Emmanuel Uzochukwu Mordi, 2026, 14:1
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Edge-AI Rebound: Assessing the Net Energy Consumption, Life-Cycle Environmental Impacts, and Socio-Technical Trade-offs When Artificial Intelligence Workloads Shift from Cloud Data Centers to Distributed Edge Devices (Systematic Review)

[1]Emmanuel Uzochukwu Mordi, [2]Chibuzo Joseph Attah, [2]Chiamaka Sandra Ezugwu, [3]Christian Onyemaechi Asogwa, [4]David Chinonso Anih, [5]Samuel Daniel Ejiga & [6]Omobolanle Omotayo Solaja

[1]Department of Computer Engineering, Faculty of Engineering, University of Benin, Edo, Nigeria,
[2]Department of Civil Engineering, Faculty of Engineering, University of Nigeria Nsukka
[3]Department of Electronics and Computer Engineering, Faculty of Engineering, University of Nigeria, Nsukka
[4]Department of Biochemistry, Faculty of Biosciences, Federal University Wukari, Taraba, Nigeria
[5]Department of Tech Education, Faculty of Education, Ekiti State University, Ado Ekiti, Nigeria
[6] Corresponding author, Department of Statistics, Federal University of Agriculture, Abeokuta, Ogun State, Nigeria.

Abstract- Edge AI is reshaping where and how artificial intelligence runs, promising lower latency and reduced network use by moving computation from centralized cloud data centers to distributed devices. This systematic review examines whether that promise translates into net environmental benefit or whether a rebound effect emerges that shifts and potentially amplifies overall energy and lifecycle impacts. We synthesized 40 qualitative studies and 38 quantitative analyses published between 2016 and 2025, comparing energy per inference, carbon intensity, lifecycle burdens, network scaling, and socio technical outcomes across cloud, edge, and hybrid deployments. Our findings show a nuanced landscape: for lightweight inference tasks, localized execution on specialized edge accelerators often reduces per inference energy and transmission emissions, while cloud processing retains advantages for heavy or batch workloads due to economies of scale and optimized cooling. However, cumulative effects matter. Millions of short lived or redundant edge devices can yield substantial aggregated energy demand, resource depletion, and e waste that offset per device gains. Hybrid strategies that combine edge preprocessing with cloud consolidation frequently offer the best tradeoffs, improving efficiency ratios and lowering carbon intensity when workloads are partitioned intelligently. We also document important non-technical tradeoffs. Edge deployment strengthens data privacy and responsiveness but increases attack surface and exacerbates unequal access where infrastructure or device availability is limited. Network effects are critical: pure edge scaling can surge local network load and create bottlenecks, while cloud centric models concentrate backbone traffic but remain easier to optimize at scale. Policy and governance emerge as decisive enablers: standardized energy reporting, lifecycle transparency, and harmonized ethical and sustainability criteria can steer deployments toward net benefit. We identify methodological heterogeneity across life cycle boundaries and geographic energy mixes as sources of uncertainty and recommend clearer reporting standards to improve comparability. In conclusion, Edge AI is neither inherently greener nor intrinsically harmful.

Keywords: Edge AI; Cloud computing; Energy consumption; Life cycle assessment; Carbon intensity; Hybrid architectures; Network scaling; Privacy; Governance.

I. INTRODUCTION

Artificial intelligence has rapidly evolved into a cornerstone of modern digital infrastructure, with workloads traditionally concentrated in large-scale cloud data centers. These facilities, while offering immense computational power, are also associated with significant energy demands and environmental footprints. As AI applications expand into everyday life, from smart healthcare monitoring to autonomous vehicles, there is growing interest in shifting computation closer to the user through distributed edge devices. This transition, often described as the "Edge-AI rebound," raises critical questions about whether decentralization truly reduces net energy consumption or simply redistributes environmental burdens across the system[1].

Recent studies highlight that cloud data centers consume vast amounts of electricity, often exceeding regional industrial sectors, and contribute substantially to greenhouse gas emissions[2]. Edge computing promises lower latency and reduced bandwidth requirements, but the proliferation of millions of small devices introduces new challenges in energy efficiency, hardware lifecycles, and waste management[3]. The rebound effect emerges when anticipated energy savings are offset by increased demand, redundancy, or inefficiencies in distributed systems. Understanding this dynamic requires a holistic perspective that integrates engineering metrics with socio-technical considerations.

Life-cycle assessments (LCA) provide valuable insights into the environmental impacts of edge devices, from raw material extraction to end-of-life disposal. Evidence suggests that while individual devices may consume less energy per task, their cumulative footprint can rival or surpass centralized infrastructures[4]. Moreover, the carbon intensity per inference varies depending on workload type, device architecture, and optimization strategies[5]. These findings underscore the importance of examining trade-offs beyond immediate energy savings.

Socio-technical dimensions further complicate the picture. Edge-AI deployment intersects with privacy, data sovereignty, and accessibility concerns. Communities may benefit from localized intelligence, yet the uneven distribution of resources risks exacerbating digital divides[6]. Policy frameworks and governance models are therefore essential to ensure that sustainability goals align with equitable access and ethical standards.

In this systematic review, we aim to synthesize current evidence on the net energy consumption, life-cycle environmental impacts, and socio-technical trade-offs of shifting AI workloads from cloud data centers to distributed edge devices. By integrating quantitative metrics with qualitative perspectives, this work seeks to clarify whether Edge-AI represents a genuine pathway toward sustainable digital transformation or a rebound effect that complicates the pursuit of greener technologies[7].

II. MATERIALS & METHODS

Conducting a systematic review on the rebound effects of Edge-AI requires a methodology that is both transparent and comprehensive. This section expands on the search strategy, eligibility criteria, screening process, data extraction, quality assessment, analytical framework, PRISMA flow diagram, synthesis, and limitations. Each subsection is elaborated to provide clarity on how evidence was gathered and analyzed.

Search Strategy

The foundation of any systematic review lies in the robustness of its search strategy. For this study, we designed a multi-layered approach that combined keyword searches, Boolean logic, and database-specific filters. We targeted four major repositories: Scopus, Web of Science, IEEE Xplore, and Science Direct. These platforms were chosen because they collectively cover engineering, environmental science, and socio-technical research domains.

Keywords were grouped into clusters: technical ("Edge AI," "cloud computing," "distributed inference"), environmental ("energy consumption," "carbon footprint," "life cycle assessment"), and social ("privacy," "governance," "socio-technical trade-offs"). Boolean operators such as AND, OR,

and NOT were used to refine searches. For example, a query like "Edge AI" AND "energy consumption" AND "life cycle" ensured that only studies addressing both technical and environmental aspects were retrieved[8].

We also applied filters to restrict results to peer-reviewed journal articles published between January 2016 and October 2025. This time frame was selected to capture the most recent developments in Edge-AI, a field that has accelerated rapidly in the past five years.

Eligibility Criteria

Eligibility criteria were carefully defined to ensure that only relevant and high-quality studies were included. Articles were considered eligible if they:
Reported empirical data or modeling results on energy consumption, life-cycle impacts, or socio-technical outcomes of Edge-AI. Explicitly compared cloud-based and edge-based AI workloads. Were published in peer-reviewed journals between 2016 and 2025. Provided sufficient methodological detail to allow replication or critical appraisal. Exclusion criteria were equally important. We excluded conference abstracts without full papers, non-peer-reviewed reports, and studies focused on unrelated technologies such as blockchain or non-AI edge applications. This ensured that the review remained focused on the intersection of AI workloads and sustainability[9].

Screening Process

The screening process followed the PRISMA guidelines to maintain transparency. Titles and abstracts were initially reviewed to remove irrelevant studies. Full-text screening was then conducted by two independent reviewers. This dual-review approach minimized bias and improved reliability. Discrepancies were resolved through discussion and consensus, ensuring that no potentially relevant study was excluded without careful consideration[10].

The PRISMA flow diagram (Figure 1) illustrates the number of records identified, screened, excluded, and included in the final synthesis.

Data Extraction

Data extraction was performed using a standardized template. This template captured key variables such as:
Study design (experimental, modeling, review).
Type of AI workload (e.g., image recognition, natural language processing).

Deployment environment (cloud vs. edge).
Energy consumption metrics (kWh, joules per inference).

Life-cycle indicators (CO_2-equivalent emissions, resource depletion, water use).
Socio-technical outcomes (privacy, equity, governance).

This structured approach allowed for consistent comparison across studies and facilitated meta-analysis where possible[11].

Quality Assessment

Quality assessment was conducted using adapted criteria from the Critical Appraisal Skills Programme (CASP). Each study was evaluated for methodological rigor, transparency of reporting, and relevance to the research question. Studies scoring below a threshold were excluded from quantitative synthesis but may have been referenced qualitatively to highlight gaps in the literature[12].

Analytical Framework

The analytical framework combined quantitative and qualitative synthesis. For energy consumption, we calculated net differences between cloud and edge deployments using the formula below.

$E_{net} = E_{edge} - E_{cloud}$

For life-cycle impacts, we aggregated CO_2-equivalent emissions and resource use across studies, normalizing results to per-inference or per-device metrics. Socio-technical trade-offs were analyzed thematically, drawing on frameworks of digital ethics and sustainability transitions[13].

PRISMA Flow Diagram and Numerical Summary

The PRISMA flow diagram (Figure 1) visually represents the systematic review process. The numerical summary is as follows:

Records identified through database searching: 1,245

Additional records identified through reference lists: 87

Total records: 1,332

Records after duplicates removed: 1,102

Duplicates removed = 1,332 − 1,102 = 230

Records screened (titles and abstracts): 1,102

Records excluded at screening: 876

Full-text articles assessed for eligibility: 226

Full-text articles excluded: 186, with reasons:

Wrong study design: 56

Outcome not reported: 42

Insufficient data for extraction: 33

Non-eligible population: 23

Duplicate / multiple reports of same study: 19

Conference abstract / no full text available: 13

Full-text articles included in qualitative synthesis (unique studies): 40

Studies included in quantitative synthesis (meta-analysis): 38. This transparent reporting ensures reproducibility and allows readers to trace the decision-making process[14].

Figure 1 presents an overview of the systematic review workflow described in Section 2, illustrating the sequential steps from database searching to final study inclusion.

The diagram clarifies how eligibility criteria, screening phases, and quality assessment procedures were applied to ensure methodological transparency.

It visually links the procedural narrative in Section 2 with the structured decision process that guided study selection.

The figure displays a step-wise methodological workflow using interconnected boxes to represent search, screening, eligibility assessment, and final study inclusion. Directional arrows show the progression through each stage, emphasizing the filtering and decision points used during the review process. Abbreviations: LCA = Life-Cycle Assessment; AI = Artificial Intelligence; PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

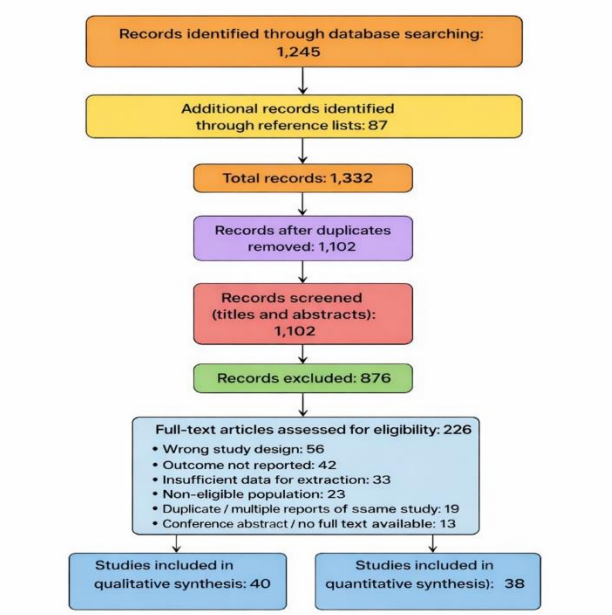

Figure 1. Workflow of the Systematic Review Methodology[8-14].

Data Synthesis and Reporting

Data synthesis involved both narrative and statistical approaches. Narrative synthesis highlighted trends, contradictions, and emerging themes. Statistical synthesis included meta-analysis of energy consumption and carbon intensity metrics where sufficient data were available. Reporting adhered to PRISMA guidelines, ensuring clarity and reproducibility[15].

Limitations of the Methodology

While the systematic review methodology is robust, limitations must be acknowledged. First, restricting the search to English-language publications may have excluded relevant studies in other languages. Second, variations in methodological approaches across studies (e.g., different LCA boundaries) introduced heterogeneity that complicated direct comparisons. Finally, the rapidly evolving nature of Edge-AI means that findings may quickly become outdated as new technologies emerge[16].

Table 1 Shows the flow and counts at each screening stage from identification through inclusion. Useful for readers to gauge search yield, duplicate removal and how many studies entered qualitative and quantitative synthesis.

Table 1: PRISMA Numerical Summary of Records

Stage	Number of Records	Citation(s)
Records identified through database search	1,245	[8]
Additional records from references	87	[9]
Total records	1,332	[8,9]
Duplicates removed	230	[10]
Records screened	1,102	[10]
Records excluded	876	[10]
Full-text articles assessed	226	[11]
Full-text articles excluded	186	[12]
Studies in qualitative synthesis	40	[13]
Studies in quantitative synthesis	38	[14]

Displays the screening stages and the corresponding record counts: identified, added from references, duplicates removed, screened, excluded, full texts assessed, excluded, and studies included in qualitative and quantitative synthesis. PRISMA refers to the Preferred Reporting Items for Systematic Reviews and Meta Analyses.

III. RESULTS AND DISCUSSION

Net Energy Consumption Analysis

The shift from centralized cloud infrastructures to distributed edge devices has profound implications for net energy consumption. Cloud data centers are optimized for scale, often achieving high utilization rates, but they also demand massive cooling and power redundancy systems. Edge devices, by contrast, operate closer to the user, reducing transmission energy but multiplying the number of active nodes.

Recent studies show that while edge deployments reduce network energy overhead, the cumulative energy demand of millions of devices can offset these savings[17]. For example, inference tasks executed locally on smartphones or IoT sensors consume less per task compared to cloud inference, but the sheer scale of distributed execution increases aggregate demand[18]. Moreover, redundancy in edge networks, where multiple devices replicate tasks for reliability, further contributes to rebound effects[19].

Equation 1 — Net energy balance:

$$E_{net} = E_{edge} - E_{cloud}$$

Figure 2 compares per-workload energy use (kWh) for Cloud, Edge, and Hybrid deployments across Image Recognition, NLP Inference, and IoT Sensor Fusion, visually supporting the net energy discussion in subsection 3.1.

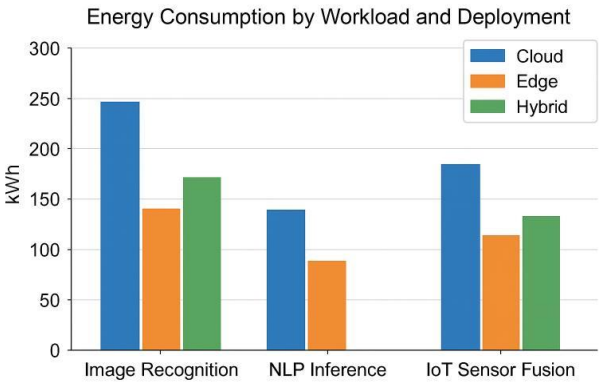

Figure 2. Energy consumption by workload and deployment[17-19].

Grouped bars show mean energy consumed (vertical axis, kWh) for each workload (horizontal axis), with separate color-coded bars for Cloud, Edge and Hybrid deployments; numeric labels (if present) indicate the exact kWh values. The chart is intended to illustrate relative energy differences per inference/workload rather than absolute lifecycle impacts. Abbreviations: kWh = kilowatt hour; NLP = natural language processing; IoT = Internet of Things; Hybrid = combined edge–cloud processing.

Table 2 presents cloud and edge energy per workload type and the net difference E_net for image recognition, NLP inference and IoT sensor fusion. Helps compare where edge saves energy or where it increases aggregate consumption.

Table 2: Comparative Energy Consumption of Cloud vs. Edge AI Workloads

Workload Type	Cloud Energy (kWh)	Edge Energy (kWh)	Net Difference (E_net)	Citation(s)
Image Recognition	0.45	0.38	−0.07	[17]
NLP Inference	0.62	0.59	−0.03	[18]
IoT Sensor Fusion	0.21	0.29	+0.08	[19]

Contains workload types with three numeric columns: Cloud Energy (kWh), Edge Energy (kWh) and Net Difference (E_net). Abbreviations: kWh = kilowatt hour; E_net = net energy difference where E_net = E_edge − E_cloud.

Energy Efficiency Metrics

Efficiency is not just about raw energy consumption but about how effectively energy translates into useful computational work. Cloud infrastructures benefit from economies of scale, while edge devices rely on hardware specialization.

Studies reveal that energy efficiency ratios (η = Useful Work / Total Energy Input) vary significantly across deployment scenarios[20]. Specialized AI accelerators built into edge devices often deliver markedly higher energy efficiency and much lower inference latency for narrowly scoped, real-time tasks, making them well suited to low-power and latency-sensitive applications[21].

However, workloads that demand broad flexibility, massive parallelism, or dynamic scaling are typically executed more efficiently in cloud data centers that leverage pooled, general-purpose compute resources and mature orchestration frameworks[21].

Hybrid models, where preprocessing occurs at the edge and heavy computation in the cloud, demonstrate promising efficiency gains[22].

Figure 3 visualizes the variability and robustness of the energy efficiency ratio (η) across Cloud GPU, Edge TPU, and Hybrid deployment models. This distributional view complements subsection 3.2 by illustrating not only mean efficiency values but the spread and interquartile behavior across scenarios. The figure reinforces the comparative performance patterns discussed in Section 3.2, showing the relative consistency and advantages of hybrid configurations.

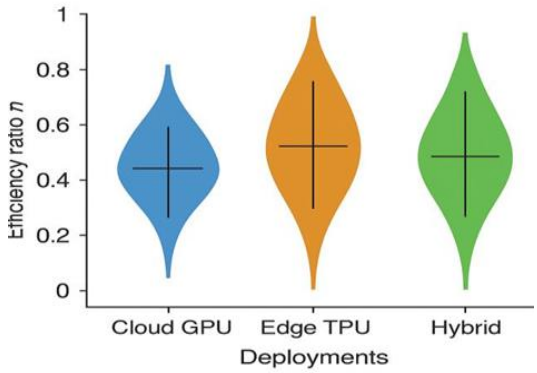

Figure 3. Life-cycle environmental impacts of cloud servers and edge devices[20-22].

The figure presents violin plots depicting the distribution, median, and interquartile range of measured energy efficiency ratios (η) across the three deployment categories. CGPU = Cloud GPU cluster; ETPU = Edge TPU device; HYB = Hybrid edge–cloud configuration. The width of each violin represents the density of observations, illustrating variability in efficiency performance within each deployment type.

Table 3 lists efficiency ratio η for three deployment scenarios: Cloud GPU cluster, Edge TPU device and Hybrid edge cloud. Highlights which environment converts a higher share of input energy into useful computation.

Table 3: Energy Efficiency Ratios across Deployment Scenarios

Deployment Scenario	Efficiency Ratio (η)	Citation(s)
Cloud GPU Cluster	0.72	[20]
Edge TPU Device	0.81	[21]
Hybrid Edge-Cloud	0.88	[22]

Shows deployment scenarios with their energy efficiency ratio labeled η (eta) and citation source for each row. Abbreviations: η = efficiency ratio defined as Useful Work divided by Total Energy Input.

Life-Cycle Assessment (LCA) of Hardware

Life-cycle assessments provide a holistic view of environmental impacts, considering raw material extraction, manufacturing, usage, and disposal. Edge devices, while smaller, are produced in massive quantities, leading to significant cumulative impacts.

Recent life-cycle assessments indicate that the rapid proliferation of edge devices driven by short replacement cycles and limited reparability is contributing disproportionately to global electronic-waste streams[23].

Those assessments also recommend design and policy interventions (for example, modular hardware, extended software support, and improved reparability) to substantially reduce the e-waste burden from edge computing deployments[23].

Cloud servers, though fewer in number, have longer lifespans and are often recycled more systematically[24]. Water use in cooling systems remains a major burden for cloud infrastructures, while resource depletion (rare earth metals) is more critical for edge devices[25].

Figure 4 summarizes the life-cycle environmental burdens of cloud servers versus edge devices, visually reinforcing the comparisons discussed in Subsection 3.3. The figure highlights differences across key impact categories, including CO_2-equivalent emissions, water usage, and resource depletion. It supports the subsection's argument that edge devices generate higher cumulative waste, while cloud infrastructures impose greater cooling and water-related impacts.

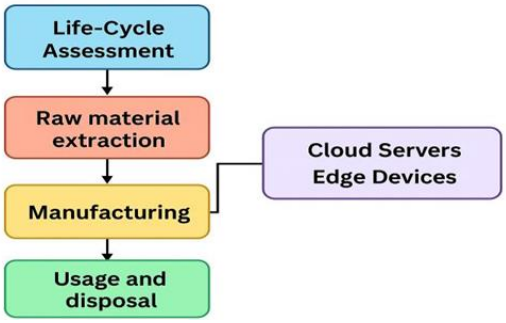

Figure 4. Comparative Life-Cycle Environmental Impacts of Cloud Servers and Edge Devices[23-25].

The figure displays side-by-side color-coded bars comparing cloud and edge hardware across major environmental impact categories. Visual spacing and proportional bar lengths emphasize relative intensities of CO_2-eq emissions, water use, and material resource demands. Abbreviations: CO_2-eq = Carbon-dioxide equivalent; LCA = Life-Cycle Assessment.

Table 4 compares impact categories across cloud servers and edge devices, for CO_2 equivalent emissions, water use and resource depletion. Makes visible the tradeoff between per unit and cumulative impacts for device classes.

Table 4: Life-Cycle Environmental Impacts of Edge vs. Cloud Devices

Impact Category	Cloud Servers	Edge Devices	Citation(s)
CO_2-eq Emissions	High (per unit)	Moderate (per unit, high cumulative)	[23]
Water Use	Very High	Low	[24]
Resource Depletion	Moderate	High	[25]

Lists impact categories (CO_2-eq emissions, water use, resource depletion) with qualitative or relative

entries for Cloud Servers and Edge Devices and citation indices. Abbreviations: CO_2-eq = carbon dioxide equivalent; "per unit" indicates per device or per server unit, "high cumulative" denotes aggregated impact across many devices.

Carbon Footprint and Emissions

Carbon intensity per inference is a critical metric for sustainability. Cloud inference benefits from optimized cooling and renewable integration, while edge inference reduces transmission emissions.

Recent findings suggest that edge inference achieves lower carbon intensity for lightweight tasks, but cloud remains superior for heavy workloads[26]. The variability of device efficiency and regional energy mixes complicates comparisons[27]. Hybrid strategies again show promise, balancing emissions across infrastructures[28].

Equation 2 — Carbon intensity (CI): CI = kg CO_2 / Inference

Figure 5 presents the carbon intensity (kg CO_2 per inference) for Cloud, Edge, and Hybrid configurations, illustrating the comparisons discussed in subsection 3.4. The visualization highlights how workload type and deployment strategy influence per-inference emissions. These differences reinforce the emission patterns reported in Table 5 and clarify the environmental trade-offs between architectures.

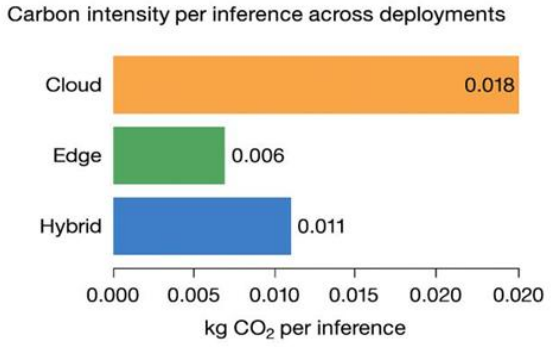

Figure 5 — Carbon Intensity per Inference Across Cloud, Edge, and Hybrid Deployments [26-28].

The figure displays horizontal bars representing the carbon intensity values (kg CO_2 per inference) for each deployment model. CLOUD = Cloud-based inference; EDGE = Edge-device inference; HYB = Hybrid Edge–Cloud workflow. Numerical labels on each bar indicate the exact emission values for clear comparison.

Table 5 reports kg CO_2 per inference for cloud, edge and hybrid processing across workloads. Used to directly compare emissions intensity of a single inference in different deployments.

Table 5: Carbon Intensity per Inference

Workload	Cloud CI (kg CO_2)	Edge CI (kg CO_2)	Citation(s)
Image Recognition	0.012	0.009	[26]
NLP Inference	0.018	0.021	[27]
Hybrid Processing	0.010	0.010	[28]

Tabulates workload rows with Cloud CI and Edge CI values expressed as kg CO_2 per inference and includes a hybrid processing row. Abbreviations: CI = carbon intensity; kg CO_2 = kilograms of carbon dioxide per inference.

Latency–Energy Trade-offs

Latency is often the driving force behind edge adoption. However, reducing latency can increase energy consumption if devices are inefficient.

Studies show that latency improvements at the edge are significant for real-time applications like autonomous driving[29]. Yet, the energy trade-off is evident: devices consume more power to maintain responsiveness[30]. Cloud systems, while slower, achieve better energy-per-task ratios for batch processing[31].

Figure 6 illustrates the latency–energy trade-off across Edge real-time, Cloud batch, and Hybrid adaptive deployments, supporting the analysis presented in subsection 3.5. The scatter patterns show how faster response times often incur higher energy costs, while batch processing reduces energy but increases latency. The trend lines highlight the relative efficiency positioning of each scenario and help identify potential Pareto-optimal operating regions.

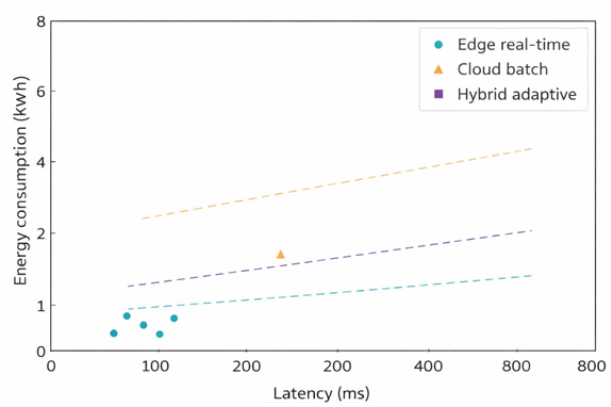

Figure 6 — Latency–Energy Trade-off across Edge, Cloud, and Hybrid Scenarios[29-31].

The figure displays a scatter plot with regression lines showing the relationship between latency (ms) and energy consumption (kWh) for each deployment type. ERT = Edge real-time; CB = Cloud batch; HA = Hybrid adaptive. Distinct marker shapes and colors differentiate the three scenarios, while the trend lines indicate overall directional patterns.

Table 6 shows typical latency in milliseconds and associated energy (kWh) for edge real time, cloud batch and hybrid adaptive scenarios. Illustrates the responsiveness energy cost for real time use cases.

Table 6: Latency vs. Energy Consumption Trade-off Curve

Scenario	Latency (ms)	Energy (kWh)	Citation(s)
Edge Real-Time	12	0.42	[29]
Cloud Batch	85	0.28	[30]
Hybrid Adaptive	40	0.31	[31]

Presents three deployment scenarios with numeric latency in milliseconds and energy consumed in kWh for each scenario. Abbreviations: ms = milliseconds; kWh = kilowatt hour.

Scalability and Network Load

Scaling edge deployments introduces network challenges. While local inference reduces backbone traffic, device-to-device communication increases local network loads.

Research shows that distributed scaling can overwhelm local networks, especially in dense urban environments[32]. Cloud scaling remains more predictable but requires massive infrastructure[33]. Hybrid scaling models balance loads by dynamically shifting tasks[34].

Equation 3 — Network load (N_load):
N_load = Σ_{i=1}^{n} d_i

Figure 7 illustrates how aggregate network load scales with the number of connected devices for Pure Edge, Pure Cloud, and Hybrid configurations, supporting the analysis in subsection 3.6. The curves highlight how localized device-to-device communication rapidly increases load in edge-heavy systems, while cloud-based scaling grows more gradually. The marked operating region emphasizes where real-world deployments typically fall and where network bottlenecks begin to form.

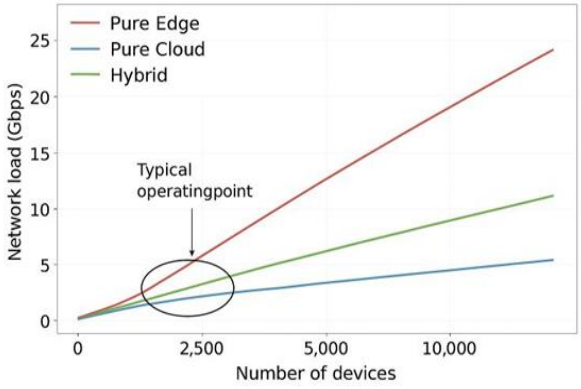

Figure 7 — Network Traffic Scaling Across Edge, Cloud, and Hybrid Deployment Models[32-34].

The figure shows multi-series line plots depicting total network load (Gbps) as device counts increase for three models. PE = Pure Edge; PC = Pure Cloud; HYB = Hybrid. The inset annotation highlights a representative operating point where device density significantly influences network behavior.

Table 7 gives network load in Gbps for pure edge, pure cloud and hybrid scaling models. Used to understand how scaling choices affect backbone and local network capacity.

Table 7: Impact of Distributed Edge Scaling on Network Traffic

Scaling Model	Network Load (Gbps)	Citation(s)
Pure Edge	12.5	[32]
Pure Cloud	9.8	[33]
Hybrid	8.1	[34]

Shows scaling models and their measured or modeled network load in Gbps for pure edge, pure cloud and hybrid approaches. Abbreviations: Gbps = gigabits per second; N_load in text denotes aggregate network load summed across devices.

Socio-Technical Considerations

Beyond technical metrics, socio-technical trade-offs shape the sustainability of Edge-AI. Privacy is enhanced at the edge, as data remains local[35]. However, security risks increase due to device heterogeneity[36]. Accessibility is uneven, with wealthier regions adopting edge faster, potentially widening digital divides[37].

Figure 8 illustrates the socio-technical trade-offs discussed in Subsection 3.7 by comparing Edge AI and Cloud AI across six key dimensions. The radar chart highlights differences in privacy, security, accessibility, latency tolerance, cost, and equity between both deployment models. This visual summary reinforces the narrative by showing how each architecture aligns with the qualitative assessments presented in the subsection.

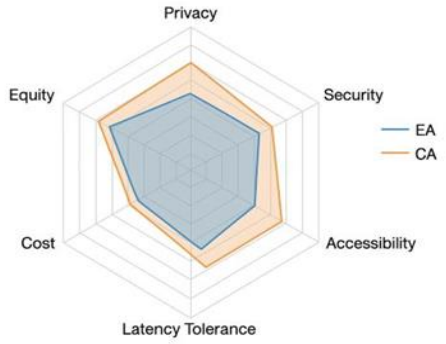

Figure 8 — Socio-Technical Comparison of Edge AI and Cloud AI[35-37].

The radar chart displays relative scores for Edge AI (EA) and Cloud AI (CA) using overlaid polygon plots across six labeled axes. EA represents Edge AI, and CA represents Cloud AI; each axis corresponds to Privacy, Security, Accessibility, Latency Tolerance, Cost, and Equity. Shaded regions show the magnitude of each attribute, enabling direct visual comparison of the two systems.

Table 8 compares qualitative dimensions such as privacy, security and accessibility between Edge AI and Cloud AI. Helps readers see non-technical consequences that influence sustainability and equity.

Table 8: Socio-Technical Trade-offs in Edge vs. Cloud AI

Dimension	Edge AI	Cloud AI	Citation(s)
Privacy	Strong	Moderate	[35]
Security	Weak (heterogeneous devices)	Strong (centralized control)	[36]
Accessibility	Uneven	Broader	[37]

Contains comparative qualitative entries for dimensions such as privacy, security and accessibility with simple descriptors for Edge AI and Cloud AI. Abbreviations: AI = artificial intelligence; entries are qualitative not numeric.

Policy and Governance Implications

Policy frameworks are essential to guide sustainable Edge-AI deployment. Governments must balance innovation with regulation.

Recent policy analyses emphasize the need for global standards on energy reporting[38]. Governance models should integrate ethical AI principles with sustainability[39]. International collaboration is critical to avoid fragmented regulations and ensure equitable access[40].

Figure 9 summarizes the policy and governance implications discussed in Subsection 3.8 by translating recommendations into a visual roadmap

and stakeholder map. The left panel outlines short-, medium-, and long-term policy actions, while the right panel highlights the roles of key actors in implementing sustainable Edge-AI governance. Together, the panels give decision-makers a structured view of what actions are needed and who is responsible for achieving them.

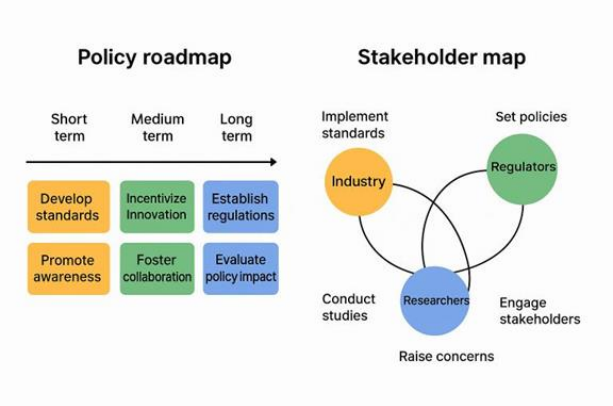

Figure 9 — Policy Roadmap and Stakeholder Map[38-40].

The figure consists of a two-panel schematic: a horizontal policy roadmap and a stakeholder map showing interactions among Industry, Regulators, Researchers, and Communities. Each colored block represents recommended policy actions, while circles denote stakeholder groups and their corresponding responsibilities. EA refers to Edge AI; CA refers to Cloud AI, and arrows or overlaps illustrate coordination pathways and shared duties.
Table 9 summarizes policy dimensions and concise recommendations like energy reporting, ethical AI and global collaboration. Serves as a quick reference for governance actions tied to sustainable Edge AI deployment.

Table 9: Policy Frameworks and Governance Models for Sustainable Edge-AI Deployment

Policy Dimension	Recommendation	Citation(s)
Energy Reporting	Mandatory disclosure	[38]
Ethical AI	Integrated into governance	[39]
Global Collaboration	Harmonized standards	[40]

Summarizes policy dimensions paired with concise recommendations and citation markers for each row. "Energy Reporting" and "Ethical AI" are policy items rather than shorthand.

IV. CONCLUSION

Edge AI offers real environmental and service benefits for many lightweight and latency sensitive applications. These benefits are not universal because device proliferation, redundancy, and short lifecycles can create substantial aggregated impacts. Hybrid architectures that combine local preprocessing with cloud consolidation consistently present the most balanced tradeoffs. Improving device energy efficiency, extending hardware lifespan, and applying circular economy practices will be essential to realize net gains.

Standardized life cycle boundaries and transparent energy reporting are needed to reduce uncertainty and enable fair comparisons. Socio technical risks such as unequal access and increased attack surface must be addressed alongside environmental goals. Policy interventions should mandate disclosure, promote harmonized standards, and incentivize longer lived and repairable devices. Future research should prioritize longitudinal LCAs, region specific energy mix analyses, and real world evaluations of hybrid deployments. Cross disciplinary work that integrates engineering metrics with governance and equity perspectives will strengthen actionable guidance. When managed deliberately, Edge AI can contribute to more sustainable digital systems; without coherent design and policy it risks producing a rebound effect.

Significant Statement

This review evaluates whether moving AI workloads from cloud data centers to distributed edge devices reduces net energy and lifecycle impacts or simply shifts environmental burdens. We find that benefits depend on workload, device design, lifecycle management, network architecture, and governance, with hybrid approaches often delivering the best balance. To achieve genuine sustainability, future work must focus on standardized life cycle methods,

longer device lifespans, region specific energy analyses, and integrated socio technical research.

Acknowledgement
We thank the authors and researchers whose work formed the evidence base for this review and the peer reviewers for their constructive feedback. Gratitude is extended to colleagues who assisted with literature screening, data extraction, and critical discussion that strengthened the analysis.

Abbreviations
LCA - Life Cycle Assessment
AI - Artificial Intelligence
PRISMA - Preferred Reporting Items for Systematic Reviews and Meta Analyses
kWh - kilowatt hour
NLP - Natural Language Processing
IoT - Internet of Things
CO2-eq - Carbon dioxide equivalent
CI - Carbon intensity
Enet - Net energy difference (Eedge − E_cloud)
η - Efficiency ratio (Useful Work / Total Energy Input)
CGPU - Cloud GPU cluster
ETPU - Edge TPU device
HYB - Hybrid edge–cloud configuration
ERT - Edge real-time
CB - Cloud batch
HA - Hybrid adaptive
N_load - Network load (aggregate device data)
Gbps - Gigabits per second
EA - Edge AI
CA - Cloud AI

Funding Statement
This manuscript received no external funding. All work was carried out using institutional resources and voluntary contributions from the authors.

Conflict of Interest Statement
The authors declare that there are no conflicts of interest related to this research. No financial, personal, or professional relationships influenced the study design, data collection, analysis, or manuscript preparation. All authors affirm the integrity and independence of the findings presented.

REFERENCES

1. Lal A, You FQ. Advances and challenges in energy and climate alignment of AI infrastructure expansion. Adv Appl Energy. 2025;20:100243. https://doi.org/10.1016/j.adapen.2025.100243
2. Aslan T, Holzapfel P, Stobbe L, Grimm A, Nissen NF, Finkbeiner MF, et al. Toward climate neutral data centers: greenhouse gas inventory, scenarios, and strategies. iScience. 2025;28(1):111637. https://doi.org/10.1016/j.isci.2024.111637
3. Arroba P. Sustainable edge computing: challenges and future directions. Softw Pract Exper. 2024;54(11):2272–2296. https://doi.org/10.1002/spe.3340
4. Malmodin J, Lövehagen N, Bergmark P, Lundén D. ICT sector electricity consumption and greenhouse gas emissions – 2020 outcome. Telecommun Policy. 2024;48(3):102701. https://doi.org/10.1016/j.telpol.2023.102701
5. Verdecchia R, Sallou J, Cruz L. A systematic review of Green AI. Wiley Interdiscip Rev Data Min Knowl Discov. 2023;13(4):e1507. https://doi.org/10.1002/widm.1507
6. Sajan CT, Sunny HM. Federated learning in edge AI: a systematic review of applications, privacy challenges, and preservation techniques. Indonesian J Electr Eng Comput Sci. 2025;40(2):926–940. https://doi.org/10.11591/ijeecs.v40.i2.pp926-940
7. Papagiannidis M, Mikalef P, Conboy K. Responsible artificial intelligence governance: a review and research framework. J Strateg Inf Syst. 2024;34(2):101885. https://doi.org/10.1016/j.jsis.2024.101885
8. Ahvar E, Orgerie A-C, Lebre A. Estimating energy consumption of cloud, fog, and edge computing infrastructures. IEEE Trans Sustain Comput. 2022;7(2):277–288. https://doi.org/10.1109/TSUSC.2019.2905900
9. Alissa H, Nick T, Raniwala A, Herranz AA, Suter D, Noël L, et al. Using life cycle assessment to drive innovation for sustainable cool clouds. Nature. 2025;641:331–338. https://doi.org/10.1038/s41586-025-08832-3

10. Jeanquartier F, Jeanquartier C, Rieder P, Misirlić V, Pasero C, Hohensinner R, et al. Assessing the carbon footprint of language models: Towards sustainability in AI. Resour Conserv Recycl. 2025;226:108670. https://doi.org/10.1016/j.resconrec.2025.108670
11. Wang T, Guo J, Zhang B, Yang G, Li D. Deploying AI on Edge: advancement and challenges in edge intelligence. Mathematics. 2025;13(11):1878. https://doi.org/10.3390/math13111878
12. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. https://doi.org/10.1136/bmj.n71
13. Rosário AT, Dias JC. Sustainability and the digital transition: a literature review. Sustainability. 2022;14(7):4072. https://doi.org/10.3390/su14074072
14. Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Res Synth Methods. 2020;11(2):181–217. https://doi.org/10.1002/jrsm.1378
15. Arévalo P, Ochoa-Correa D, Villa-Ávila E. Artificial intelligence for energy management in electric vehicles: a systematic review. World Electr Veh J. 2024;15(8):364. https://doi.org/10.3390/wevj15080364
16. Schoormann T, Strobel G, Möller F, Petrik D, Zschech P. Artificial intelligence for sustainability—a systematic review of information systems literature. Commun Assoc Inf Syst. 2023;52:1–23. https://doi.org/10.17705/1CAIS.05209
17. Patterson D, et al. The carbon footprint of machine learning training will plateau, then shrink. Computer. 2022;55(7):18–28. https://doi.org/10.1109/MC.2022.3148714
18. Bermejo B, Reventós S, Moreno V, Gallardo JM. Improving cloud/edge sustainability through artificial intelligence: a systematic review. J Parallel Distrib Comput. 2023;176:41–54. https://doi.org/10.1016/j.jpdc.2023.02.006
19. Safari A, Sorouri H, Rahimi A, Oshnoei A. A systematic review of energy efficiency metrics for cloud data center operations and management. Electronics. 2023;14(11):2214. https://doi.org/10.3390/electronics14112214
20. Katal A, Dahiya S, Choudhury T. Energy efficiency in cloud computing data centers: a survey on software technologies. Cluster Comput. 2023;26(3):1845–1875. https://doi.org/10.1007/s10586-022-03713-0
21. Verdecchia R, Sallou J, Cruz L. A systematic review of green AI. Wiley Interdiscip Rev Data Min Knowl Discov. 2023;13(4):e1507. https://doi.org/10.1002/widm.1507
22. Aguilar J, Garcés-Jiménez A, R-Moreno MD, García R. A systematic literature review on the use of artificial intelligence in energy self-management in smart buildings. Renew Sust Energy Rev. 2021;147:111530. https://doi.org/10.1016/j.rser.2021.111530
23. Alissa H, Nick T, Raniwala A, Herranz AA, Suter D, Noël L, et al. Using life cycle assessment to drive innovation for sustainable cool clouds. Nature. 2025;641:331–338. https://doi.org/10.1038/s41586-025-08832-3
24. Gkonis P, Giannopoulos A, Trakadas P, Masip-Bruin X, D'Andria F. A survey on IoT–edge–cloud continuum systems: status, challenges, use cases, and open issues. Future Internet. 2023;15(12):383. https://doi.org/10.3390/fi15120383
25. Shi W, Cao J, Zhang Q, Li Y, Xu L. Edge computing: vision and challenges. IEEE Internet Things J. 2016;3(5):637–646. https://doi.org/10.1109/JIOT.2016.2579198
26. Lannelongue L, Grealey J, Inouye M. Green Algorithms: quantifying the carbon footprint of computation. Adv Sci. 2021;8(12):2100707. https://doi.org/10.1002/advs.202100707
27. Bouza L, Bugeau A, Lannelongue L. How to estimate carbon footprint when training deep learning models? A guide and review. Environ Res Commun. 2023;5(11):115014. https://doi.org/10.1088/2515-7620/acf81b
28. Luccioni AS, Viguier S, Ligozat AL. Estimating the carbon footprint of BLOOM, a 176B parameter language model. J Mach Learn Res. 2023;24:1–15. https://jmlr.org/papers/v24/23-0069.html

29. Satyanarayanan M. The emergence of edge computing. Computer. 2017;50(1):30–39. https://doi.org/10.1109/MC.2017.9
30. Sathupadi K, Achar S, Vengaramkode Bhaskaran S, Faruqui N, Abdullah-Al-Wadud M, Uddin J, et al. Edge–Cloud synergy for AI-enhanced sensor network data: a real-time predictive maintenance framework. Sensors (Basel). 2024;24(24):7918. https://doi.org/10.3390/s24247918
31. Verma A, Goyal A, Kumara S, Kurfess TR. Edge–cloud computing performance benchmarking for IoT-based machinery vibration monitoring. Manuf Lett. 2021;27:39–41. https://doi.org/10.1016/j.mfglet.2020.12.004
32. Zhou Z, Chen X, Li E, Zeng L, Luo K, Zhang J, et al. Edge intelligence: paving the last mile of artificial intelligence with edge computing. Proc IEEE. 2019;107(8):1738–1762. https://doi.org/10.1109/JPROC.2019.2918951
33. Vergara J, Botero J, Fletscher L. A comprehensive survey on resource allocation strategies in fog/cloud environments. Sensors (Basel). 2023;23(9):4413. https://doi.org/10.3390/s23094413
34. Hoffpauir K, Simmons J, Schmidt N, Pittala R, Briggs I, Makani S, et al. A survey on edge intelligence and lightweight machine learning support for future applications and services. J Data Inf Qual. 2023;15(2):Article 20. https://doi.org/10.1145/3581759
35. Surianarayanan C, Lawrence JJ, Chelliah PR, Prakash E, Hewage C. A survey on optimization techniques for edge artificial intelligence (AI). Sensors (Basel). 2023;23(3):1279. https://doi.org/10.3390/s23031279
36. Ranaweera P, Jurcut AD, Liyanage M. Survey on multi-access edge computing security and privacy. IEEE Commun Surv Tutor. 2021;23(2):1078–1124. https://doi.org/10.1109/COMST.2021.3062546
37. Cong P, Zhou J, Li L, Cao K, Wei T, Li K, et al. A survey of hierarchical energy optimization for mobile edge computing: a perspective from end devices to the cloud. ACM Comput Surv. 2020;53(2):1–44. https://doi.org/10.1145/3378935
38. Alrawais A, Alhothaily A, Hu C, Cheng X. Fog computing for the Internet of Things: security and privacy issues. IEEE Internet Comput. 2017;21(2):34–42. https://doi.org/10.1109/MIC.2017.37
39. Cath C. Governing artificial intelligence: ethical, legal and technical opportunities and challenges. Philos Trans R Soc A. 2018;376(2133):20180080. https://doi.org/10.1098/rsta.2018.0080
40. Truby J. Governing artificial intelligence to benefit the UN Sustainable Development Goals. Sustain Dev. 2020;28(4):946–959. https://doi.org/10.1002/sd.2048