# AI ENERGY CONSUMPTION

## GLOBAL RISK ASSESSMENT

Energy Constraints, Territorial Shortfalls & the Investment Return Question

**Sparse Supernova Research — March 2026**

operations@sparse-supernova.com

# Executive Summary

Global AI spending is projected to reach $1.5 trillion in 2025 and exceed $2 trillion by 2026 [1]. At the core of this growth lies a critical dependency: energy. Data centres currently consume approximately 415 terawatt-hours (TWh) of electricity annually, representing around 1.5% of global electricity consumption. The IEA projects this will more than double to approximately 945 TWh by 2030, equivalent to the entire electricity demand of Japan [2].

The IEA identifies AI as “the most important driver of this growth, alongside growing demand for other digital services” [2]. AI-optimised data centre electricity demand is projected to more than quadruple by 2030. In the United States alone, data centre power consumption is on course to account for nearly half of the growth in total electricity demand between now and 2030 [2]. By 2030, the US is set to consume more electricity for data centres than for manufacturing all energy-intensive goods combined [2].

This report examines three interconnected risks: the physical energy constraints that threaten AI infrastructure deployment, the territorial inequalities where certain regions face acute energy shortfalls, and growing evidence that the investment thesis underpinning frontier AI companies may be fundamentally challenged by these energy realities. Where claims rest on estimates with uncertain methodology, these are explicitly quarantined and labelled.

# Global AI Energy Consumption: The Scale of the Problem

## Current State

Data centres, AI workloads, and cryptocurrency operations together consumed approximately 460 TWh of electricity in 2024, representing nearly 2% of global demand [3]. AI workloads account for an estimated 5–15% of total data centre electricity use under the IEA’s own framing [10]; secondary aggregators (FlowingData, AllAboutAI) place the range at 11–20% using broader definitions of AI-adjacent workloads [3]. Both ranges point in the same direction: AI’s share is growing rapidly. Data centre electricity demand has grown at 12% per year for the past five years, more than four times faster than total electricity demand growth from all other sectors [2].

Electricity consumption in AI-accelerated servers is projected to grow at 30% annually, compared to 9% for conventional servers. These accelerated servers account for almost half of the net increase in global data centre electricity consumption [2].

| Metric | 2024 | 2030 (Projected) |
|---|---|---|
| Global data centre electricity (TWh) | ~415 TWh [2] | ~945 TWh [2] |
| Share of global electricity | ~1.5% [2] | ~3% [2] |
| AI share of data centre power | 5–15% (IEA) / 11–20% (aggregators) [3][10] | 35–50% [10] |

| | | |
|---|---|---|
| Annual growth rate (data centres) | 12% [2] | ~15% [2] |
| AI server growth rate | 30% [2] | 30%+ [2] |

## Per-Query and Per-Training Energy: From Estimates to Benchmarks

> ⚠️ **Methodological Note: Per-query and training energy figures**
>
> The figures below draw on company claims, third-party analysis, academic modelling, and — new in this revision — infrastructure-aware benchmarking from Jegham et al. (2025) [26]. Methodologies vary significantly. OpenAI and other frontier labs have not disclosed verified per-query energy measurements. These numbers should be treated as order-of-magnitude guides, not precise constants.

Sam Altman has claimed a standard ChatGPT query consumes approximately 0.34 Wh [4]; the methodology behind this figure has not been publicly disclosed. Recent infrastructure-aware benchmarking by Jegham et al. provides independent cross-validation: their framework estimates approximately 0.42 Wh (±0.13 Wh) for a short GPT-4o prompt, or 0.37 Wh excluding data centre overhead [26]. This falls within 19% of Altman's figure, lending credibility to the general magnitude while confirming it should be treated as a range rather than a constant.

Third-party estimates for advanced reasoning models (e.g. OpenAI o3) suggest 7–40 Wh per query [5], but these have not been independently verified by a primary measurement study. Image generation is estimated to demand 20–40 times more energy than text, and video generation 1,000–3,000 times more [5].

Jegham et al. benchmarked 30 models across three prompt sizes and found the most energy-intensive systems exceed 29 Wh per long prompt, over 65× the most efficient systems [26]. Even within a single provider's lineup, infrastructure choices dominate: GPT-4o mini consumes approximately 20% more energy than GPT-4o on long queries because it runs on older A100 hardware rather than H100/H200 nodes [26]. Model architecture alone does not determine real-world energy consumption — the data centre stack matters as much or more.

**GPT-5 adaptive routing:** Jegham et al. include a dedicated GPT-5 case study (Section 7 of their paper) analysing adaptive model routing, where the system dynamically scales reasoning depth per prompt. Energy consumption ranges from 0.67 Wh for a short minimal-reasoning query to 33.8 Wh for a long high-reasoning query — a roughly 50× spread within a single system [26]. As models increasingly incorporate dynamic reasoning depth, per-query energy becomes harder to predict and harder to cap.

As Andy Wu of Harvard Business School observed: "While generative AI can do amazing things, it is also perhaps the most wasteful use of a computer ever devised." [6]

### Infrastructure-Aware Energy Benchmarks (Selected Models)

*Source: Jegham et al. (2025) [26]. Energy in Wh, mean ± std dev. Short = 100 input / 300 output tokens. Long = 10k input / 1.5k output tokens.*

| Model | Short (Wh) | Long (Wh) | Host |
|---|---|---|---|

| | | | |
|---|---|---|---|
| LLaMA-3.1-8B | 0.052 ± 0.008 | 0.443 ± 0.028 | AWS |
| GPT-4.1 nano | 0.207 ± 0.047 | 0.827 ± 0.094 | Azure |
| GPT-4o (Mar ’25) | 0.423 ± 0.085 | 2.875 ± 0.421 | Azure H200 |
| Claude-3.7 Sonnet | 0.950 ± 0.040 | 5.671 ± 0.302 | AWS |
| o3 | 1.177 ± 0.224 | 12.222 ± 1.082 | Azure |
| DeepSeek-R1 (Azure) | 2.353 ± 1.129 | 7.410 ± 2.159 | Azure |
| DeepSeek-R1 (DS own) | 19.251 ± 9.449 | 29.078 ± 9.725 | DeepSeek DC |
| LLaMA-3.1-405B | 2.226 ± 0.142 | 25.202 ± 0.526 | AWS |

*The highlighted row illustrates infrastructure dominance: DeepSeek-R1 on its own servers consumes over 29 Wh per long prompt, while the identical model on Azure consumes 7.4 Wh — a reduction of more than 70%, driven by hardware efficiency, cooling, PUE, and regional carbon intensity [26].*

## Training Costs and Efficiency

Training frontier models requires enormous compute. GPT-4’s training cost is estimated at $50–100 million [7], though neither the exact compute budget nor energy consumption has been publicly disclosed by OpenAI. Third-party estimates place it above 50 GWh, but this figure lacks a primary citation chain and should be treated as approximate.

DeepSeek-V3 represents a significant efficiency counterpoint. Its technical report discloses 2.788 million H800 GPU hours for the final training run, at an estimated cost of $5.6 million [8]. This represents a roughly 90–95% reduction in training compute cost compared to GPT-4’s estimated expenditure. However, this is a cost and compute comparison, not a verified energy measurement. Actual energy consumption depends on GPU power draw, utilisation rates, and data centre PUE, which were not disclosed. The $5.6 million figure also excludes prior research, ablation experiments, and infrastructure amortisation [8].

**Inference now dominates.** While training receives most public attention, inference is emerging as the primary contributor to lifecycle energy. Recent estimates suggest inference can account for up to 90% of a model’s total lifecycle energy use [26]. Jegham et al.’s GPT-4o case study projects approximately 391–463 GWh of annual energy from GPT-4o inference alone in 2025 — equivalent to the electricity consumption of 35,000 US households [26]. As deployment scales, inference’s share will only grow.

## Water and Material Consumption

By 2030, global data centre water use is projected to reach 450 million gallons per day, equivalent to the daily use of approximately 5 million people, up from 292 million gallons in 2022 [1]. Large AI data centres consume 300,000 to 5 million gallons daily. By 2027, global AI demand is projected to withdraw 1.1–1.7 trillion gallons of freshwater [5]. Two-thirds of US data centres are located in high-stress water regions [1].

**Per-query water footprint:** Jegham et al. quantify water consumption at the prompt level using WUE multipliers for both on-site cooling and off-site electricity generation. The most efficient models consume under 4 mL per long prompt, while DeepSeek-R1 on its own servers exceeds 200 mL per long query [26]. Scaled to GPT-4o's estimated 772 billion queries in 2025, annual water consumption reaches 1,335–1,580 megalitres — the annual drinking needs of 1.2 million people [26]. Infrastructure choice changes water outcomes massively: the same DeepSeek model on Azure consumes only 34 mL per long query, a reduction of nearly 85% [26].

# Territorial Energy Shortfalls: Where the Grid Cannot Cope

The energy challenge is not uniformly distributed. Data centres are highly geographically concentrated, and certain territories face acute constraints that threaten both AI deployment and broader energy security.

## United States

**Virginia (Northern Virginia / Data Center Alley):** The world's largest data centre hub, already consuming approximately 26% of the state's electricity [9][10]. Peak demand is expected to grow 70% by 2045, prompting a projected $90 billion grid upgrade plan [11]. PJM Interconnection capacity market prices increased tenfold from $28.92 to $329.17 per MW-day between the 2024/25 and 2026/27 planning years, directly attributable to data centre demand [5]. Northern Virginia is approaching saturation, with new projects being pushed south and west [12].

**Texas:** ERCOT reserve margins could fall into risky territory after 2028 [12]. Senate Bill 6 (effective 2025) redefines interconnection for large electrical loads exceeding 75 MW, requiring disclosure, financial commitments, and grid infrastructure cost coverage [13]. Texas data centres face substantial water stress, with projections of an 870% water demand increase by 2030 [5].

**Broader US:** Fifteen states account for approximately 80% of national data centre load [13]. US power demand is expected to reach record levels of 4,179 billion kWh in 2025 and 4,239 billion kWh in 2026 [14]. Goldman Sachs estimates approximately $720 billion in grid upgrades through 2030 [14]. Under constrained renewable scenarios, the IMF models US electricity prices increasing by up to 8.6% [15].

## Ireland and Europe

Ireland represents the most extreme case of AI-driven energy stress in Europe. CSO Ireland data: data centres consumed 21% of total national metered electricity in 2023 (up from 5% in 2015) and 22% in 2024 [16][17]. EirGrid projects this could reach 31–32% by 2026–2027 [18]. Oeko-Institut analysis (cited by Carbon Brief) estimates data centres consume approximately 79% of Dublin's electricity [10]. This figure reflects Dublin-area concentration; almost all

contracted capacity is in the Greater Dublin region [19]. The methodology has not been independently replicated.

The CRU imposed conditional grid-connection restrictions in 2021, creating a moratorium on new Dublin data centres until 2028. In December 2025, the CRU published its final decision requiring new data centres to provide dispatchable generation or battery storage matching import capacity, and to source 80% from renewables [20].

European data centre demand is projected to grow by more than 45 TWh (up 70%) through 2030 [2]. The EU Energy Efficiency Directive requires data centres above 500 kW to report detailed energy metrics. The EU AI Act mandates energy disclosure for general-purpose AI models [5].

## Asia-Pacific

China and the US account for nearly 80% of global AI electricity consumption growth to 2030 [2]. China mandates a PUE of 1.25 for large data centres. Southeast Asia is an emerging hotspot — Singapore, southern Malaysia — demand expected to more than double by 2030 [2]. In Japan, data centres could account for more than half of electricity demand growth [2]. In Malaysia, data centres could consume a fifth of total electricity growth [2].

## The Infrastructure Bottleneck

The constraint extends beyond total energy supply to the physical infrastructure to deliver it. Many regional grids cannot accommodate large-scale data centres without upgrades requiring 5–10 years for planning, permitting, and construction [13]. High-voltage transformers that once had 6-month lead times now take 3–4 years [11]. The IEA estimates approximately 20% of planned data centre projects face delay risk due to grid constraints [2].

| Territory | DC Energy Share | Key Constraint | Risk Level |
|---|---|---|---|
| Virginia, US | ~26% state | Grid saturation, 10× capacity price [5] | **CRITICAL** |
| Ireland | 22% national | CRU moratorium, 31%+ by 2027 [18] | **CRITICAL** |
| Dublin (local) | ~79% [10]* | All DC capacity concentrated here [19] | **CRITICAL** |
| Texas, US | Large, growing | Reserve margins post-2028, water [12] | **HIGH** |
| Singapore/Malaysia | Emerging hub | Grid capacity, 1/5 of growth [2] | **HIGH** |
| Japan | Growing | >50% of demand growth to DCs [2] | **HIGH** |

** Oeko-Institut estimate via Carbon Brief; methodology not independently replicated.*

# Infrastructure Dominates Outcomes

A central finding of Jegham et al. (2025) — reinforcing this report's territorial analysis — is that the same model deployed on different infrastructure produces vastly different energy, water, and carbon outcomes [26].

| Metric (long prompt) | DeepSeek-R1 (own DC) | DeepSeek-R1 (Azure) | Reduction |
|---|---|---|---|
| Energy (Wh) | 29.08 | 7.41 | **>70%** |
| Water (mL) | >200 | ~34 | **~85%** |
| Carbon ($gCO_2e$) | ~17 | ~2.5 | **~85%** |

The gap is driven by hardware generation, cooling efficiency, PUE, regional carbon intensity, and water usage effectiveness. Where a data centre is built, and on what infrastructure, determines its environmental footprint far more than which model it runs.

Jegham et al. use three environmental multipliers — Power Usage Effectiveness (PUE), Water Usage Effectiveness (WUE), and Carbon Intensity Factor (CIF) — applied per provider and region [26]. Their framework integrates public API performance data with inferred hardware configurations and Monte Carlo simulation to produce probabilistic per-query estimates.

The eco-efficiency analysis (cross-efficiency DEA) found OpenAI's smaller reasoning models (o3-mini score 0.884, o1-mini 0.836) and Anthropic's Claude 3.7 Sonnet (0.825) achieved the highest sustainability-adjusted performance. DeepSeek models on their own servers scored lowest (0.067 and 0.059) despite strong raw capability [26].

# The Investment Return Question

## The Scale of the Bet

The four largest hyperscalers are on track to spend more than $325 billion collectively in 2025 on AI infrastructure, an increase of roughly $100 billion from expectations at the start of the year [21]. Including neoclouds, sovereign clouds, and private clouds, total AI spending reaches approximately $600 billion in 2025 and could hit $1 trillion by 2027–2028 [21]. AI capex accounted for over 1 percentage point of US GDP growth in the first half of 2025, exceeding the consumer as the primary driver of economic growth [22][23].

The share of GDP devoted to AI investment is nearly a third greater than internet-related investment during the dot-com bubble [23]. The Manhattan Project was approximately 0.4% of GDP; Apollo was similar [23]. The current AI infrastructure build-out dwarfs both.

## The Circular Financing Problem

OpenAI has committed $300 billion in computing power with Oracle, averaging $60 billion per year, despite projected revenues of only $13 billion in 2025. CNBC reporting suggests Oracle

expects to “lose considerable sums of money” on this arrangement [22]. OpenAI took a 10% stake in AMD while Nvidia invested $100 billion in OpenAI. Microsoft is a major shareholder in OpenAI but also a major customer of CoreWeave, in which Nvidia holds significant equity [22].

Howard Marks at Oaktree called it “bubble-like behaviour.” Yale published “This Is How the AI Bubble Bursts” mapping the circular structure [22]. When everyone depends on everyone else, a disruption to any node cascades.

## The Enterprise Return Problem

MIT Media Lab’s NANDA project (August 2025): “Despite $30–40 billion in enterprise investment into generative AI, 95% of organisations are getting zero return” [24]. This measures deployed pilots generating revenue or operational savings. The gap between investment and measurable returns is stark, and the demand assumptions behind infrastructure investment rest on explosive adoption. If 95% of enterprise buyers aren’t seeing returns, the demand curve may be softer than supply is building for.

## The Energy-Returns Contradiction

The investment thesis assumes exponential growth in AI adoption, requiring exponential growth in energy consumption, in a world where energy infrastructure operates on decade-long build cycles. If every AI company’s growth projections materialised simultaneously, global data centre demand would far exceed the IEA’s projections. The Stargate Project alone aims at 10 gigawatts [11]. Grid infrastructure takes 5–10 years. Transformer lead times are 3–4 years. The IEA estimates 20% of planned projects are already at risk of delay [2].

## The Chip Obsolescence Factor

AI chips become technically obsolete in 10–12 months and face physical stress requiring replacement every 2.5 years [23]. As one analysis observed, those who compare data centre buildouts to railroads should consider how railroads would have developed if track gauge changed every year [23]. Capital deployed today may need replacement before generating returns.

## Counterarguments and Mitigating Factors

JPMorgan argues the sector does not meet classic bubble criteria [24]. BlackRock notes the S&P 500 IT Index trades at 30× forward earnings, well below the 55× at dot-com peak [25]. Jerome Powell has argued AI capex is a genuine engine of economic growth [24].

Big Tech balance sheets are substantially stronger than dot-com era companies. The risk concentrates in the non-hyperscaler layer: OpenAI, xAI, CoreWeave, and Oracle are taking on significant debt-financed commitments [22].

# Carbon and Paris Alignment: Reporting Without Greenwash

The IEA estimates data centre emissions will reach approximately 1% of global $CO_2$ by 2030, growing from 220 Mt in 2024 to 300–320 Mt by 2035 [2][10]. Data centres are one of only three sectors where emissions are set to grow — alongside road transport and aviation [10].

Harvard's T.H. Chan School found data centre electricity is 48% more carbon-intensive than the US average, because they draw 24/7 baseload power [4].

**Per-query carbon benchmarks:** Jegham et al. quantify per-query carbon using regional CIF. The most efficient models emit less than 0.3 $gCO_2e$ per long prompt; DeepSeek-R1 on its own servers emits ~17 $gCO_2e$ [26]. GPT-4o inference alone generates 138,000–163,000 tonnes $CO_2e$ annually — comparable to 30,000 petrol cars [26].

## Paris-Aligned Reporting Framework

This report recommends AI deployment scenarios disclose:

- kWh per 1,000 inferences: Standardised energy intensity, with model architecture and hardware specifications.
- Location-based AND market-based $kgCO_2e$: Both figures, side by side, to prevent greenwash through RECs.
- Carbon budget benchmarked to 1.5°C pathway: Year-by-year trajectory, not vague 2050 promises.
- Absolute emissions targets alongside intensity targets.

**Jegham et al.'s framework aligns directly with these recommendations** by providing per-prompt energy, water, and carbon metrics with explicit PUE/WUE/CIF multipliers and Scope 2 boundary definitions [26]. Their live Power BI dashboard offers a model for continuous, transparent reporting.

## Jevons Paradox Mitigation

Efficiency improvements in AI compute have historically been absorbed by increased demand rather than reducing total energy consumption. Governance rule: efficiency gains count toward climate targets only if total facility emissions decrease year-on-year. Where efficiency gains enable expanded capacity, additional emissions must be separately accounted for and offset.

Jegham et al. confirm this empirically: "as AI becomes cheaper and faster, total usage expands far more rapidly, amplifying net resource consumption" [26]. Their GPT-4o case study demonstrates that at 0.42 Wh per query, 772 billion annual queries produce a cumulative burden that overwhelms per-query efficiency gains.

# Conclusions and Implications

- Energy and grid delivery are increasingly binding constraints in key regions. The physical inability to deliver enough electricity fast enough will determine which AI companies succeed and which face stranded-asset risk.
- Not every AI company can win the energy race. Aggregate energy demands implied by collective investment theses exceed what is physically deliverable in the timeframes assumed.
- Territorial energy inequalities will create winners and losers. Regions with grid constraints face hard limits. Companies with early energy infrastructure access hold a decisive competitive advantage.
- The circular financing structure amplifies systemic risk. When AI companies are simultaneously each other's investors, customers, and suppliers, energy constraints could trigger cascading disruptions.
- The enterprise return gap demands attention. If 95% of enterprise AI investment is not generating measurable returns, demand growth assumptions may be more fragile than assumed.
- Sovereign and edge AI architectures may prove prescient. Systems designed for energy efficiency and edge deployment are better positioned for a world with hard constraints.
- Carbon reporting must separate energy from emissions. Without Paris-aligned disclosure, claims of "sustainable AI" risk greenwash.
- **Infrastructure choice is an environmental policy lever.** Jegham et al.'s finding that the same model on different infrastructure produces 70–85% variation in energy, water, and carbon outcomes means data centre siting, hardware procurement, and cooling design are environmental decisions of the first order [26].

The current AI investment trajectory assumes a world with essentially unlimited energy available on demand. That world does not exist. The question is not whether energy will constrain AI growth in key regions, but when — and which companies and territories will be caught unprepared.

# References


[1] World Economic Forum, "The AI-Energy Nexus Will Dictate AI's Future," December 2025.

[2] IEA, Energy and AI, 2025. iea.org/reports/energy-and-ai.

[3] IEA Global Energy Review 2025; AllAboutAI aggregation. AI workloads 11–20% from FlowingData (2025).

[4] Sam Altman ~0.34 Wh/query claim; Harvard T.H. Chan School carbon intensity preprint, May 2025.

[5] Medium/aggregator estimates (Asrar Jan 2026; AllAboutAI Dec 2025). o3 energy: third-party, unverified.

[6] Harvard Gazette, "Should U.S. Be Worried About AI Bubble?" Andy Wu interview, December 2025.

[7] GPT-4 training cost: estimated $50–100M (BentoML, Epoch AI). Exact cost/energy undisclosed.

[8] DeepSeek-V3 Technical Report (arXiv:2412.19437v1). 2.788M H800 GPU hours, $5.576M.

[9] PEI/Power Magazine, “Power Generation in the Age of AI,” December 2025.

[10] Carbon Brief, “AI: Five Charts,” September 2025. Dublin 79% attributed to Oeko-Institut.

[11] CSG Talent, “Data Centre Grid Challenges.” Northern Virginia $90B upgrade, transformer lead times.

[12] BloombergNEF, “AI and the Power Grid,” December 2025.

[13] arXiv:2509.07218v2, “Electricity Demand and Grid Impacts of AI Data Centers,” September 2025.

[14] US EIA via Morgan Lewis, February 2025; Goldman Sachs $720B grid upgrade estimate.

[15] IMF WP/2025/081, “Power Hungry: How AI Will Drive Energy Demand,” April 2025.

[16] CSO Ireland, Data Centres Metered Electricity 2023, July 2024. 21% share.

[17] CSO Ireland, Data Centres Metered Electricity 2024, June 2025. 22% share.

[18] EirGrid forecasts: 31–32% by 2026–2027.

[19] EirGrid Capacity Outlook 2022–2031; Irish Government Statement on Data Centres, July 2022.

[20] CRU Final Decision on large energy user connections, December 2025.

[21] TCW, “AI Investment Potential Accelerates,” November 2025.

[22] Yale Insights, “This Is How the AI Bubble Bursts,” October 2025.

[23] GARP, “AI Bubble or Boom?” November 2025.

[24] MIT NANDA “The GenAI Divide,” August 2025; Bank of England warnings; Altman/Dalio quotes.

[25] BlackRock, “Are We in a Bubble?” October 2025. S&P 500 IT Index 30× vs 55× dot-com peak.

[26] Jegham, N., Abdelatti, M., Koh, C.Y., Elmoubarki, L., & Hendawi, A. (2025). “How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference.” arXiv:2505.09598v6, November 2025. University of Rhode Island / University of Tunis / Providence College.

# Methodological Appendix: Data Sources and Estimation Practices

This report draws on three tiers of evidence. Key figures are classified to help readers interpret the numbers and avoid over-claiming precision.

## Tier 1: Primary Statistical Agencies and Official System Operators

Hard data subject to normal revision risk: national electricity shares from statistical offices (CSO Ireland 2015–2024), grid forecasts from system operators (EirGrid, US EIA), regulatory decisions from energy regulators (CRU connection conditions).

## Tier 2: Multilateral Agencies, Central-Scenario Models, and Sector-Wide Surveys

Model-based estimates with documented methods: IEA scenarios (415 TWh → ~945 TWh, $CO_2$ pathways), IMF electricity price projections, industry surveys (Goldman Sachs, Deloitte, BloombergNEF).

## Tier 3: Reconstructed, Inferred, or Single-Study Estimates

Order-of-magnitude indicators, explicitly flagged: per-query energy for frontier models (undisclosed methodologies), GPT-4 training costs (analyst inferences), Dublin 79% (single-institute, unreplicated), water-use projections (sensitive to siting/cooling assumptions).

## New: Infrastructure-Aware Benchmarking (Jegham et al., 2025)

This revision incorporates Jegham et al. [26], an infrastructure-aware benchmarking framework that:

- Integrates public API performance data (latency, tokens-per-second) with inferred hardware configurations and environmental multipliers (PUE, WUE, CIF).
- Uses Monte Carlo simulation (10,000 correlated samples per model) for probabilistic per-query estimates.
- Focuses on Scope 2 operational emissions, excluding Scope 3 embodied emissions.
- Validates against disclosures: GPT-4o estimate within 19% of Altman's figure; Mistral Large 2 within one standard deviation of Mistral's published LCA.

Classification: between Tier 2 and Tier 3. More rigorous than Fermi estimates but reliant on inferred hardware configurations. Use as best-available benchmarks, not ground truth.

## Interpretation Guideline

- Tier 1: Point estimates anchored in observed data (with normal statistical uncertainty).
- Tier 2: Scenario outcomes, useful for direction and scale, not precise forecasts.
- Tier 3: Indicative ranges only. Quarantined with caveats. Not for fine-grained calibration.
- Jegham et al. benchmarks: Best available inference-level data, with the caveat that hardware configurations are inferred rather than disclosed.