

Sustainable and Synergistic Energy Management for Large Model Computing

Jingbo Hao ^{1,*} and Yang Tao ²

¹ Hunan Chaoneng Robot, Changsha 410003, China

² School of Digital Intelligence, Nanchang Vocational University, Nanchang 330599, China

* Correspondence: jbhao@126.com

How To Cite: Hao, J.; Tao, Y. Sustainable and Synergistic Energy Management for Large Model Computing. *AI Engineering* **2026**, *2*(2), 16. <https://doi.org/10.53941/aieng.2026.100016>

Received: 18 June 2026

Revised: 18 September 2026

Accepted: 20 September 2026

Published: 10 October 2026

Abstract: The rapid growth of foundation models, most prominently large language models (LLMs), is turning the electricity consumption of computing infrastructure into a first-order constraint on the artificial intelligence (AI) industry. This review synthesizes recent literature on three levers for reconciling large-model computing with power-system limits: green energy supply, green computing, and computing-energy synergy (CES), i.e., the coordinated operation of computing infrastructure and power systems. Beyond earlier descriptive accounts, the review makes four contributions: (1) comparative assessments that condense the energy-supply options for AI data centers and the principal energy-efficiency techniques, including quantization, pruning, knowledge distillation, batching, dynamic voltage and frequency scaling (DVFS), and scheduling, into single-view tables with their performance, latency, cost, energy, and carbon trade-offs; (2) an explicit introduction of standard sustainability metrics (PUE, CUE, WUE, energy per token, and carbon emissions per workload) as the quantitative backbone for comparing the surveyed approaches; and (3) a substantially deepened treatment of CES, comparing centralized, distributed, and hybrid algorithmic paradigms, formulating a representative joint scheduling problem, and analyzing feasibility under engineering constraints such as renewable-supply uncertainty, storage limits, grid congestion, workload deadlines, latency requirements, and electricity and carbon price volatility; and (4) a conceptual framework, that links supply-side options, the sustainability-metrics backbone, demand-side workload flexibility, and power-system engineering constraints into a single coordination view. Three research gaps are identified: integrating 24/7 carbon-free electricity procurement with flexible workload scheduling; co-scheduling delay-tolerant training with latency-sensitive inference; and embedding power-system constraints and market signals into CES algorithms. The evidence reviewed indicates that green energy lowers the operational emissions of large-model computing but does not by itself deliver continuous carbon-free operation; green computing techniques can substantially reduce energy per unit of computation at accuracy or latency costs; and CES is a promising coordination paradigm whose large-scale feasibility still needs to be demonstrated under realistic engineering constraints.

Keywords: large model; green energy; green computing; computing-energy synergy; data center; sustainability metrics



1. Introduction

As the frontier of artificial intelligence (AI), the development of foundation models, represented by large language models (LLMs), inevitably brings about significant challenges with respect to computing power and energy consumption [1]. Researchers need to continuously optimize the design of hardware systems to improve energy efficiency and resource utilization, and at the same time promote the application of green technology in algorithm optimization, model optimization, and lifecycle management to minimize energy consumption. For enterprises and research institutions, this is both a way to reduce costs and a path toward sustainable development.

The urgency of this agenda is grounded in well-documented evidence. The Intergovernmental Panel on Climate Change has concluded that human activities have unequivocally caused global warming and that deep, rapid greenhouse-gas emission reductions are required in the coming decades [2]. The globally averaged atmospheric CO₂ concentration reached 423.9 ppm in 2024, the highest level in at least 800,000 years [3]. Under the Paris Agreement, signatories committed to holding global warming well below 2 degrees C and pursuing efforts toward 1.5 degrees C, with carbon-neutrality targets concentrated in the 2050–2070 timeframe. At COP28, more than 130 countries endorsed the pledge to triple global installed renewable capacity to roughly 11,000 GW by 2030 [4]. The European Union is implementing its “Fit for 55” package; China pursues its dual-carbon goals (carbon peaking before 2030 and carbon neutrality before 2060) together with the “Eastern Data, Western Computing” national project; and the United States supports clean technologies through the Inflation Reduction Act. The principal challenges include the intermittency of wind and solar power, the cost of long-duration storage and green hydrogen, and grid-connection backlogs.

At the same time, large models are bringing significant changes to the energy sector through massive data analysis and intelligent optimization [5]. They can predict electricity demand, optimize grid dispatch, and improve the utilization of renewable energy; assist carbon capture and oil and gas exploration in the traditional energy industry; reduce energy waste through smart building and industrial energy-saving systems; and accelerate research and development on novel batteries and nuclear fusion. Large models are thus becoming key tools for the low-carbonization and intelligentization of the energy sector itself.

From a global perspective, the energy consumption of AI computing keeps rising with the advancement of large-model technology. The International Energy Agency (IEA) estimates that data centers consumed about 415 TWh of electricity in 2024, roughly 1.5% of global electricity demand, and projects this to more than double, to about 945 TWh, by 2030, with the United States, China and Europe accounting for the largest shares of the increase [6]. In Ireland, data centers already account for more than one fifth of the country’s metered electricity consumption [6]. Although AI technology is advancing toward lower cost and higher speed, the result of its global deployment is a disproportionate strain on energy and environmental resources [7].

To address the energy dilemma of AI computing, efforts must be made across multiple aspects, from computing-efficiency gains in new algorithms to the widespread adoption of clean energy. The coordinated development of computing power and electricity is a crucial pathway for achieving energy transition and improving societal energy efficiency. Green computing power has emerged as a defining trend: it is a holistic concept covering the entire lifecycle of computing resources from production and supply to service delivery, with a focus on low-carbon sustainability. Computing power and electric power, as core productive forces of the digital economy and the energy economy, have seen their coordination become part of national strategy.

This review goes beyond topic-level description in three respects. First, it consolidates scattered qualitative discussion into comparative tables: the energy-supply options for AI data centers (Table 2), the green-computing techniques and their trade-offs (Table 3), the standard sustainability metrics (Table 4), the workload classes and their flexibility characteristics (Table 5), the CES algorithmic paradigms (Table 6), and the CES engineering constraints and corresponding design responses (Table 7). Second, it introduces standard sustainability metrics (PUE, CUE, WUE, energy per token and carbon per workload) as a quantitative backbone for the whole survey (Section 3.3). Third, it deepens the treatment of CES, the most distinctive part of this review, by explicitly separating delay-tolerant training from latency-sensitive inference workloads (Section 4.2) and analyzing how engineering constraints affect CES feasibility (Section 4.4). The remainder of this paper is organized as follows. Section 2 reviews green energy with an emphasis on the options relevant to large-model computing. Section 3 reviews green computing. Section 4 explores computing-energy synergy. Section 5 discusses large-model-specific evidence and development trends, and Section 6 concludes.

Positioning and research questions. Existing surveys address parts of this landscape separately: green deep learning and green AI [8–10], data-center practices and green-aware management [11,12], and carbon-aware computing [13]. To the best of our knowledge, none connects the supply-side options, the standardized sustainability metrics, the demand-side workload flexibility of large-model computing, and the engineering

constraints of power systems in a single synthesis. This survey is positioned to fill that gap: its analytical frame, summarized in Figure 1, treats computing demand as a set of workload classes with distinct flexibility, benchmarks supply options and efficiency techniques on a common metrics backbone, and evaluates their coordination (CES) under explicit engineering constraints. Three research questions guide the review. RQ1: Which energy-supply options and procurement models can realistically deliver low-carbon, and ideally 24/7 carbon-free, operation for large-model data centers (addressed in Sections 2, 3.1 and 5.2)? RQ2: Through which efficiency levers and scheduling mechanisms can computing demand be made flexible without violating performance requirements (addressed in Sections 3.2, 4.2 and 5.1)? RQ3: Under which engineering constraints can computing-energy synergy be implemented at scale, and which research gaps remain (addressed in Sections 4.3–4.5 and 5.5)?

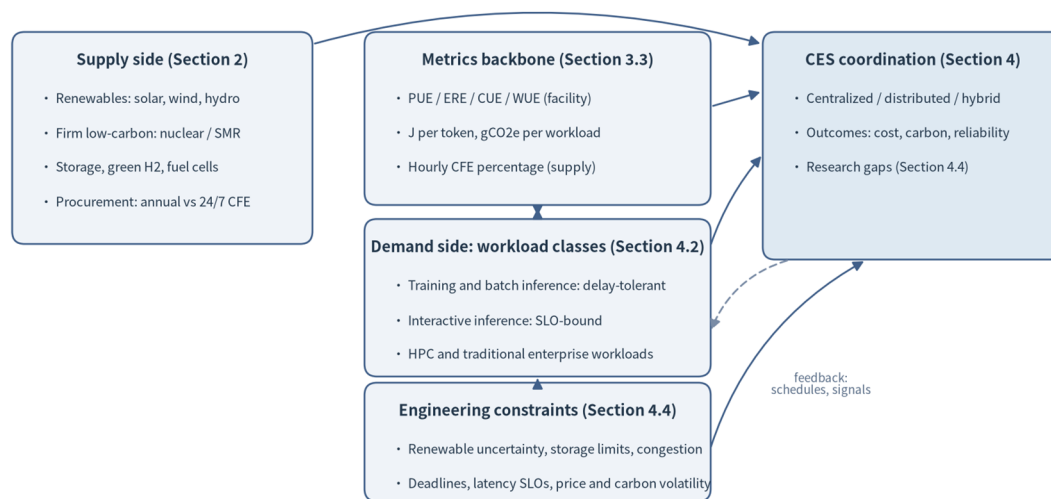


Figure 1. Conceptual framework of this survey: supply-side options (Section 2) and workload classes (Section 4.2) are benchmarked on a common metrics backbone (Section 3.3) and coordinated through CES (Section 4) under engineering constraints (Section 4.4); dashed arrows denote feedback of schedules and procurement signals.

Review Methodology

This is a narrative review rather than a systematic review. We searched Web of Science, Scopus, IEEE Xplore, the ACM Digital Library, and arXiv for English-language literature, and CNKI for Chinese-language policy and industry literature, covering documents published between January 2019 and September 2025; a small number of earlier seminal works were retained where foundational. Search terms combined model-side keywords (“large language model”, “foundation model”, “deep learning”, “green computing”) with energy-side keywords (“energy consumption”, “carbon footprint”, “renewable energy”, “data center”, “power system”, “demand response”). Two inclusion criteria were applied: (i) peer-reviewed journal or conference publications, or authoritative institutional and governmental reports (e.g., IEA, WMO, IPCC, China Academy of Information and Communications Technology); and (ii) clear relevance to the energy or environmental aspects of computing. Preprints were retained only when widely cited or authored by established groups, given the fast-moving nature of the field. Records were screened by title and abstract and, where relevant, full text; backward snowballing from the reference lists of key surveys was used to complete coverage. Because the review is narrative in nature, no PRISMA-style screening statistics are reported; this limitation is acknowledged in Sections 5.5 and 6. Table 1 summarizes the methodology.

Table 1. Summary of the literature search methodology.

Item	Detail
Databases	Web of Science, Scopus, IEEE Xplore, ACM Digital Library, arXiv (English); CNKI (Chinese policy and industry literature)
Publication period	January 2019 to September 2025; selected earlier seminal works retained
Search terms	Model-side: large language model, foundation model, deep learning, green computing. Energy-side: energy consumption, carbon footprint, renewable energy, data center, power system, demand response
Inclusion criteria	(i) Peer-reviewed journal/conference papers or authoritative institutional and governmental reports; (ii) clear relevance to energy or environmental aspects of computing
Exclusion criteria	Studies without verifiable quantitative or methodological grounding; duplicate coverage of the same technique without added evidence
Supplementary strategy	Backward snowballing from reference lists of key surveys and reports

2. Green Energy

The development of green energy provides sustainable power support for the training and deployment of large models and can substantially reduce their operational carbon footprint. With the rapid adoption of wind and solar power, global computing hubs are increasingly sited in regions abundant in green energy, from hydro-rich Guizhou and wind-rich Gansu and Inner Mongolia under China's "Eastern Data, Western Computing" project [14] to the Nordic countries, because the availability of low-carbon power has become a primary siting criterion for AI data centers. Since general expositions of renewable technologies are available elsewhere [15], this section keeps the technology overview brief and focuses on the aspects most relevant to powering large-model computing; Section 2.5 compares the options in a single view.

2.1. Renewable Energy and Storage

Solar photovoltaics generate electricity with near-zero operational emissions and are available almost everywhere, at steadily declining cost. For data centers, solar power is a good daytime complement in sunny regions, but its diurnal and weather-driven variability means that it cannot by itself provide the round-the-clock, uninterruptible supply that large-model training and inference require; pairing with storage or grid backup is essential. Lifecycle emissions from module manufacturing and end-of-life panel waste, although small relative to fossil generation, are non-zero and are discussed in Section 2.4.

Wind power offers abundant resources, wide distribution, low operational emissions, and competitive cost. Its output is intermittent and synoptically variable, so wind works best for data centers when combined with solar generation, storage, and firm capacity. For computing hubs in wind-rich regions, direct corporate power purchase agreements with wind farms have become a standard procurement instrument [16]. Grid integration requires careful coordination with existing power systems, particularly in regions with weak transmission.

Hydropower is dispatchable, large-scale low-carbon generation and is therefore an attractive firm anchor for large computing clusters, as in Guizhou and the Nordic countries. Its limitations are ecological and climatic: large dams disrupt river ecosystems and biodiversity, reservoirs flood land and can cause substantial evaporative water losses, and output depends on seasonal and hydrological conditions, with drought years reducing firm capacity [17].

For energy storage, pumped hydro storage and lithium-ion battery storage are complementary: pumped hydro suits large-scale, long-duration storage, while lithium-ion batteries suit medium- to short-term buffering [15]. For data centers, on-site storage bridges renewable-generation gaps and provides uninterruptible-power functions, but multi-day coverage of gigawatt-scale AI campuses remains beyond current storage economics.

Renewable energy and the smart grid are in a close, dynamic relationship [18]. By monitoring and forecasting renewable generation in real time, adjusting generation and load distribution, integrating distributed generation, and dispatching storage, smart grids alleviate the integration challenges of variable renewables. This monitoring and forecasting capability is also the technical foundation of carbon-aware workload scheduling discussed in Section 3.1.

For large computing facilities, however, the binding constraints are increasingly transmission and interconnection rather than generation itself. Renewable-rich regions are often remote from the metropolitan areas where data centers concentrate for latency, staffing and fibre reasons, so large new loads face interconnection queues that can take years, and grid operators must reinforce corridors that are already congested; the IEA projects grid-connection bottlenecks to be a principal constraint on data-center growth in several markets [6]. Curtailment of solar and wind in generation-rich, transmission-poor regions, and the long lead times of high-voltage direct-current corridors of the kind underlying China's West-to-East transmission, mean that the geographic matching of computing demand to renewable supply, the premise of China's "Eastern Data, Western Computing" siting strategy [14] and of Nordic and US hub siting, is as important as the temporal matching discussed later. Voltage-stability and frequency-support requirements of gigawatt-scale campuses further couple data-center operation with power-system operation, which is the starting point of CES in Section 4.

2.2. Nuclear Energy and Small Modular Reactors

Nuclear energy provides substantial, weather-independent power with near-zero operational greenhouse-gas emissions, which makes it a candidate firm baseload for large data-center campuses whose load is flat and continuous. Its drawbacks are equally well known: lifecycle emissions from mining, enrichment and construction are non-zero; radioactive waste requires long-term secure containment; accidents carry severe consequences; and facilities involve high capital costs, long construction times and high technical barriers. A sustainable deployment path must therefore weigh firm low-carbon output against safety, waste and cost.

Small modular reactors (SMRs), with capacities up to about 300 MW and factory fabrication, are a flexible low-carbon option better suited to integration with smaller grids and localized demands; more than 80 commercial designs are under development worldwide, with some units already operating in Russia and China [19]. Industrial interest in nuclear-powered computing is visible: several hyperscale operators have recently announced agreements to procure power from restarted or advanced reactors for their data centers. Even so, SMR deployment timelines largely extend beyond the 2030 horizon in which most AI-driven load growth is projected.

2.3. Frontier Technologies: Green Hydrogen and Fuel Cells

Green hydrogen, produced by water electrolysis powered by renewables, and fuel-cell technology, which converts hydrogen into electricity efficiently, are relevant to large-model computing mainly as long-duration storage and clean backup power, niches where batteries are uneconomic. Both remain expensive, and the supporting infrastructure for production, storage and distribution is immature; major economies have incorporated hydrogen into national strategies and continue to drive breakthroughs in electrolyzers, hydrogen-storage materials and fuel cells. In the near term, green hydrogen is best viewed as a complementary option for AI data centers rather than a primary supply route.

2.4. Environmental Impacts

Environmental statements about electricity generation must distinguish operational impacts from lifecycle impacts. Operationally, wind, solar, and hydro emit essentially no CO₂ or air pollutants during generation; lifecycle analyses, however, attribute non-zero emissions to equipment manufacturing, transport, construction, and decommissioning [17]. Three impact categories deserve particular emphasis for data-center siting. First, land use: wind farms and utility-scale photovoltaic plants require significant land or water surfaces, which can reduce available land resources and affect local ecosystems. Second, water use: although wind and photovoltaics consume little water in operation, hydropower reservoirs can cause substantial evaporative losses, and data-center cooling is itself a significant water consumer; data centers in the United States were estimated to directly and indirectly consume water on the order of hundreds of billions of litres per year [20,21]. Third, end-of-life waste: decommissioned photovoltaic panels, turbine blades and batteries require recycling or disposal at scale to prevent pollution. A fourth category, upstream material extraction, is specific to lifecycle accounting: lifecycle assessment (LCA) methodology attributes impacts to all stages from raw-material extraction through manufacturing, operation and end-of-life [17], and for renewable technologies these stages include the mining of critical materials, notably rare-earth elements for permanent-magnet wind turbines and lithium and cobalt for batteries, whose extraction carries localized environmental and social impacts that operational accounting does not capture [17]. Claims of “zero-emission” or “zero-water” generation therefore hold only under operational accounting and for specific technologies; lifecycle accounting provides a more accurate basis for comparing supply options, and the same distinction between operational and lifecycle accounting applies to the computing side (Section 3.3).

2.5. Comparative Overview

Table 2 condenses the discussion above into a one-page comparison of the energy alternatives for AI data centers across availability and dispatchability, carbon intensity, intermittency, cost trends, land and water footprint, and suitability for large-model workloads.

Table 2. Comparative overview of energy options for AI data centers.

Option	Dispatchability	Operational Carbon	Intermittency	Cost Trend	Land/Water Footprint	Suitability for Large-Model DCs
Solar PV	None (daytime only)	Near zero; lifecycle emissions from manufacturing	High (diurnal, weather)	Steeply declining	High land use; low operational water	Good daytime complement in sunny regions; needs storage or grid backup for 24/7 operation
Wind	None (variable)	Near zero	High (synoptic variability)	Declining; offshore costlier	Moderate land (co-use possible); low water	Strong complement to solar; fits wind-rich hub siting; needs firm backup
Hydropower	High (with reservoir)	Near zero; some reservoir methane	Seasonal; drought risk	Site-dependent; low operating cost	Reservoir land flooding; evaporative water loss	Attractive firm low-carbon anchor for large clusters; constrained by ecology

Table 2. *Cont.*

Option	Dispatchability	Operational Carbon	Intermittency	Cost Trend	Land/Water Footprint	Suitability for Large-Model DCs
Nuclear/SMR	Very high (firm baseload)	Near zero; lifecycle emissions from fuel cycle	None (except refuelling outages)	High capital; SMR costs uncertain	Small land; cooling water needed	Strong fit for flat DC load; deployment timelines largely post-2030
Green H ₂ /fuel cells	High (stored fuel)	Near zero if H ₂ is renewable	None	High today, declining	Moderate (electrolysis water use)	Niche: long-duration storage and clean backup; not a near-term primary supply
Grid + annual renewable matching	High (grid)	Depends on grid mix	None for the DC	Lowest cost path	Delegated to generators	Cheapest headline-green option, but weak real-world emission impact [22]
Grid + 24/7 carbon-free matching	High (grid)	Lowest effective	None for the DC	Premium (can exceed \$20/MWh in some regions)	Delegated to generators	Strongest system-level emission impact; drives storage and firm clean-energy investment [22]

3. Green Computing

Green computing aims to employ energy-efficient hardware, optimized algorithms, and clean energy to reduce the energy consumption and carbon emissions of computing systems [23]. As carbon-neutral technologies mature, green computing may emerge as one of the key differentiators in large-model competition. This section reviews green data centers (Section 3.1), green algorithms and systems together with their trade-offs (Section 3.2), the standard sustainability metrics needed to compare them (Section 3.3), and quantum computing as a longer-horizon option (Section 3.4).

3.1. Green Data Centers

Traditional data centers rely heavily on fossil-based electricity and contribute substantially to greenhouse-gas emissions. Green data centers adopt energy-efficient hardware, renewable-energy integration, advanced cooling, and resource-utilization optimization to reduce energy consumption and carbon footprint, while lowering operating costs and complying with environmental regulations [24]. Sustainable hardware is a key lever: vendors are introducing more powerful tensor and graphics processing units for large-scale training and inference, low-power variants and domain-specific accelerators for particular workloads [25]. Beyond chips, advanced cooling, in particular direct liquid cooling and immersion cooling, reduces the overhead between IT load and facility load, which is captured by the PUE metric introduced in Section 3.3 [12].

Renewable-energy procurement has become standard practice for large operators. Google, Apple and Microsoft have pursued high renewable coverage through integrated wind-solar-storage portfolios and power purchase agreements [16]; the renewable-energy mix differs markedly across operators and years because it follows grid conditions at each site. Google reports a global average hourly carbon-free-energy match of 66% across its data centers in 2024, up from 64% in 2023, with nine of its twenty grid regions at 80% or higher [26]; Microsoft reported 34 GW of contracted renewable energy and 29.8 MtCO₂e of contracted carbon removal by 2024 [27]. As shown in Figure 2, Microsoft reports substantial progress since 2020 on its environmental-sustainability commitments across carbon, energy, water, waste and land. Such figures also illustrate the accounting caveat discussed in Section 5.2: contracted volumes and hourly-matched percentages measure different things, and only the latter approximates the temporal reality of supply.

A critical distinction, often blurred in corporate reporting, is that between annual renewable-energy matching and continuous 24/7 carbon-free electricity (CFE). Annual matching purchases enough clean energy over a year (through certificates or power purchase agreements) to equal annual consumption, but the data center still draws fossil-based power at night, in windless periods, and in low-output seasons. Twenty-four-seven CFE instead matches consumption with carbon-free supply in every hour at every grid where the operator runs. The two are not equivalent in impact: system-level modelling of United States procurement strategies shows that annual volumetric matching yields little additional long-run emission reduction, because the wind and solar it contracts would in many cases have been built anyway, whereas hourly 24/7 CFE matching is the strategy most consistently associated with lower system-wide emissions, driving investment into storage and firm clean-energy technologies at a cost premium that can exceed \$20/MWh in some regions [22]. For AI data centers, this distinction matters because their loads are large, flat, and growing, precisely the profile that stresses the hourly matching of variable renewables.

On the demand side, Google's carbon-intelligent computing system (CICS) proactively minimizes carbon footprint and infrastructure cost by time-shifting flexible workloads [13]. At its core, analytical processes collect

next-day carbon-intensity forecasts, train daily prediction models, and use risk-aware optimization to generate next-day carbon-aware virtual capacity curves, which impose hourly limits on the resources available to flexible workloads while preserving total daily capacity. Operationally, this delays flexible jobs out of carbon-intensive hours with a high probability of completion within the day. CICS demonstrates that even without new generation, demand-side flexibility can be converted into carbon savings, and it prefigures the tighter coupling of computing and power systems discussed in Section 4.

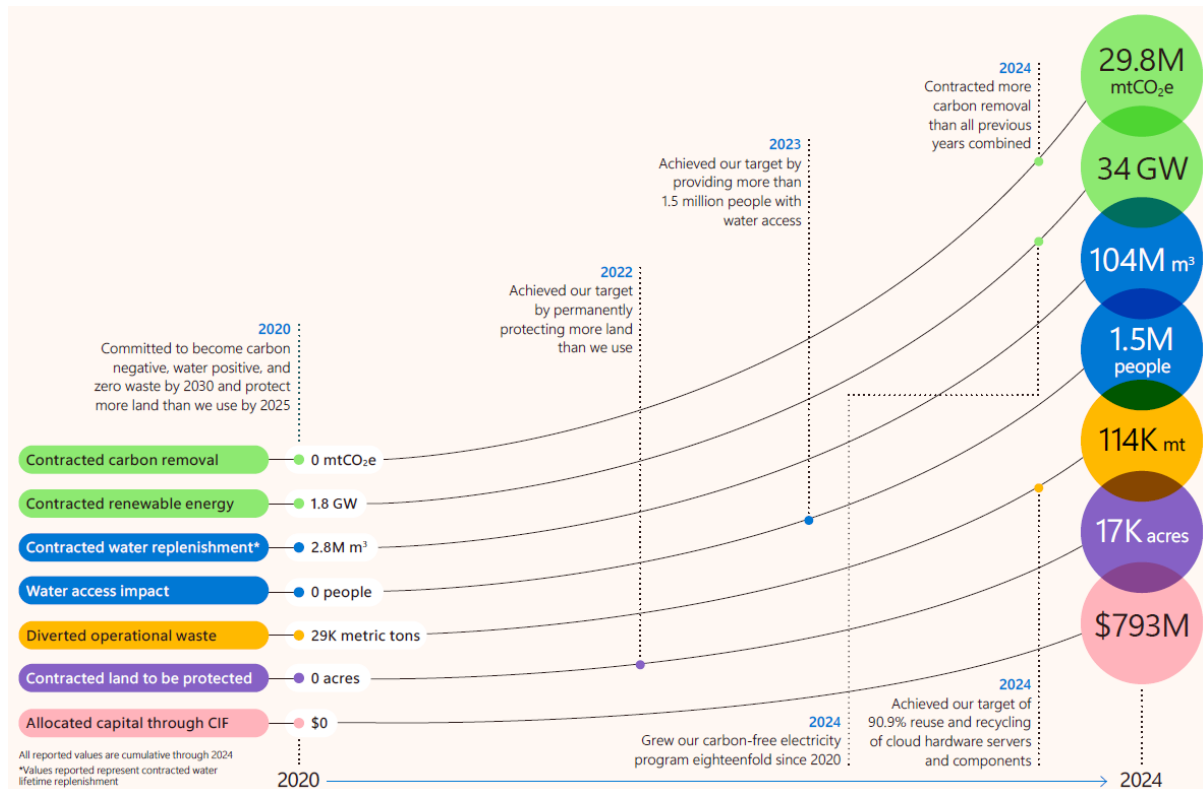


Figure 2. Microsoft's reported environmental-sustainability commitments and cumulative contracted volumes, 2020–2024: contracted carbon removal (29.8 MtCO₂e), contracted renewable energy (34 GW), contracted water replenishment (104 million m³), people reached with water access (1.5 million), diverted operational waste (114 thousand tonnes), land protection (17 thousand acres), and Climate Innovation Fund allocations (\$793 million) [27].

Complementary data-center-level techniques include digital twins, which create virtual models of facilities for real-time monitoring, predictive maintenance, and AI-driven optimization of energy use and emissions [11], and waste-heat recovery, in which district heating, on-site power production from low-temperature heat, and absorption cooling convert the large heat output of data centers into useful energy, subject to geographic and infrastructure constraints [12].

3.2. Green Algorithms and Systems

On the algorithm side, green deep learning covers model compression, knowledge distillation, compact networks, energy-efficient training, and efficient data utilization [8], and energy-efficient model design builds lightweight architectures with lower parameter counts and computational complexity [9]. To measure and manage carbon footprints, tools such as CarbonTracker, CodeCarbon, Green Algorithms, and PowerTop estimate the energy consumption and emissions of AI workloads [10]. On the system side, the design of LLM inference services balances performance, energy, and cost: batching combines concurrent requests for joint processing, parallelism tuning adapts GPU counts to workloads, frequency scaling lowers GPU power, and scheduling and software optimization raise utilization and throughput [28].

Table 3 summarizes the principal techniques in a structured comparison, spanning model-level (quantization, pruning, knowledge distillation, compact and mixture-of-experts architectures) and system-level (batching, DVFS, carbon- and price-aware scheduling) methods. The central message is that every technique exchanges one scarce resource for another: quantization and distillation exchange accuracy for energy and memory; batching exchanges per-request latency for throughput and energy efficiency; DVFS exchanges speed for power; and workload

scheduling exchanges temporal immediacy for cheaper or cleaner energy. Quantifying and managing these trade-offs, rather than assuming energy savings are free, is a prerequisite for deploying them in production LLM services [28]. Token-efficiency techniques are becoming more valuable as reasoning-enhanced models multiply output token counts, and thus inference energy, per query [29,30].

Table 3. Structured comparison of green-computing techniques for large-model workloads.

Technique	Mechanism	Typical Energy Effect	Main Trade-Off	Best Suited for
Quantization	Lower-precision weights and activations (e.g., INT8/FP8/INT4)	Lower memory traffic and compute per token; higher throughput per accelerator	Accuracy degradation at aggressive bit-widths; kernel support varies	Inference, especially memory-bound models
Pruning	Remove redundant weights, channels or experts	Smaller compute and memory footprint	Requires re-tuning; unstructured sparsity hard to realize on GPUs	Deployment under tight resource budgets
Knowledge distillation	Train a smaller student model from a larger teacher	Order-of-magnitude fewer parameters at deployment	Capability gaps versus the teacher; upfront training cost	High-volume inference services
Compact/MoE architectures	Efficient architectures and conditional computation	Better quality per FLOP for training and inference	Training complexity; routing and memory overhead	Both training and inference at scale
Batching and request scheduling	Group concurrent requests; schedule to utilization valleys	Higher GPU utilization; lower energy per token	Higher and more variable per-request latency	Latency-tolerant or heavily loaded services
DVFS/frequency scaling	Lower processor frequency and voltage	Near-linear power savings per frequency reduction	Slower execution; possible SLO violations	Power-capped or non-urgent jobs
Carbon-/price-aware scheduling	Shift flexible jobs to cleaner or cheaper hours and regions	Little direct energy saving; reduces emissions and energy cost	Requires delay tolerance and reliable forecasts	Training and batch inference with deadlines

Note that Table 3 deliberately reports energy effects qualitatively: reported savings vary widely with model, hardware and serving stack, and cross-paper quantitative comparisons are often not reproducible, which is itself a motivation for the standardized metrics of Section 3.3.

For LLMs specifically, the dominant decoder-only transformer architecture has a characteristic energy profile that shapes which techniques pay off. First, self-attention cost grows with context length, and autoregressive decoding is memory-bandwidth-bound: each generated token requires reading the key-value (KV) cache of all previous tokens, so energy per token is dominated by memory traffic rather than arithmetic, which is why quantization of weights and KV caches yields savings that FLOP counts alone do not predict [28,30]. Second, architecture families differ systematically: dense models spend all parameters on every token, whereas mixture-of-experts (MoE) models activate only a subset per token, exchanging memory capacity for per-token compute efficiency, and speculative decoding uses a small draft model to let the large model verify multiple tokens per forward pass, trading a cheap draft for fewer expensive passes. Third, model scale translates into energy: across the open-weight models measured in Section 5.1, the 7-8B class emits roughly an order of magnitude less CO₂ per query than the 70B class, so choosing the smallest capability-sufficient model is the single highest-leverage efficiency decision. This is also where LLM-specific distillation and pruning matter: distilling a 70B-class teacher into a 7-8B-class student preserves much of the instruction-following and reasoning capability at an order-of-magnitude lower per-token energy, and INT8/INT4 quantization of open-weight LLMs has become standard deployment practice [8,9]. The trade-off is always capability versus energy: reasoning-enhanced variants multiply output tokens, and thus energy, per query [29], so token-efficiency and output-length control belong in the same toolbox as architecture selection.

3.3. Sustainability Metrics

Comparing the approaches above requires standard metrics. At the facility level, power usage effectiveness (PUE), the ratio of total facility energy to IT equipment energy, is the dominant efficiency metric [31]; modern hyperscale facilities operate near 1.1–1.2, while older facilities commonly exceed 1.5. The energy reuse effectiveness (ERE) extends PUE when waste heat is recovered (Section 3.1). Carbon usage effectiveness (CUE) multiplies facility energy by the carbon intensity of the supplying grid and normalizes the result by IT equipment energy, and water usage effectiveness (WUE) measures litres of water consumed per kilowatt-hour of IT energy; both link facility efficiency to the environmental characteristics of the site [11,31]. At workload level, energy per token or per inference (joules per generated token or per completed query) and carbon per workload (grams of CO₂-equivalent per training run or per query) combine hardware power draw, PUE, and grid carbon intensity, and are the units in which LLM energy studies now report [7,32–34]. Finally, hourly carbon-free-energy percentage measures the temporal matching quality of supply under 24/7 CFE procurement [22]. Table 4 summarizes these metrics.

Table 4. Standard sustainability metrics for data centers and large-model workloads.

Metric	Definition	Unit	Scope and Notes
PUE	Total facility energy divided by IT equipment energy	dimensionless	Facility level; lower is better; blind to energy source and water
ERE	Extension of PUE crediting recovered waste heat	dimensionless	Facility level; relevant where heat reuse exists
CUE	CO ₂ -equivalent emissions per unit of IT energy	kgCO ₂ e/kWh	Facility level; couples PUE with grid carbon intensity
WUE	Water consumed per unit of IT energy	L/kWh	Facility level; site-dependent; includes cooling water
Energy per token/inference	Energy per generated token or per completed inference	J/token, J/query	Workload level; hardware-, stack- and model-dependent
Carbon per workload	CO ₂ e per training run or per query	gCO ₂ e/run, gCO ₂ e/query	Workload level; combines energy per token with PUE and grid carbon intensity [33,34]
Hourly CFE %	Share of consumption matched by carbon-free supply in each hour	%	Supply level; measures quality of 24/7 matching [22]

Two caveats apply. First, metric values are highly site- and grid-dependent, so comparisons across studies require care. Second, metrics can conflict: a low-PUE data center on a carbon-intensive grid may emit more than a high-PUE facility on a clean grid, which is precisely why facility efficiency and energy supply should be optimized jointly, the subject of Section 4. It should also be noted that operational metrics omit embodied carbon in hardware manufacturing; lifecycle assessments of AI systems attempt to close this gap [35].

3.4. Quantum Computing: Prospects and Caveats

Quantum computing is often proposed as a long-term solution to the energy cost of AI computing: qubit superposition and entanglement, together with quantum algorithms, promise speedups in parallel processing, search and optimization, and quantum simulation could accelerate materials research relevant to batteries and energy technologies [36]. These prospects should be weighed against three caveats. First, current quantum hardware remains limited in qubit count, fidelity, and error correction, and practical quantum advantage for workloads relevant to large-model computing is still an open question rather than an established fact [37]. Second, energy accounting must include overheads: many leading platforms require cryogenic cooling, in the millikelvin range for superconducting qubits, whose energy demand, together with control electronics and classical co-processing, can offset claimed savings; whether a net energy-efficiency quantum advantage materializes is itself an active research question [37]. Third, projections of gigaton-scale CO₂ abatement by 2035, illustrated in Figure 3, rest on optimistic assumptions about hardware scaling and technology diffusion and should be read as upper-bound estimates [38]. Quantum computing is therefore best treated as a longer-horizon option for large-model energy management rather than a near-term solution.

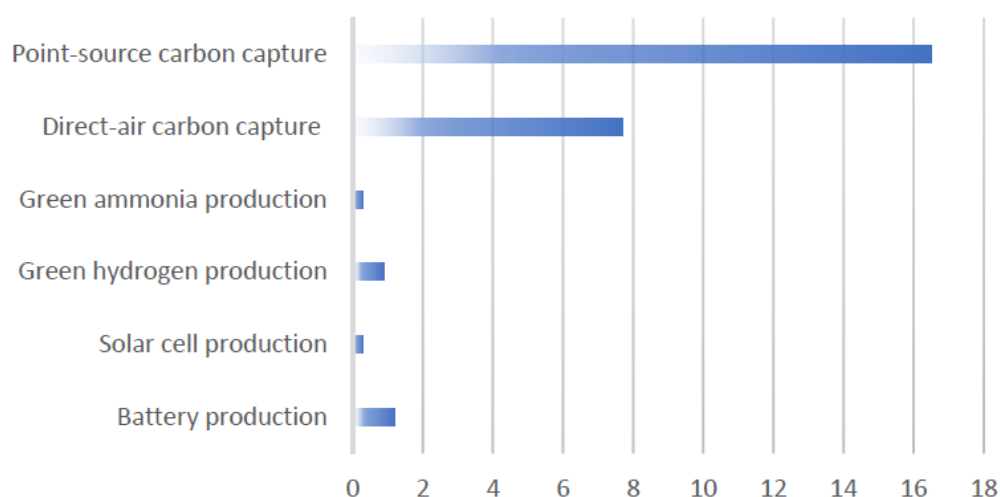


Figure 3. Projected application-level CO₂ emission-reduction potential by 2035 (gigatonnes) under optimistic technology-adoption assumptions [38].

4. Computing-Energy Synergy

Computing-energy synergy (CES) leverages new-type power systems as its foundation and is guided by the high-quality development of computing infrastructure. With advances in intelligent scheduling, source-grid-load-storage coordination, power supply and backup systems, and green-power aggregation, CES promotes innovation

in both technologies and mechanisms. Its goal is holistic optimization of computing and power systems across industrial planning, operational management, resource allocation, and market design. This section describes the CES framework (Section 4.1), analyses the workload flexibility that makes coordination possible (Section 4.2), compares CES algorithms (Section 4.3), and examines the engineering constraints that determine feasibility (Section 4.4).

4.1. CES Framework

As shown in Figure 4, CES involves four key components: power source, power grid, energy storage, and computing load [39]. Innovative paradigms for integrated source-grid-load-storage systems are built through bilateral coordination and interaction among the components: flexible adjustment of computing workload helps optimize the allocation of power resources, and green-electricity and green-certificate trading mechanisms improve the utilization of renewable energy. CES is a progressive development process, from exploration to maturity and from coordination to full integration, commonly divided into four stages [40]: (1) initial exploration, in which the two systems begin to coordinate; (2) early development, in which policy support shapes a preliminary coordination mechanism; (3) deep coordination, in which the industry chain is largely established; and (4) full integration, in which clean energy is deeply integrated, and computing and energy systems merge into an efficient, green, and intelligent paradigm.

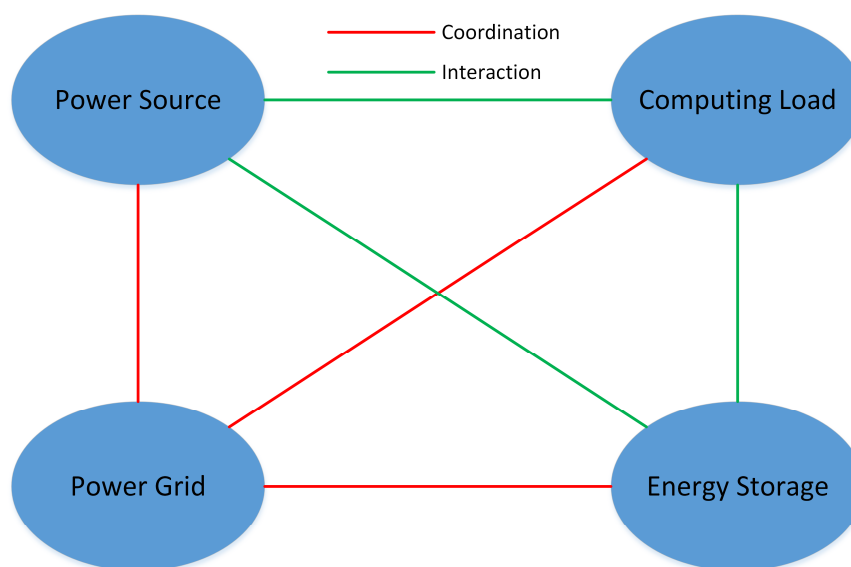


Figure 4. Basic framework of computing-energy synergy.

In China, the computing-power industry is large and fast-growing but energy-intensive, regionally concentrated, and thus a substantial burden on local grids; the computing load shows stable, seasonally cyclical patterns whose flexible-adjustment potential remains underutilized, and green-electricity integration is urgently needed [14]. The existing “West-to-East Power Transmission” and “Eastern Data, Western Computing” projects operate largely independently, with separate optimization for the power system and the computing network. Yang et al. [41] therefore proposed a CES framework as an enhancement and supplement to the two projects, aiming at more efficient, low-carbon energy utilization through coordinated scheduling, multi-spatiotemporal optimization, virtualization technologies, data-flow collaboration, and improved market mechanisms.

4.2. Workload Classes and Their Flexibility

Flexible computing demand is the linchpin of CES, but flexibility differs radically across workload classes, and conflating them leads to either infeasible schedules or wasted flexibility. A useful separation is between delay-tolerant and latency-sensitive workloads, summarized in Table 5.

Foundation-model pre-training and large-scale fine-tuning run for weeks to months, consume GWh-scale energy per run, and are structurally delay-tolerant: periodic checkpointing allows interruption, migration and restart, and batch scheduling allows start-time deferral. Training is therefore the primary target of carbon- and price-aware temporal shifting, and its tolerance for interruption makes it an analog of deferrable industrial loads in demand-response programmes [13]. Interactive LLM inference (chatbots, coding assistants, agents) faces strict latency requirements, typically hundreds of milliseconds for first-token response, and 24/7 arrival patterns; its

temporal flexibility is small, and the levers that remain are request batching, model downsizing and distillation, and geographic load balancing across data centers [7,28]. Offline or batch inference (document processing, large-scale evaluation, media generation) carries deadlines of minutes to hours and is largely delay-tolerant, making it, together with training, the dispatchable portion of AI demand. HPC simulations have variable deadlines and scheduling flexibility, and traditional enterprise workloads are mostly batch processes that are easily shifted. Since inference dominates the lifecycle energy of deployed AI services [29,35], efficiency levers (Section 3.2) matter more for interactive inference, while scheduling levers matter more for training and offline inference; a CES framework should treat the former as efficiency-constrained firm demand and the latter as dispatchable demand-side resources.

Table 5. Workload classes for large-model computing and their flexibility characteristics.

Workload Class	Latency Requirement	Temporal Flexibility	Energy Profile	Grid-Interaction Mechanisms
Foundation-model pre-training/large fine-tuning	None beyond project deadline	High (checkpoint, migrate, defer start)	GWh-scale per run; sustained multi-week draw	Carbon-/price-aware shifting; interruption for curtailment; capacity bidding
Interactive LLM inference	Strict SLOs (hundreds of ms first token)	Low	Continuous 24/7; scales with traffic and output length	Batching; model downsizing; geographic load balancing; efficiency levers
Offline/batch inference	Minutes-to-hours deadlines	High	Large blocks of shiftable compute	Temporal shifting to clean/cheap hours; queueing
HPC simulation	Variable deadlines	Moderate	Sustained node-hours	Scheduling within deadlines; curtailment participation
Traditional enterprise/batch	Hours to days	High	Small blocks, easily deferred	Conventional demand response

4.3. CES Algorithms

CES algorithms couple power flow and data flow for the joint optimization of energy and computing resources [42–45]. By system architecture and control method, they fall into three families with distinct characteristics and application scopes; to date, the field has evolved from one-way scheduling to two-way interaction, and from centralized optimization to distributed coordination.

The centralized family aggregates information from all computing and power nodes to a central scheduler that performs global optimization. Its advantage is the theoretically global optimal solution; its challenges are high computational complexity, difficult privacy protection, and heavy communication overhead. It suits regional, small- to medium-scale systems. The distributed family adopts a multi-agent architecture in which each computing or power node is an independent decision maker and system-wide coordination emerges from information exchange and local optimization; it scales well and protects privacy better, but suffers slow convergence and high coordination complexity, fitting multi-cloud data centers and cross-regional computing hubs. The hybrid family combines the two, typically with hierarchical control: an upper layer for global coordination and macro-optimization, a lower layer for local optimization and autonomous decision-making. It balances global optimality and local autonomy at the price of high system complexity and interface-standardization challenges, fitting complex systems such as a nationwide integrated computing network. Table 6 compares the three paradigms.

Table 6. Comparison of CES algorithmic paradigms.

Paradigm	Scalability	Communication Requirement	Privacy	Computational Complexity	Typical Applications
Centralized	Low to moderate	High (all node information to a center)	Weak	High at the center; grows with system size	Regional, small- to medium-scale systems; single-operator campuses
Distributed	High	Moderate (peer exchange of local states)	Strong (local data stays local)	Moderate per node; slow convergence	Multi-cloud data centers; cross-regional computing hubs
Hybrid	High	Layered (aggregate up, dispatch down)	Moderate	Balanced across layers	Nationwide integrated computing networks; provincial-national coordination

The three families can be placed on a common mathematical footing with a representative joint scheduling problem. Let T denote the scheduling horizon and t an interval within it; let g_t denote the grid import, s_t and d_t the storage charging and discharging power, D_t the computing demand of the schedule, and r_t the flexible workload deferred beyond the horizon. The operator minimizes the sum of electricity cost, carbon cost, and deferral penalty:

$$\min \sum_{t \in T} [\pi_t g_t + \lambda^{CO_2} e_t g_t] + \mu \sum_{t \in T} r_t \quad (1)$$

where π_t is the electricity price in interval t , e_t the grid carbon intensity (tCO₂e per MWh), λ^{CO_2} the carbon price (currency per tCO₂e), and μ the deferral penalty per unit of deferred workload, subject to four constraint families: (i) power balance, $g_t + d_t - s_t = D_t \cdot \text{PUE}$, with the PUE factor covering cooling overhead (Section 3.3); (ii) storage dynamics, battery state evolving with charging and discharging efficiencies within capacity and degradation limits; (iii) computing feasibility, each training or batch job completing by its deadline and interactive inference served within its latency SLO (Section 4.2); and (iv) network feasibility, power flows on each corridor remaining within transmission limits (Section 2.1). The formulation makes explicit where the three algorithm families diverge: the centralized family solves the resulting optimization problem globally, typically in a model-predictive rolling-horizon fashion [42]; the distributed family decomposes it by Lagrangian duality, ADMM or auction mechanisms so that nodes exchange only prices or boundary variables [44,45]; and the hybrid family solves a coarse master problem centrally, and delegates constrained subproblems to local controllers [41]. Published coordinated-scheduling studies of data centers and power systems follow exactly this pattern, differing mainly in how realistically the constraint families are modelled [42,43].

Key design factors include the spatiotemporal flexibility of computing tasks (Section 4.2), the security constraints of the power system, the variability of renewable energy and the coordination of interests among multiple stakeholders [42,43]. With advances in AI and the Internet of Things, CES algorithms are evolving from static planning to dynamic perception, from single-objective to multi-objective optimization, and from rule-based to data-driven methods [44,45]; as computing demand grows explosively and power-system transformation deepens, algorithmic innovation and the energy transition reinforce each other.

4.4. Engineering Constraints and Feasibility

The CES literature often optimizes abstract cost functions under simplified assumptions; in practice, six engineering constraints shape what coordination can achieve, and each implies specific design responses (Table 7). (a) Renewable-supply uncertainty: wind and solar forecasts are imperfect at operational horizons, so carbon- or price-aware schedules must be robust or stochastic, with reserve margins and recourse actions rather than deterministic plans [13]. (b) Storage-capacity limits: on-site batteries cover minutes to a few hours, so storage buffers short mismatches but cannot bridge multi-day renewable droughts; computing load itself, being dispatchable, can act as a virtual storage that reduces the required physical capacity [42]. (c) Grid congestion: transmission limits between generation-rich and demand-rich regions, salient in the contrast between western generation and eastern demand in China, mean that locational signals, not just temporal ones, must enter the scheduling objective, and geographic load balancing must respect network capacity [41]. (d) Workload deadlines: training runs and offline-inference batches have completion deadlines that bound the depth of delay; schedules must preserve probabilistic completion guarantees, as CICS does through risk-aware capacity curves [13]. (e) Latency requirements: interactive inference SLOs constrain curtailment participation and geographic shifting, since network latency adds to service latency; only efficiency levers remain feasible during tight grid conditions [28]. (f) Electricity- and carbon-price volatility: price spikes and carbon-price swings expose data-center operators to cost risk, which long-term power purchase agreements hedge but which also weakens the price signal that CES scheduling exploits; market design must reconcile these two functions [45].

Table 7. Engineering constraints on CES and corresponding design responses.

Constraint	Impact on CES Operation	Design Responses
Renewable-supply uncertainty	Forecast errors make deterministic schedules infeasible or carbon-suboptimal	Stochastic/robust optimization; reserve margins; rolling-horizon re-optimization [13]
Storage-capacity limits	Batteries buffer only short mismatches; degradation limits cycling	Computing load as virtual storage; hybrid storage sizing; duration-aware dispatch [42]
Grid congestion	Temporal shifting alone cannot deliver power across congested corridors	Locational signals; geographic load balancing within network limits; siting coordination [41]
Workload deadlines	Bounded delay depth for training and batch inference	Risk-aware capacity curves; probabilistic completion guarantees [13]
Latency requirements	Interactive inference cannot be delayed or moved freely	Efficiency levers (batching, DVFS within SLO); admission control; edge offloading [28]
Electricity/carbon price volatility	Cost risk to operators; unstable incentives for flexibility	Long-term PPAs as hedge; separate hedging from dispatch signals; market-rule design [45]

Against this background, three research gaps emerge. Gap 1: coupling 24/7 CFE procurement with workload scheduling. Procurement decides the supply side and CICS-style scheduling decides the demand side, but joint

optimization of the two is largely unexplored; the procurement literature treats demand as fixed, while the scheduling literature takes the supply portfolio as given [13,22]. Gap 2: co-scheduling delay-tolerant training with latency-sensitive inference. A single framework that respects both energy signals and inference SLOs, with training as the flexible resource that absorbs forecast errors created by rigid inference demand, is still missing. Gap 3: embedding realistic power-system constraints into CES algorithms. Network limits, storage state dynamics, and price/carbon volatility are often abstracted away in existing coordinated-scheduling models [41–45]; closing this gap requires co-design between power-system and computer-systems research.

4.5. Implementation Cases, International Perspectives and Adoption Barriers

Although large-scale, fully integrated CES does not yet exist, several implementations approximate its building blocks. Google’s carbon-intelligent computing system is the most documented demand-side case: carbon-aware capacity curves time-shift flexible workloads across hours based on next-day carbon-intensity forecasts, in production since 2020–2021 and reported to shift substantial flexible compute to cleaner hours [13]. Chinese coordinated-scheduling studies have modelled data-center demand management and bidirectional optimization between computing nodes and power nodes under the country’s “Eastern Data, Western Computing” topology, demonstrating cost and renewable-utilization gains in simulation and pilot settings [42–44]. Voluntary corporate procurement in the United States shows the supply side of the synergy: power purchase agreements at hyperscale, and the shift of leading operators from annual towards hourly 24/7 carbon-free matching, which system-level studies identify as the procurement strategy with consistent emission impact [16,22]. Regulatory approaches differ: China’s program is state-led and plan-coordinated through national computing-network and power-sector planning [14,40]; the United States relies on corporate procurement and electricity-market signals [22]; and the European Union is regulation-led, with mandatory energy and water reporting for data centers and national efficiency requirements (for example, binding PUE-oriented rules in Germany) that push facilities towards measurable efficiency targets [6]. These international perspectives suggest that CES can be entered from either the demand side (scheduling), the supply side (procurement), or the regulatory side (reporting and efficiency mandates), but that its full value is realized only when the three are aligned.

Adoption barriers are correspondingly economic and institutional as much as technical. Split incentives: the data-center operator bears the cost of flexible scheduling and on-site storage, while much of the benefit, reduced curtailment, deferred grid reinforcement, accrues to the power system and is not compensated unless market rules reward demand-side participation [41]. Market access: computing loads are not uniformly eligible to offer flexibility products in electricity markets, and participation rules were designed for conventional industrial demand response [45]. Data confidentiality: effective coordination requires the power system to learn about computing schedules and vice versa, raising commercially sensitive information issues that privacy-preserving scheduling only partially resolves [45]. Regulatory and price uncertainty: volatile electricity and carbon prices make long-term contracts costly to price, and inconsistent carbon-pricing coverage across jurisdictions distorts siting decisions [40]. Overcoming these barriers is a policy-design problem as much as an algorithm-design problem, and it interacts with the rebound and accounting issues discussed in Section 5.2.

5. Discussion

With the acceleration of digital transformation, the demand for computing power is surging, and computing power has become a crucial metric for assessing a nation’s technological capability and overall competitiveness. This section discusses large-model-specific quantitative evidence (Section 5.1), then offers critical analyses that the survey literature usually leaves implicit: efficiency growth and its rebound (Section 5.2), carbon accounting and offsetting (Section 5.2), water and other externalities (Section 5.3), international policy differences (Section 5.4), and the limitations both of existing approaches and of this survey itself (Section 5.5), before outlining development trends (Section 5.6).

5.1. Large Model-Specific Analysis

Assessing the sustainability of LLMs requires a holistic evaluation of their environmental impact across the entire lifecycle, from hardware manufacturing through training, inference, deployment, and decommissioning. Because of transparency limitations, research has focused on measurable energy consumption in the training and inference phases, converting greenhouse-gas emissions into CO₂ equivalents for comparability [33]. Targeting high-performance computing (HPC), the open-source Python package perun evaluates the energy consumption of HPC applications by measuring power usage, memory utilization, and response time during execution [46]. The energy consumption of a compute node is calculated with:

$$E_{\text{node}} = \int P_{\text{cpu}}(t) + P_{\text{gpu}}(t) + P_{\text{mem}}(t) + P_{\text{other}}(t) dt \quad (2)$$

where $P(\cdot)$ denotes the power draw of individual hardware components and the terms are integrated over the operating interval. With PUE, the power-usage-effectiveness factor of the data center (Section 3.3), the total energy consumption of an HPC application running on multiple nodes is:

$$E_{\text{total}} = \text{PUE} \times \sum E_{\text{node}}^i \quad (3)$$

Training-side measurements complement this picture. Patterson et al. [34] estimated that training GPT-3 consumed about 1287 MWh of energy and produced about 552 tCO₂e of emissions, and Luccioni et al. [33] measured about 433 MWh and roughly 25 tCO₂e (operational) for the 176B-parameter BLOOM model, figures that are material but one-off per model, whereas the inference phase of a deployed model repeats its energy and water cost on every query and, over a service lifetime, dominates the lifecycle footprint [29,35]. This asymmetry justifies the workload-class separation of Section 4.2: training flexibility is about when and where to spend large one-off blocks, while inference efficiency is about the recurring cost curve.

Based on perun measurements, the total CO₂-equivalent emissions from processing a subset of the MMLU dataset [47] for typical open-weight LLMs are shown in Figure 5. Two patterns stand out. First, emissions scale with model size at fixed accuracy targets, which is why the compression techniques of Section 3.2 translate directly into carbon savings. Second, reasoning-capable variants generate several times higher per-query emissions than their non-reasoning counterparts, because longer generated outputs multiply token-level energy [29,32]. This makes a strong case for token-efficiency techniques and for reporting energy per token as a standard metric, as advocated in Section 3.3.

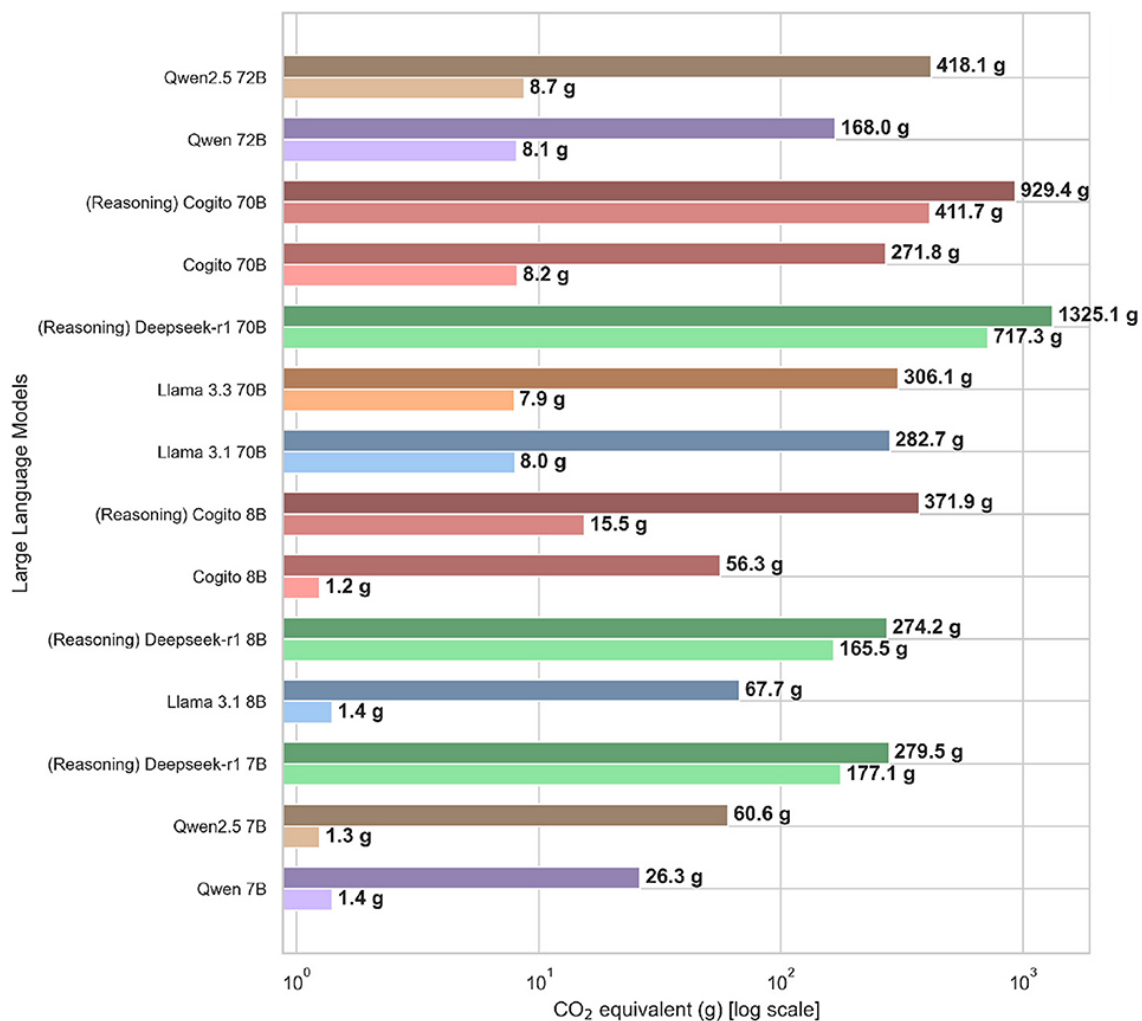


Figure 5. Per-request inference CO₂-equivalent emissions (log scale, grams) of selected open-weight LLMs, with and without reasoning, measured with perun [46] while processing a subset of the MMLU dataset [47]; source: Dauner and Socher [32].

5.2. Critical Analysis I: Efficiency, the Rebound Effect, and Carbon Accounting

The rebound effect. A recurring assumption in the green-computing literature is that per-computation efficiency gains translate into lower aggregate energy use. The evidence reviewed above cautions otherwise. Data-center electricity consumption grew by roughly 12% per year over the past five years even as efficiency per computation improved steadily [6], and economic history, from Jevons' observation on coal onward, suggests that cheaper computation stimulates demand: efficiency gains in AI compute have so far been reinvested into larger models, longer contexts and wider deployment rather than into absolute reductions [30]. Efficiency techniques (Section 3.2) therefore function as a brake on growth, with a brake pedal held by demand; they are necessary for sustainability but demonstrably insufficient, which strengthens, rather than weakens, the case for coordinating computing demand with low-carbon supply through CES. Reviews that report per-model savings without asking what happens to aggregate demand overstate their practical contribution.

Carbon accounting and verification. Much of the discrepancy between corporate claims and system-level outcomes traced in Section 3.1 is an accounting artefact. Under market-based Scope 2 accounting of the GHG Protocol [48], a data center can report near-zero emissions by buying unbundled renewable certificates equal to its consumption, even while drawing fossil power hour by hour; system-level studies find that such annual volumetric matching yields little additional emission reduction [22]. Hourly-matched 24/7 CFE accounting aligns claims with physical reality far better but is more expensive and still voluntary. Carbon offsetting faces analogous quality problems: offsets vary widely in permanence, additionality and leakage, and using them to neutralize continuing operational emissions embeds those problems in the AI supply chain. Three implications follow for the field: first, papers should state explicitly which accounting basis (location-based, market-based, or hourly-matched) underlies reported footprints; second, third-party verification of AI carbon claims remains immature and should precede regulation-by-reporting; and third, the hourly CFE percentage should be reported as a default metric alongside PUE and energy per token (Section 3.3).

5.3. Critical analysis II: Water and Other Environmental Externalities

Carbon dominates the discussion, but it is not the only externality. Water is the most under-reported: LLM inference consumes water both indirectly, through the electricity mix, and directly, through evaporative cooling, and inference-scale studies now report water footprints per query that, multiplied across billions of daily queries, amount to substantial volumes, concentrated in the very arid regions where solar-rich siting is attractive [7,20,21]. Data centers in the United States were estimated to consume water on the order of hundreds of billions of litres per year directly and indirectly [20], and siting decisions that optimize for carbon can therefore worsen water stress, a trade-off that single-metric optimization misses and that the joint-metrics view of Section 3.3 is designed to expose. Embodied carbon in accelerators and the e-waste stream from rapid hardware refresh cycles are likewise growing with fleet turnover [35], and land, noise, and community impacts of hyperscale campuses increasingly provoke local opposition that can delay or cancel projects. The practical lesson for large-model computing is that multi-criteria siting and reporting- carbon, water, materials- is not environmental decoration but a risk-management requirement.

5.4. International Policy Perspectives

The policy environment for computing-energy coordination is fragmenting along regional lines, with consequences for where large-model computing locates and how CES can be implemented. China pursues state-led coordination: national computing-network planning ("Eastern Data, Western Computing"), green-electricity and green-certificate trading, and coordinated development programmes for computing and electricity [14,40]. The United States relies primarily on market signals and voluntary corporate procurement, with power purchase agreements and the emerging practice of hourly 24/7 carbon-free matching [16,22]. The European Union is regulation-led: mandatory energy and water reporting for data centers under energy-efficiency legislation, and national rules such as Germany's binding efficiency requirements, push operators towards measurable targets regardless of procurement strategy [6]. For the CES agenda, this heterogeneity cuts both ways: it creates natural experiments from which researchers can extract comparative evidence, but it also means that scheduling algorithms, procurement instruments and reporting standards are not portable across jurisdictions, and that carbon- or price-driven load shifting may simply relocate emissions across regulatory boundaries rather than eliminate them. Comparative studies across these three regimes are scarce and constitute a research gap in their own right.

5.5. Limitations of Existing Approaches and of This Survey

Limitations of the approaches reviewed. On the supply side, annual renewable matching is weakly additional, and even 24/7 CFE remains voluntary, costly, and dependent on hourly metering infrastructure that many markets lack [22]. On the efficiency side, reported savings are rarely reproduced across hardware and serving stacks, benchmarks can be gamed, and rebound effects erode aggregate gains (Section 5.2). On the CES side, the literature overwhelmingly reports simulation or pilot studies under abstracted constraints, network limits, storage state, price volatility and, crucially, adversarial or multi-stakeholder behaviors are seldom modelled (Sections 4.3 and 4.4), and no fully integrated national-scale CES deployment yet exists to validate the coordinated-scheduling gains claimed in simulation [41–43]. On the metrics side, site dependence and the absence of standardized reporting make cross-study comparison hazardous (Section 3.3).

Limitations of this survey. This is a narrative review, not a systematic one: no PRISMA-style screening statistics are reported, study quality was not formally scored, and coverage is limited to English- and Chinese-language literature through September 2025; given the pace of the field, some preprint results cited here may be revised or superseded. The quantitative evidence assembled is heterogeneous in metrics and sites, which prevents meta-analysis; we have therefore confined quantitative comparison to internally consistent sources (for example, the perun-based measurements of Figure 5). The engineering-constraint analysis (Section 4.4) and the conceptual framework (Figure 1) are qualitative syntheses, not validated models. These limitations are reflected in the research gaps of Section 4.4.

5.6. Development Trends and Policy Directions

Building on new-type power systems, efforts are directed toward reducing energy consumption and carbon emissions across computing infrastructure. Four priorities follow from the evidence reviewed. First, increase the share of verifiable green energy in data centers, integrating green-energy usage into energy-saving evaluations and moving from annual matching toward hourly 24/7 CFE accounting, which is the procurement strategy most consistently associated with real emission reductions [22]. Second, deepen the integration of computing networks and power grids by exploring direct green-supply models such as local delivery, on-site consumption and aggregated trading, so that computing infrastructure obtains a stable renewable supply. Third, unlock demand-side flexibility along the workload taxonomy of Section 4.2, treating training and offline inference as dispatchable resources in electricity markets and ancillary-service programmes [14]. Fourth, provide institutional support: a multidimensional policy system covering organizational coordination, monitoring and evaluation, pilot projects, incentive mechanisms and risk prevention, together with a clear long-term roadmap, is a precondition for large-scale CES deployment [40].

6. Conclusions

This review synthesized recent evidence on green energy, green computing and computing-energy synergy for large-model computing. Four evidence-based conclusions emerge. First, green energy provides low-carbon electricity for large-model computing and reduces operational emissions, but annual renewable matching alone does not guarantee continuous carbon-free operation; hourly 24/7 carbon-free-energy matching is currently the procurement strategy most consistently associated with system-level emission reductions, at a non-trivial cost premium [22]. Second, green computing reduces energy per unit of computation through efficient architectures, model compression and system-level techniques, with trade-offs in accuracy or latency that must be made explicit and measured with standardized metrics (Tables 3 and 4) [28]; because efficiency gains are partly reinvested into larger models and wider deployment, these gains brake rather than reverse aggregate demand growth [6,30]. Third, CES couples power flow and data flow for the joint optimization of energy and computing resources; given that data-center electricity demand is projected to more than double by 2030 [6], supply-side and efficiency measures alone are unlikely to suffice, and the workload flexibility of training and batch inference makes CES a promising coordination paradigm. Fourth, the feasibility of CES at scale depends on engineering constraints, in particular renewable-supply uncertainty, storage limits, grid congestion, workload deadlines, latency requirements and price volatility, which the literature has only begun to address (Section 4.4).

CES should therefore be seen as one important pathway among several, alongside green-energy deployment, efficiency improvement and siting strategy, whose distinctive value lies in unlocking demand-side flexibility that the other pathways cannot provide; claims that it is the only path toward sustainable computing are not supported by the current evidence. This review has limitations: it is narrative rather than systematic; the field is fast-moving; and cross-study comparison is hindered by heterogeneous metrics (Section 3.3). Future work should pursue the three research gaps identified in Section 4.4: quantitative co-optimization of 24/7 carbon-free procurement and

workload scheduling, unified frameworks for delay-tolerant and latency-sensitive co-scheduling, and CES algorithm design under realistic power-system constraints, so that the coordination of computing and energy systems can be assessed on measurable, comparable, and reproducible grounds.

Author Contributions

J.H.: conceptualization, methodology, investigation, and original draft preparation; Y.T.: data curation. All authors have read and agreed to the published version of the manuscript.

Funding

The work was partially supported by the Hunan Provincial Key Research and Development Program Project (2025QK3018) and the Hunan Provincial Major Science and Technology Project (2026QK2004).

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

This manuscript is a review, and no novel experimental data, raw datasets, or samples were generated during the preparation of this paper. All information and evidence discussed in this work derive from previously published studies listed in the reference section. All cited publications are publicly accessible.

Conflicts of Interest

J.H. was employed by Hunan Chaoneng Robot. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used AI tools to assist with structural drafting and editorial refinement. After using AI tools, the authors reviewed and edited all content as needed and take full responsibility for the content of the publication.

References

1. Bommasani, R.; Hudson, D.A.; Adeli, E.; et al. On the Opportunities and Risks of Foundation Models. *arXiv* **2021**, arXiv:2108.07258.
2. Calvin, K.; Dasgupta, D.; Krinner, G.; et al. Climate change 2023: Synthesis report. In *Contribution of Working Groups I, II and III to the Sixth Assessment Report*; IPCC: Geneva, Switzerland, 2023.
3. *Greenhouse Gas Bulletin No. 21: The State of Greenhouse Gases in the Atmosphere Based on Global Observations through 2024*; World Meteorological Organization, Geneva, Switzerland, 2025.
4. Global Renewables and Energy Efficiency Pledge. COP28 UAE. December 2023. <https://www.cop28.com/en/global-renewables-and-energy-efficiency-pledge> (accessed on 15 January 2026).
5. Hao, J.; Tao, Y. Power grid inspection based on multimodal foundation models. *EAI Endorsed Trans. Energy Web* **2025**, 12, 9087. <https://doi.org/10.4108/ew.9087>.
6. *Energy and AI*. International Energy Agency; International Energy Agency: Paris, France, 2025. Available online: <https://www.iea.org/reports/energy-and-ai> (accessed on 20 February 2026).
7. Jegham, N.; Abdelatti, M.; Koh, C.Y.; et al. How Hungry Is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference. *arXiv* **2025**, arXiv:2505.09598.
8. Xu, J.; Zhou, W.; Fu, Z.; et al. A survey on green deep learning. *arXiv* **2021**, arXiv:2111.05193.
9. Zhou, Y.; Lin, X.; Zhang, X.; et al. On the opportunities of green computing: A survey. *arXiv* **2023**, arXiv:2311.00447.
10. Bolón-Canedo, V.; Morán-Fernández, L.; Cancela, B.; et al. A review of green artificial intelligence: Towards a more sustainable future. *Neurocomputing* **2024**, 599, 128096. <https://doi.org/10.1016/j.neucom.2024.128096>.
11. Boscariol, M.; Cacciaguerra, E.; Gasbarri, P.; et al. A review of current practices and challenges in green data centers: Renewable energy sources, waste heat recovery, and intelligent management systems. In Proceedings of the 2025 33rd

- Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP), Turin, Italy, 12–14 March 2025; pp. 478–485. <https://doi.org/10.1109/pdp66500.2025.00074>.
12. Lin, W.; Lin, J.; Peng, Z.; et al. A systematic review of green-aware management techniques for sustainable data center. *Sustain. Comput. Inform. Syst.* **2024**, *42*, 100989. <https://doi.org/10.1016/j.suscom.2024.100989>.
 13. Radovanovic, A.; Koningstein, R.; Schneider, I.; et al. Carbon-aware computing for datacenters. *IEEE Trans. Power Syst.* **2023**, *38*, 1270–1280. <https://doi.org/10.1109/tpwrs.2022.3173250>.
 14. China Academy of Information and Communications Technology. Professional Committee for Cooperative Development of Electric Power and Computing Power 2024. In *Blue Book on Cooperative Development of Electric Power and Computing Power*; China Academy of Information and Communications Technology: Beijing, China, 2024.
 15. Sayed, E.; Olabi, A.; Alami, A.; et al. Renewable energy and energy storage systems. *Energies* **2023**, *16*, 1415. <https://doi.org/10.3390/en16031415>.
 16. Kwon, T.; Kim, S. Pioneering renewable energy transition in US: A comparative case study of Google, Apple, and Microsoft's utilization of power purchase agreements. In *Cultural Perspectives of Human-Centered and Technological Innovations, Proceedings of the International Workshop on Cultural Perspectives of Human-Centered and Technological Innovations, Tokyo, Japan, 27–28 January 2024*; Springer: Cham, Switzerland, 2024; pp. 237–249.
 17. Gayen, D.; Chatterjee, R.; Roy, S. A review on environmental impacts of renewable energy for sustainable development. *Int. J. Environ. Sci. Technol.* **2024**, *21*, 5285–5310. <https://doi.org/10.1007/s13762-023-05380-z>.
 18. Khalid, M. Smart grids and renewable energy systems: Perspectives and grid integration challenges. *Energy Strategy Rev.* **2024**, *51*, 101299. <https://doi.org/10.1016/j.esr.2024.101299>.
 19. Small Nuclear Power Reactors. World Nuclear Association. 2025. Available online: <https://world-nuclear.org> (accessed on 5 March 2026).
 20. Siddik, M.A.B.; Shehabi, A.; Marston, L. The environmental footprint of data centers in the United States. *Environ. Res. Lett.* **2021**, *16*, 064017. <https://doi.org/10.1088/1748-9326/abfb1>.
 21. Mytton, D. Data centre water consumption. *npj Clean Water* **2021**, *4*, 11. <https://doi.org/10.1038/s41545-021-00101-w>.
 22. Xu, Q.; Ricks, W.; Manocha, A.; et al. System-level impacts of voluntary carbon-free electricity procurement strategies. *Joule* **2024**, *8*, 374–400. <https://doi.org/10.1016/j.joule.2023.12.007>.
 23. Paul, S.G.; Saha, A.; Arefin, M.S.; et al. A comprehensive review of green computing: Past, present, and future research. *IEEE Access* **2023**, *11*, 87445–87494. <https://doi.org/10.1109/access.2023.3304332>.
 24. Chidolue, O.; Ohenhen, P.E.; Umoh, A.A.; et al. Green data centers: Sustainable practices for energy-efficient IT infrastructure. *Eng. Sci. Technol. J.* **2024**, *5*, 99–114. <https://doi.org/10.51594/estj.v5i1.730>.
 25. Singh, A.; Patel, N.; Ehtesham, A.; et al. A survey of sustainability in large language models: Applications, economics, and challenges. In *Proceedings of the Annual Computing and Communication Workshop and Conference, Las Vegas, NV, USA, 6–8 January 2025*; pp. 8–14.
 26. Google. *Environmental Report 2025*; Google: Mountain View, CA, USA, July 2025.
 27. Microsoft. *2025 Environmental Sustainability Report*; Microsoft: Redmond, WA, USA, May 2025.
 28. Stojkovic, J.; Choukse, E.; Zhang, C.; et al. Towards greener LLMs: Bringing energy-efficiency to the forefront of LLM inference. *arXiv* **2024**, arXiv:2403.20306.
 29. Luccioni, S.; Jernite, Y.; Strubell, E. Power hungry processing: Watts driving the cost of AI deployment? In *Proceedings of the FAccT '24: The 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, Brazil, 3–6 June 2024*; pp. 85–99. <https://doi.org/10.1145/3630106.3658542>.
 30. Desislavov, R.; Martínez-Plumed, F.; Hernández-Orallo, J. Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. *Sustain. Comput. Inform. Syst.* **2023**, *38*, 100857. <https://doi.org/10.1016/j.suscom.2023.100857>.
 31. Avelar, V.; Azevedo, D.; French, A. (Eds.) *PUE: A Comprehensive Examination of the Metric*; The Green Grid: Washington, DC, USA, 2012.
 32. Dauner, M.; Socher, G. Energy costs of communicating with AI. *Front. Commun.* **2025**, *10*, 1572947. <https://doi.org/10.3389/fcomm.2025.1572947>.
 33. Luccioni, A.; Viguier, S.; Ligozat, A. Estimating the carbon footprint of BLOOM, a 176B parameter language model. *J. Mach. Learn. Res.* **2023**, *24*, 1–15.
 34. Patterson, D.; Gonzalez, J.; Le, Q.; et al. Carbon emissions and large neural network training. *arXiv* **2021**, arXiv:2104.10350.
 35. Wu, C.; Raghavendra, R.; Gupta, U.; et al. Sustainable AI: Environmental implications, challenges and opportunities. *arXiv* **2022**, arXiv:2111.00364.
 36. Ajagekar, A.; You, F. Quantum computing and quantum artificial intelligence for renewable and sustainable energy: A emerging prospect towards climate neutrality. *Renew. Sustain. Energy Rev.* **2022**, *165*, 112493. <https://doi.org/10.1016/j.rser.2022.112493>.

37. Jaschke, D.; Montangero, S. Is quantum computing green? An estimate for an energy-efficiency quantum advantage. *Quantum Sci. Technol.* **2023**, *8*, 025001. <https://doi.org/10.1088/2058-9565/acae3e>.
38. Ricciardi Celsi, M.; Ricciardi Celsi, L. Quantum computing as a game changer on the path towards a net-zero economy: A review of the main challenges in the energy domain. *Energies* **2024**, *17*, 1039. <https://doi.org/10.3390/en17051039>.
39. Li, J.; Wang, Y.; Jiang, L. Study on definition and key measures of computing and electric power coordinated development. *Inf. Commun. Technol. Policy* **2025**, *51*, 2–8. <https://doi.org/10.12267/j.issn.2096-5931.2025.02.001>.
40. China Academy of Information and Communications Technology. *Research Report on Cooperative Development of Computing Power and Electric Power (2025)*; China Academy of Information and Communications Technology: Beijing, China, 2025.
41. Yang, P.; Yu, S.; Ren, F. Collaborative innovation framework for computing power and electricity. *Control. Theory Appl.* **2024**, *41*, 1181–1186. <https://doi.org/10.7641/cta.2024.30589>.
42. Ding, Z.; Cao, Y.; Zhang, S.; et al. Coordinated operation for data center and power system in the context of energy Internet (I): Energy demand management model of data center. *Proc. CSEE* **2022**, *42*, 3161–3176. <https://doi.org/10.13334/j.0258-8013.pcsee.210813>.
43. Cao, Y.; Ding, Z.; Wang, P.; et al. Coordinated operation for data center and power system in the context of energy Internet (II): Opportunities and challenges. *Proc. CSEE* **2022**, *42*, 3512–3527. <https://doi.org/10.13334/j.0258-8013.pcsee.210814>.
44. Zhou, Q.; Yang, P.; Wan, S.; et al. Bidirectional collaborative optimization scheduling of adjustable resources in computing node and power node. *Electr. Power Constr.* **2025**, *46*, 13–25. <https://doi.org/10.12204/j.issn.1000-7229.2025.02.002>.
45. Sun, Y.; Ding, Z.; Yan, Y.; et al. Privacy-preserving energy sharing among cloud service providers via collaborative job scheduling. *IEEE Trans. Smart Grid* **2025**, *16*, 1168–1180. <https://doi.org/10.1109/tsg.2024.3482390>.
46. Muriedas, J.P.G.H.; Flügel, K.; Debus, C.; et al. *perun*: Benchmarking energy consumption of high-performance computing applications. In *Euro-Par 2023: Parallel Processing, Proceedings of the 29th International Conference on Parallel and Distributed Computing, Limassol, Cyprus, 28 August–1 September 2023*; Springer: Cham, Switzerland, 2023; pp. 17–31.
47. Hendrycks, D.; Burns, C.; Basart, S.; et al. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations, Vienna, Austria, 3–7 May 2021*.
48. Ranganathan, J.; Bhatia, P. *The Greenhouse Gas Protocol: A Corporate Accounting and Reporting Standard*, revised ed.; World Resources Institute and WBCSD: Washington, DC, USA, 2004.