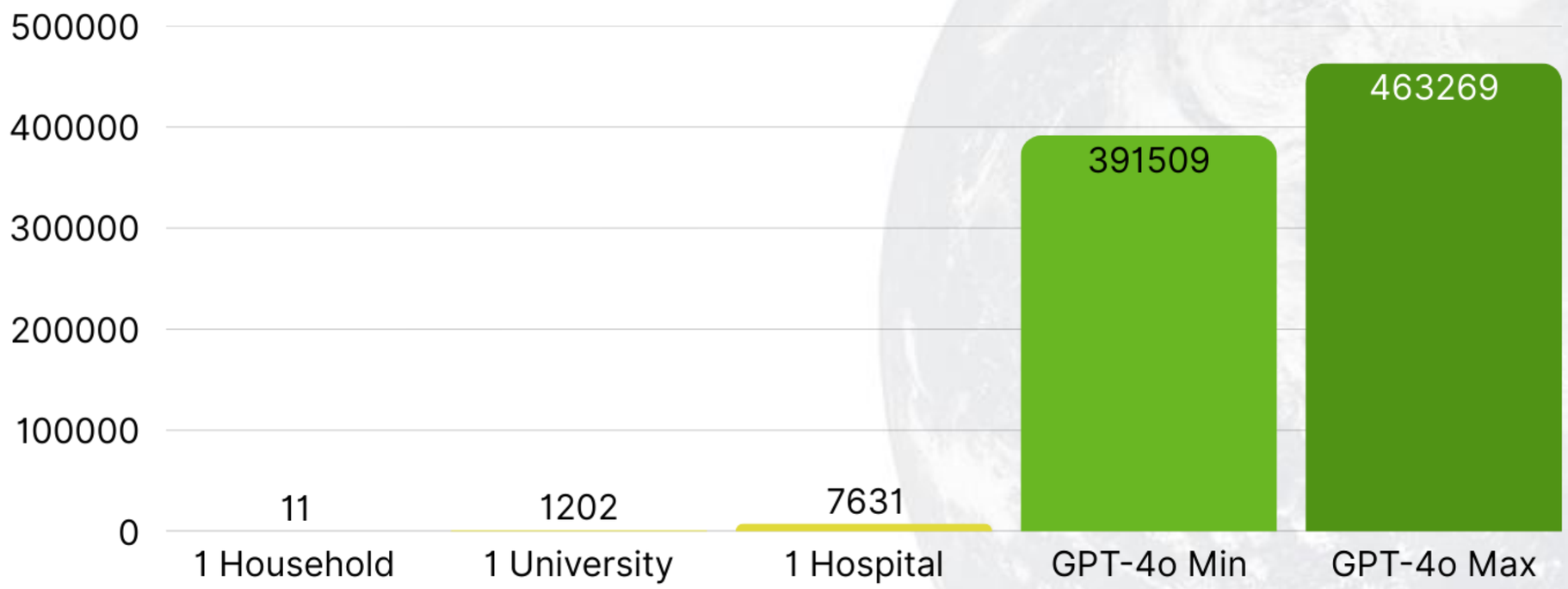


THINK BIG WE DO

1. HTTPS://WWW.EIA.GOV/ENERGYEXPLAINED/USE-OF-ENERGY/ELECTRICITY-USE-IN-HOMES.PHP
2. HTTPS://WWW.EIA.GOV/CONSUMPTION/COMMERCIAL/DATA/2018/BC/PDF/B1.PDF

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

GPT-4O ANNUAL ENERGY CONSUMPTION (MWH)

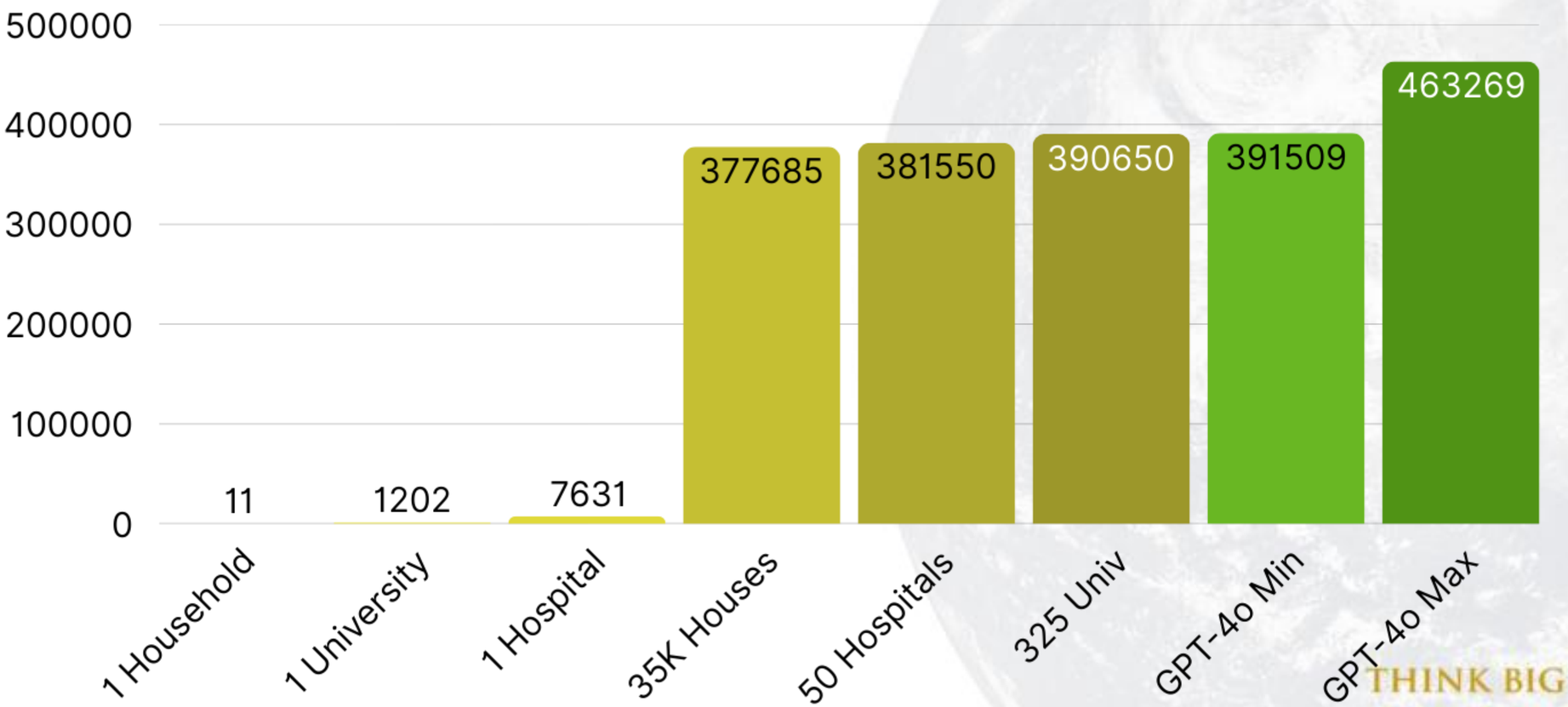


THINK BIG WE DO

1. HTTPS://WWW.EIA.GOV/ENERGYEXPLAINED/USE-OF-ENERGY/ELECTRICITY-USE-IN-HOMES.PHP
2. HTTPS://WWW.EIA.GOV/CONSUMPTION/COMMERCIAL/DATA/2018/BC/PDF/B1.PDF

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

GPT-4O ANNUAL ENERGY CONSUMPTION (MWH)



THINK BIG WE DO

1. HTTPS://WWW.EIA.GOV/ENERGYEXPLAINED/USE-OF-ENERGY/ELECTRICITY-USE-IN-HOMES.PHP
2. HTTPS://WWW.EIA.GOV/CONSUMPTION/COMMERCIAL/DATA/2018/BC/PDF/B1.PDF

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

DISCUSSION & KEY TAKEAWAYS

The Critical Role of Infrastructure in AI Sustainability

Rebound Effects and the Jevons (growing) Paradox

The Intensity of Chain-of-thought Models

THINK BIG WE DO

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

DISCUSSION & KEY TAKEAWAYS

Implementing Dielectric Liquid Cooling

Utilizing Larger Batch Sizes & Smaller Models

Relying on Renewable Energy Sources

More Transparent Reporting

THINK BIG WE DO

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

THE TEAM BEHIND HOW HUNGRY IS AI

Nidhal Jegham[1 2]

Marwan Abdelatti[1 3]

Chan Young Koh [1]

Lasaad Moubarki [2]

Abdeltawab Hendawi [1]

[1] Computer and Statistical Sciences Department, University of Rhode Island, RI, USA

[2] Tunis Business School, University of Tunis, Tunis, Tunisia

[3] Computer Science Department, Providence College, RI, USA

THINK BIG WE DO

THE UNIVERSITY OF RHODE ISLAND COOPERATIVE EXTENSION

THANK YOU FOR YOUR ATTENTION

Dashboard Link

Any Questions?

Paper Link