Energy Use per AI Inference Task: Joules and Watt-Hours

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · ai inference energy · joules per token · llm energy consumption · gpu power consumption · data center energy use · power usage effectiveness · mlperf power benchmark · ai energy efficiency

Executive Summary

Energy use per completed AI inference task has no single agreed-upon figure, because published estimates depend on which part of the system is measured: the GPU chip alone, the full server node, or the entire data center including cooling overhead. A 2026 Microsoft Research study published in the journal *Joule* used a bottom-up framework to estimate a median energy use of 0.31 watt-hours (Wh) per typical chatbot query, rising roughly thirteenfold to about 3.91 Wh for long, reasoning-style responses (Joule). An academic API-data-based estimate put a short GPT-4o prompt at 0.42 Wh, alongside OpenAI's self-disclosed estimate of 0.34 Wh per average ChatGPT query and Epoch AI's independent, first-principles estimate of roughly 0.3 Wh (arxiv.org) (blog.samaltman.com) (epoch.ai). Google has published two figures for a median Gemini prompt, a comprehensive 0.24 Wh estimate and a narrower, chip-only 0.10 Wh figure (cloud.google.com).

Hardware sets the outer bound on these numbers. NVIDIA specifications show accelerator thermal design power climbing from 300 to 400 watts on the A100 and H100 PCIe generations to up to 700 watts on H100 SXM modules and H200 SXM modules and up to 1,000 watts per GPU on the Blackwell B200 (nvidia.com). Facility overhead adds a further layer: Google discloses a global fleet power usage effectiveness (PUE) of 1.09 and Microsoft reports 1.16 to 1.17, both below the broader industry average of 1.54, so identical GPU workloads carry different total footprints depending on the host operator (google.com) (microsoft.com). The International Energy Agency estimates data centers consumed 415 terawatt-hours (TWh) of electricity in 2024, about 1.5% of world electricity use, projected to more than double to around 945 TWh by 2030 as AI-optimized servers grow electricity draw roughly 30% annually (iea.org) (iea.org).

This report documents eight measurable variables that each independently swing per-task energy by 25% or more: model size, architecture, numeric precision, batch size, context length, output length, parallelism strategy, and GPU generation. Reasoning models generating far more output tokens per response are the largest driver of higher energy per task, while quantization, larger batch sizes, and newer GPU generations are the most effective reduction levers. Standardized, audited disclosure remains immature: MLPerf Power's wall-outlet methodology and Hugging Face's fixed-batch AI Energy Score are the most rigorous public benchmarking efforts to date, but neither yet publishes a routine, segment-broken-out joules-per-task figure comparable across vendors.

Any energy-per-inference claim is only as trustworthy as its stated hardware, measurement instrument, system boundary, and task definition. IntuitionLabs describes its work as providing AI solutions designed specifically for pharmaceutical and life science organizations (intuitionlabs.ai).

Introduction and Background

Every prompt submitted to a large language model completes a measurable chain of physical work: matrix multiplications executed on accelerator chips, data moved across high bandwidth memory, and heat removed by a data center's cooling plant. Quantifying that chain in **joules** (the SI unit of energy, abbreviated J) or in **watt-hours** (Wh, equal to 3,600 joules) has become a recurring technical and policy question as generative AI usage scales toward billions of daily queries. There is no single, universally reported "energy per inference" figure. Published estimates for one chatbot query span more than two orders of magnitude, from roughly 0.1 Wh to tens of watt-hours, and the gap comes down to what each estimate actually includes rather than a real disagreement about physics.

A 2026 peer-reviewed analysis from Microsoft Research, published in the journal *Joule*, found that widely cited public energy-per-query figures are overstated by four to twenty times relative to production-optimized deployments (arxiv.org). Independent academic teams at Carnegie Mellon and the University of Michigan have separately shown that even the choice of measurement instrument, batch size, or output length can swing a reported "joules per token" number by several-fold for the exact same model and GPU, as detailed in the methodology and benchmark sections below. Meanwhile OpenAI, Google, and Mistral AI have each published vendor-side estimates that differ by roughly threefold for a comparable text prompt, a divergence examined in detail later in this report.

This report is organized as a **reproducible measurement protocol**, not a single averaged number. It separates the questions that a rigorous energy-per-task figure must answer before it means anything: What hardware power envelope is being measured, the GPU package alone or the full server node? Is idle and non-inference overhead subtracted out? Is data center cooling and power delivery loss (captured in the **power usage effectiveness**, or

PUE, ratio) included or excluded? And was the task actually completed, or does the measurement include failed or retried requests? As of September 2026, no regulator or standards body mandates disclosure of per-query energy figures, so the most defensible answers come from peer-reviewed measurement studies, official hardware specifications, and clearly labeled vendor disclosures, each addressed in turn below alongside the secondary questions readers of this report most often ask: how much energy AI inference uses in aggregate, how that energy is measured, what benchmarks exist, and what drives energy cost per completed task up or down.

Measurement Methodology and System Boundaries

A defensible energy-per-inference measurement requires explicit answers to four boundary questions, because each one can change the reported number by a large multiple without any error in the underlying physics.

- **Instrument and sampling boundary.** NVIDIA's own driver-level interface, the NVIDIA Management Library (NVML), exposes GPU power through a call that "retrieves power usage for this GPU in milliwatts and its associated circuitry" such as onboard memory, meaning it reports total board power rather than compute-die power alone (docs.nvidia.com). On Ampere-generation GPUs and newer, that same call "returns power averaged over 1 sec interval," which is too coarse to resolve the brief power spikes of a single short decode step, a limitation that led the Carnegie Mellon team to instead read NVML's cumulative energy counter, which "reports cumulative energy consumption in mJ" and derives energy from the difference between two synchronized readings rather than a sampled average (docs.nvidia.com) (arxiv.org).
- **Idle versus active power boundary.** A GPU under an inference server draws non-zero power even between requests (fan, memory refresh, unused compute units), so a measurement that fails to subtract idle draw over-attributes energy to the request it happens to be measuring.
- **System scope boundary.** Open-source tools deliberately narrow their scope for tractability. CodeCarbon, a widely used Python energy-tracking library, computes an emissions figure as measured or estimated energy multiplied by the local electricity grid's carbon intensity, stating that "carbon dioxide emissions" can "then be calculated as C * E" (codecarbon.io). The tool states plainly that it excludes disk, network, and display power because "those sources are usually much smaller, and often negligible, for local code-level experiments," and it samples power at a default interval of 15 seconds using the same underlying NVIDIA library, noting it "tracks Nvidia GPUs energy consumption using nvidia-ml-py library" (codecarbon.io) (codecarbon.io) (codecarbon.io). It also measures GPU power "at the device level," so a shared GPU's full power draw is attributed to any one process running on it rather than split proportionally (codecarbon.io). The University of Michigan's Zeus framework, introduced at USENIX NSDI 2023 for "Understanding and Optimizing" GPU energy consumption of deep learning workloads, has since been extended to cover CPU, DRAM, AMD GPUs, Apple Silicon, and NVIDIA Jetson devices, reflecting how much of the measurement burden falls outside the GPU alone (github.com) (github.com).
- **Facility overhead boundary.** Even a perfect GPU-level joule count does not yield a whole-facility figure. To estimate whole-facility energy, first measure or estimate the full IT-equipment energy attributable to the task, then apply a PUE measured for the relevant facility and period; a GPU-only result should remain labeled GPU-only (U.S. Department of Energy).

For standardized, comparable benchmarking, the MLCommons consortium's **MLPerf Power** methodology sets the strictest boundary of the tools surveyed here: it specifies that "power is measured at the system level ("at the wall")," using a power analyzer that "must be located between the Line Voltage Source and the SUT [system under

test]" (github.com) (spec.org). MLPerf built this methodology on the Standard Performance Evaluation Corporation's PTDaemon interface, which "translates vendor-specific command and hardware interfaces into a consistent API" so that different power analyzer models produce comparable readings, under a required "overall uncertainty of less than 1% (AC) and 1.5% (DC)" and an explicit rule that "the use of automatic range support should be forbidden" during a compliant run because range-switching degrades measurement accuracy (mlcommons.org) (spec.org) (spec.org) (spec.org). A compliant MLPerf Power submission requires "a power analyzer (anyone certified by SPEC PTDaemon)," most commonly a Yokogawa unit (mlcommons.org). MLPerf Inference v1.0 in 2021 was the first round to include this optional measurement, releasing 864 power results alongside performance scores (mlcommons.org), and MLCommons' Power working group describes its purpose as being to "create power measurement techniques for various MLPerf benchmarks that enable reporting and comparing energy consumption" across submitters on a consistent basis (mlcommons.org).

The practical takeaway for any reader trying to reproduce or sanity-check a published figure: ask whether the number is (a) GPU-package power only, (b) full-server power, or (c) full-facility power including cooling and PUE, because a defensible comparison can only be made within one of these three tiers, not across them.

Hardware Power Envelopes: GPUs, Nodes, and Data Center Overhead

The starting point for any bottom-up energy estimate is the accelerator's rated **thermal design power (TDP)**, the maximum sustained power draw the chip and its cooling solution are engineered to dissipate. NVIDIA's own product pages and datasheets, the Tier 1 source for this figure, show a clear generational and form-factor spread. The **H100 SXM** module is rated "up to 700W (configurable)" while the PCIe card variant is capped at 350 to 400W (nvidia.com); NVIDIA's H100 PCIe product brief specifies the card "operates unconstrained up to a maximum thermal design power (TDP) level of 350 W" and the H100 NVL brief lists "400 W" for that variant (nvidia.com) (nvidia.com). The successor **H200 NVL** card is rated "up to 600W," while NVIDIA states the H200 SXM delivers its performance gains "all within the same power profile as the H100," i.e., no TDP increase for the top-end module (nvidia.com) (nvidia.com). The older **A100** is lower-power across the board, with NVIDIA's datasheet specifying "400W TDP for standard configuration" on the SXM4 module and 300W for the PCIe card (nvidia.com). The **B200** (Blackwell) raises the ceiling substantially: NVIDIA's product carbon footprint summary lists each "individual GPU: Configurable up to 1000 W" (nvidia.com).

Table 1 below summarizes these official per-GPU TDP figures alongside the node-level assumptions used by the Microsoft Research energy study discussed in the next section.

GPU (NVIDIA)	Form Factor	Max Configurable TDP	Source
A100	SXM4	400W (standard), up to 500W (CTS variant)	NVIDIA A100 datasheet (nvidia.com)
H100	PCIe	350 to 400W	NVIDIA H100 product brief (nvidia.com)
H100	SXM	Up to 700W	NVIDIA H100 product page (nvidia.com)
H200	NVL	Up to 600W	NVIDIA H200 product page (nvidia.com)

GPU (NVIDIA)	Form Factor	Max Configurable TDP	Source
H200	SXM	Up to 700W (same envelope as H100 SXM)	NVIDIA H200 product page (nvidia.com)
B200	HGX	Configurable up to 1000W per GPU	NVIDIA HGX B200 carbon footprint summary (nvidia.com)

A single GPU's TDP is only the starting point because inference at scale runs on multi-GPU nodes with their own power ceiling. The Microsoft Research study models frontier-scale models as "hosted on 8xH100 GPUs at FP8 precision, as in the NVIDIA DGX H100 architecture," a configuration with a maximum node power draw in the 10 kilowatt (kW) class, and notes some very large models require a 10-GPU node instead (arxiv.org). At rack scale, NVIDIA markets its **GB200 NVL72** system, which "uses NVLink and liquid cooling to create a single massive 72-GPU rack" and claims this liquid-cooled architecture "delivers 25x more performance at the same power" as an air-cooled H100 generation of infrastructure (nvidia.com) (nvidia.com).

Above the rack, the facility itself adds a further overhead layer captured by PUE. The International Energy Agency (IEA) notes that within a modern data center, servers "account for around 60% of electricity demand," with the remainder split between cooling, power distribution, and other overhead, and that cooling's share specifically "varies from about 7% for efficient hyperscale data centres to over 30% for less-efficient enterprise data centres" (iea.org) (iea.org). This is precisely why liquid cooling, as used in the GB200 NVL72, is being adopted for the highest-density AI racks: air cooling becomes proportionally less efficient as per-rack power density rises into the 100kW-plus range the IEA associates with modern AI-focused facilities, which it distinguishes from "a conventional data centre" that "may be around 10-25 megawatts" in total size (iea.org).

Joules and Watt-Hours per Completed Task: Benchmark Findings

With hardware power envelopes established, the central question becomes: how many joules does a completed inference task actually consume once GPU utilization, output length, and serving configuration are accounted for? The most rigorous public answer comes from the Microsoft Research and *Joule* journal study, which built a bottom-up model using measured throughput data for five open models (including Llama 3.1 405B and DeepSeek-R1 671B) on 8-GPU H100 nodes. It found that for frontier-scale models above 200 billion parameters, "we estimate a median energy of 0.31 Wh/query" for a typical chatbot-style exchange, with the true PUE-inclusive figure modeled "as a log-normal distribution with P5 [to] P95 range of 1.05 to 1.40," consistent with the hyperscaler-disclosed PUE figures discussed in the next section (Joule) (arxiv.org). Critically, that median energy is not fixed: for reasoning-style "test-time scaling" queries with outputs roughly 15 times longer than a typical chat response, "the median energy rises 13" fold to about 3.91 Wh per query (Joule). Scaled to production volume, the same model estimates that serving one billion queries per day under a naive baseline configuration "leads to 0.73 GWh" of daily energy, a figure the authors show can fall by more than half under conservative efficiency improvements (arxiv.org).

Independent, hardware-level measurements corroborate both the scale of these numbers and their sensitivity to configuration. A Carnegie Mellon University study measured a small Llama-3.2-1B model on an H100 SXM GPU (rated, per NVIDIA, at "700 W nominal TDP") and an H200, and found that "increasing output length from 10 to 512 tokens reduces token energy from 7.46 to 0.72 J/token" even as the total energy consumed by the batch rises, illustrating that per-token efficiency and total energy consumption can move in opposite directions (arxiv.org)

(arxiv.org). At the University of Michigan, the **ML.ENERGY** research initiative benchmarked 46 generative models across seven task types in 1,858 configurations and found that "LLM task type can lead to 25" fold energy differences for the same underlying model, driven largely by how many tokens a task requires the model to generate (arxiv.org). On newer hardware, the same group found the B200 GPU delivers a "median 35% energy reduction" relative to H100 at matched latency across the majority of LLM workloads tested (arxiv.org).

Two independent academic groups have also built analytical, first-principles models that avoid the need for physical hardware access. A study titled "From Tokens to Watt-hours" derived per-operation energy coefficients "based on microarchitectural accelerator-energy measurements" for H100-class hardware and used them to estimate that a 120-billion-parameter open model "reaches 374.4 mJ/output token" (arxiv.org) (arxiv.org). Validated against physically measured results, the estimator "matches measurement-based values within approximately 5" to 27 percent, but the authors are explicit that their model excludes facility-level cooling and power usage effectiveness overhead, scoping the estimate to GPU compute and memory movement only, not the facility overhead layer (arxiv.org). Separately, University of Rhode Island and University of Tunis researchers estimated per-query energy for major commercial models from public API latency and throughput data, finding that "a single short GPT-4o query consumes 0.42 Wh," and that the most efficient tested model, "GPT-4.1 nano remains the most efficient overall," while o3, a reasoning model, "consumes 39.223 Wh" for a long, 10,000-input-token prompt, over seventy times the smallest model's footprint (arxiv.org) (arxiv.org) (arxiv.org). That same group found GPT-4o's own energy cost triples for long prompts, using "1.788 Wh for long prompts and 0.42 Wh for short ones," underscoring that prompt length, not just model choice, is a first-order driver of per-task energy (arxiv.org).

Task type outside of text generation shows similarly wide swings. A 2026 diffusion-model energy scaling study found that "a single diffusion image can consume up to 10" times the energy of an average ChatGPT-style text query, illustrating that image-generation inference is not directly comparable to text inference on a per-request basis (arxiv.org). None of these figures should be read as "the" energy cost of AI inference; they are point measurements under stated hardware, model, and prompt-length assumptions, and the single most important methodological finding across this literature is that **any figure quoted without those three variables attached is not reproducible**.

Analysis of Key Segments: Task Type, Model Size, and Hardware Generation

Breaking the aggregate numbers down by segment clarifies which levers actually move energy per task, and by how much. The **AAAI-26 TokenPowerBench study** is the most granular public source on this question, benchmarking energy across model families, serving engines, and configurations, and it opens with an industry-sourced framing that matters for scale: "inference, not training, accounts for more than 90% of total power consumption" in production large language model services, meaning the levers below apply to the majority of the energy budget, not a training-time footnote (ojs.aaai.org).

- **Model size.** TokenPowerBench found that scaling a LLaMA-3 family model from 1 billion to 70 billion parameters "increases energy per token by 7.3" fold, a markedly sub-linear relationship given the 70-fold increase in parameter count, meaning larger models are, per parameter, more energy-efficient at inference than smaller ones, even though their absolute per-token cost is higher (ojs.aaai.org).

- **Model architecture.** Sparse mixture-of-experts models change this curve. The same benchmark found that Mixtral 8x7B, a sparse MoE model, "cuts token energy by 2" to 3-fold compared with dense models of similar output quality, because only a subset of its parameters activate per token (ojs.aaai.org).
- **Serving engine.** Holding the model and GPUs fixed, TokenPowerBench found that production-grade serving engines "reduce energy per token by 25" to 40 percent relative to a baseline Transformers engine on the same 4x H100 setup, a purely software-level efficiency gain (ojs.aaai.org).
- **Numeric precision.** Quantizing a 405-billion-parameter Llama 3 model from FP16 to FP8 precision was found to cut per-batch "total energy" from about 45 kilojoules (kJ) "to 32 kJ under the heaviest load," roughly a 30 percent reduction, on the same H100 cluster (ojs.aaai.org). This benefit is not automatic across all quantization approaches; precision reduction interacts with batch size and, applied naively, can offset or reverse the expected saving.
- **Batch size.** TokenPowerBench found that increasing batch size from 32 to 256 "cuts per-token energy by about 25% in this range" for a 70-billion-parameter model, as fixed per-request overhead is amortized across more concurrent requests (ojs.aaai.org).
- **Context length.** The same study found that growing the input context from 2,000 to 10,000 tokens "raises energy per token by roughly a factor of three" for a 70-billion-parameter model, since a longer prompt increases the computationally expensive prefill stage relative to token-by-token decoding (ojs.aaai.org).
- **Parallelism strategy.** At 16-GPU scale, the gap between the best and worst GPU-splitting configuration "grows from about 40 J/token under Standard Load to more than 60 J/token under High Throughput" for state-of-the-art models, showing that parallelism choices matter more, not less, as load increases (ojs.aaai.org).
- **Reasoning versus standard chat.** The University of Michigan's ML.ENERGY study found that reasoning models "produce one to two orders of magnitude more output tokens per request compared to standard chat" models on identical benchmarks, which is the single largest identified driver of energy variance between model types (arxiv.org).

These segment-level findings collectively explain why single, undifferentiated "energy per AI query" statistics circulating in public discussion are so inconsistent: they typically collapse several of the eight independent variables above into one number.

Vendor Disclosures Versus Independent Measurement

Public discussion of AI energy use draws on two categorically different kinds of source, and this report treats them separately because they carry different evidentiary weight. **Vendor claims** are self-reported figures published by the company that operates the model, typically without a disclosed methodology open to replication. **Independent estimates** come from academic or research groups modeling the same systems from published specifications and third-party measurements, without insider access to the vendor's actual production telemetry. Each category has distinct evidentiary limits: independent estimates are transparent about method but must approximate configuration details vendors do not publish.

On the vendor side, OpenAI co-founder and CEO Sam Altman wrote on his personal blog that "the average query uses about 0.34 watt-hours" of electricity on ChatGPT, alongside a water-use estimate of "roughly one fifteenth of a teaspoon" per query (blog.samaltman.com) (blog.samaltman.com). Google has published a more detailed, though still self-reported, disclosure: its comprehensive accounting states "the median Gemini Apps text prompt uses 0.24

watt-hours" of energy, while a narrower calculation that counts only active AI-chip power puts "the median Gemini text prompt" at "0.10 Wh." (cloud.google.com) (cloud.google.com). Mistral AI has taken a different disclosure approach, publishing carbon and water figures, a 400-token Le Chat response uses "45 mL of water" and an associated carbon footprint, and stating this lifecycle assessment was "peer-reviewed by Resilio and Hubblo," two named environmental-audit consultancies (mistral.ai) (mistral.ai).

On the independent side, the research organization Epoch AI built a bottom-up estimate from GPU specifications and public utilization data and concluded that "typical ChatGPT queries using GPT-4o likely consume roughly 0.3 watt-hours," a figure close to both OpenAI's own 0.34 Wh claim and the Jegham et al. academic API-data/statistical estimate of 0.42 Wh discussed in the previous section (epoch.ai). Epoch AI also directly addressed a widely circulated older estimate of 3 Wh per query, concluding "this figure of 3 watt-hours per query is likely an overestimate" by roughly an order of magnitude given more efficient current-generation hardware, though it noted that a sufficiently long input, around 10,000 tokens, would "increase the cost per query to around 2.5 watt-hours," approaching that older figure under a different, longer-prompt assumption (epoch.ai) (epoch.ai). These estimates are not a standardized comparison and should not be treated as convergent evidence for a universal typical-query range, because products, models, workloads, and measurement approaches differ (Microsoft Research).

Data Center Demand and PUE

Zooming out from individual queries to system-wide scale clarifies why per-task efficiency matters. The IEA's *Energy and AI* report estimates that data centers worldwide "accounted for around 1.5% of the world's electricity consumption in 2024, or 415 terawatt-hours," and projects that consumption is "set to more than double to around 945 TWh by 2030" (iea.org) (iea.org). Looking further out, the IEA's Base Case "sees global data centre electricity consumption rising to around 1 200 TWh by 2035," while cautioning there is "substantial uncertainty both about data centre consumption today and in the future," reflected in a stated scenario range of roughly 700 to 1,700 TWh for that year (iea.org) (iea.org). For scale, the IEA notes that "a typical AI-focused data centre consumes as much electricity as 100 000 households," and it separately projects that electricity use in AI-optimized "accelerated servers" will grow "by 30% annually in the Base Case," far outpacing the low single digits projected for conventional enterprise servers (iea.org) (iea.org).

Table 2 summarizes the disclosed PUE figures from two hyperscale operators, Google and Microsoft, alongside the Uptime Institute's industry-survey benchmark.

Operator / Source	Disclosed PUE	Definition Basis	As Of
Google (global fleet, trailing 12 months)	1.09	"compares the amount of non-computing overhead energy" to IT energy (google.com)	2025 (google.com)
Microsoft (global fleet, owned/controlled facilities)	1.16 (FY24) to 1.17 (FY25)	"dividing the total energy needed for a datacenter facility" by IT energy (microsoft.com)	FY2024 to FY2025
Uptime Institute (industry-wide survey average, cited by Google)	1.54	Global average across surveyed operators (google.com)	2025

Water is a related, less commonly disclosed overhead metric. Microsoft defines its water usage effectiveness (WUE) figure as "dividing the annual liters of water used for humidification and cooling" by the annual electricity used to power IT equipment, mirroring the PUE ratio but for water rather than energy overhead (microsoft.com). The gap between hyperscaler-disclosed PUE (roughly 1.09 to 1.17) and the broader industry average (1.54) matters directly for any per-query energy estimate: applying the industry-average multiplier instead of a leading hyperscaler's figure to the same raw GPU energy number can add on the order of 30 to 40 percent to a "wall-to-wall" energy-per-query estimate, purely from facility overhead. This is consistent with the Uptime Institute's own finding, in its 2025 Global Data Center Survey, that "average PUE levels show little change for the sixth consecutive year" industry-wide, even as the same survey found "roughly one-third of data center owners and operators currently perform some AI training or inference" workloads, meaning AI-specific efficiency gains are concentrated in a subset of newer, hyperscale-operated facilities rather than the installed base as a whole (uptimeinstitute.com) (uptimeinstitute.com).

Standardized third-party benchmarking of model-level energy remains comparatively immature relative to performance benchmarking. Hugging Face's **AI Energy Score** project, one of the few public, model-comparable leaderboards, states that "all benchmarks are conducted exclusively on NVIDIA H100 GPUs" and rates models using star bands formed by "dividing the GPU energy range for a specific task into five equal sections" (huggingface.github.io) (huggingface.github.io). Its methodology fixes batch size to one request at a time, a simplification the University of Michigan researchers explicitly critique, noting this approach "fixes the inference batch size to 1 for all models," which does not reflect how production services actually batch concurrent requests to improve efficiency, meaning single-model energy leaderboards should be read as relative rankings under a fixed synthetic condition rather than production-representative figures (arxiv.org). That same research group also found that estimating energy from a GPU's rated TDP, rather than measuring it directly, is unreliable at the individual model level: doing so produced a "worst-case overestimation of energy consumption by a factor of 4.1" for one small model tested on H100 hardware, which is a caution against any energy claim built purely from a hardware spec sheet rather than an actual measurement (arxiv.org).

Implications and Future Directions

Several structural trends will shape how energy-per-task figures evolve over the next several years. First, hardware efficiency continues to improve within a roughly constant power envelope: NVIDIA states its H200 accelerator delivers its performance gains "all within the same power profile as the H100," and the University of Michigan's B200 comparison, cited earlier in this report, found a similar generational efficiency gain at matched latency, indicating that per-task energy is falling even as raw GPU wattage plateaus at the high end (nvidia.com). Second, the shift toward **reasoning models** that generate substantially longer chains of intermediate output works in the opposite direction, as the test-time-scaling and reasoning-token findings discussed earlier in this report both show. Whether aggregate AI energy demand rises or falls over time depends heavily on which of these two forces, per-token hardware efficiency or growing per-response token counts, dominates in practice, and the evidence surveyed here suggests both are operating simultaneously.

Third, measurement standardization itself is still developing. MLPerf Power's system-level, wall-outlet methodology and Hugging Face's model-level AI Energy Score represent two different, currently non-interoperable approaches (system throughput-normalized power versus fixed-batch per-model energy), and neither yet publishes a routine, audited joules-per-completed-task figure broken out by the eight segment variables identified in this report

(github.com) (huggingface.github.io). Enterprises evaluating AI vendors, including life-sciences organizations weighing AI-assisted drug discovery, regulatory, or commercial platforms against sustainability and total-cost-of-ownership goals, currently have no standardized per-task disclosure format to request. IntuitionLabs, which identifies itself as an AI software development company focused on the pharmaceutical and life-science industry (intuitionlabs.ai), notes on its own site that it works with life-science organizations to evaluate "cutting-edge AI solutions designed specifically for pharmaceutical and life science organizations," a category of engagement in which energy transparency is one of several operational due-diligence questions, alongside compliance and validation, that a life-sciences buyer should expect a software or AI vendor to be able to answer with a stated methodology rather than a marketing figure (intuitionlabs.ai).

Finally, the tooling gap between research-grade measurement (NVML counters, Zeus, controlled GPU-level studies) and production-scale, facility-inclusive reporting (PUE-adjusted, MLPerf Power-style wall measurement) is likely to narrow as more vendors adopt SPEC PTDaemon-class methodology for inference-specific power reporting, following the path MLPerf Inference took for training and classic inference benchmarks in 2021 (mlcommons.org). Readers seeking to reproduce or audit a specific vendor's per-query energy claim should, per the methodology outlined earlier in this report, ask for the GPU model and node configuration, the measurement instrument (sampled power versus cumulative energy counter), the PUE assumption applied, and whether the reported figure reflects a short or long prompt and a standard or reasoning-mode response, since each of these can independently move the final number by a factor of two or more.

Conclusion

Energy use per AI inference task is not a single number but a function of at least eight identifiable variables: model size and architecture, numeric precision, batch size, context length, output length, parallelism strategy, GPU generation, and data center overhead. Published values near 0.3 to 0.4 watt-hours are non-standardized point estimates, not a universal typical-query range, because products, models, workloads, and measurement approaches differ (Microsoft Research), while reasoning-mode queries with substantially longer outputs can consume ten times that amount or more. Hardware thermal design power, ranging from 300 to 400 watts on older GPU generations up to 1,000 watts on the Blackwell B200, sets the ceiling; actual per-task energy depends on how efficiently that ceiling is utilized through batching, quantization, and serving-engine choice, all of which this report's cited studies show can each independently swing per-token energy by 25 percent or more. At the facility level, disclosed power usage effectiveness ratios from leading hyperscale operators (roughly 1.09 to 1.17) remain meaningfully below the broader industry average of 1.54, meaning the same GPU-level workload can carry a substantially different total energy footprint depending on where and by whom it is hosted. A reproducible energy claim, whether made by a vendor, a researcher, or an enterprise evaluating an AI deployment, must state its hardware, its measurement instrument, its system boundary, and its task definition; absent those four disclosures, a reported energy-per-inference figure cannot be verified or meaningfully compared to any other.

Frequently Asked Questions (FAQs)

- **How much energy does AI inference use per query?** As detailed in this report's benchmark findings above, peer-reviewed bottom-up model estimates place a typical chatbot query on H100-class hardware at roughly 0.3 to 0.4 Wh, rising to several watt-hours for long, reasoning-style responses, alongside OpenAI's own disclosed 0.34 Wh and Epoch AI's independent 0.3 Wh estimate (blog.samaltman.com) (epoch.ai).

- **What is a joule per inference and how is it measured for LLMs?** A joule is the base SI unit of energy; researchers typically report either joules per generated token or watt-hours per completed request. Direct measurement uses a GPU's cumulative energy counter or a certified external power analyzer, as specified in the MLPerf Power and SPEC PTDaemon methodologies, rather than a single instantaneous power reading (mlcommons.org).
- **What AI inference power consumption benchmarks exist?** MLPerf Power provides system-level, wall-outlet power benchmarking under a SPEC-certified methodology, while Hugging Face's AI Energy Score benchmarks individual models exclusively on H100 GPUs at a fixed batch size of one (mlcommons.org) (huggingface.github.io).
- **What is the energy cost of running AI models at scale?** The IEA estimates global data center electricity consumption was 415 TWh in 2024, projected to reach around 945 TWh by 2030, with AI-optimized servers growing electricity use at roughly 30% annually, well above the rate for conventional servers (iea.org) (iea.org) (iea.org).
- **How do researchers measure AI model energy efficiency?** Common approaches include NVIDIA's NVML power and energy counters, the open-source Zeus and CodeCarbon measurement libraries, and system-level power analyzers built to the SPEC PTDaemon standard used in MLPerf Power submissions (github.com) (codecarbon.io) (spec.org).
- **How much energy does a GPU consume per inference?** It depends on the GPU's rated TDP (300 to 400W for A100 and H100 PCIe, up to 700W for H100 and H200 SXM modules, up to 1,000W for B200) and how efficiently that power is utilized; per-token energy for a given model can vary by more than sixfold depending on batch size and context length alone (nvidia.com) (nvidia.com).
- **How much data center energy does a single AI query use, including cooling?** To estimate whole-facility energy, first measure or estimate the full IT-equipment energy attributable to the task and then apply a PUE measured for the relevant facility and period; label a GPU-only figure as GPU-only (U.S. Department of Energy).
- **What are the standard metrics for LLM inference energy?** The two most common metrics are joules (or watt-hours) per generated token and watt-hours per completed request, with a growing minority of studies also reporting a task-success-adjusted figure that excludes failed or retried requests from the energy accounting, though no single metric has yet been adopted as a mandatory industry or regulatory standard as of September 2026.
