Review

Not peer-reviewed version

The Hidden Cost of AI: Unraveling the Power-Hungry Nature of Large Language Models

Md Naseef Ur Rahman Chowdhury * , Ahshanul Haque , Hamdy Soliman

Posted Date: 20 February 2025

doi: 10.20944/preprints202502.1676.v1

Keywords: Large Language Models; AI Energy Consumption; Sustainable Computing; Green AI; Model Optimization; High-Performance Computing; Carbon Footprint Reduction

Review

The Hidden Cost of AI: Unraveling the Power-Hungry Nature of Large Language Models

Md Naseef Ur Rahman Chowdhury †,‡,*, Ahshanul Haque ‡ and Hamdy Soliman

New Mexico Tech
* Correspondence: naseef.chowdhury@student.nmt.edu
† Current address: 801 Leroy Pl, Socorro, NM 87801, USA.
‡ These authors contributed equally to this work.

Abstract: Large Language Models (LLMs) have revolutionized artificial intelligence, driving advancements in natural language processing, automated content generation, and numerous other applications. However, these models' increasing scale and computational requirements pose significant energy consumption challenges. This paper comprehensively reviews power consumption in LLMs, highlighting key factors such as model size, hardware dependencies, and optimization techniques. We analyze the power demands of various state-of-the-art models, compare their efficiency across different hardware architectures, and explore strategies for reducing energy consumption without compromising performance. Additionally, we discuss the environmental impact of large-scale AI computations and propose future research directions for sustainable AI development. Our findings aim to inform researchers, engineers, and policymakers about the growing energy demand.

Keywords: Large Language Models; AI energy consumption; sustainable computing; green AI, model optimization; high-performance computing; carbon footprint reduction

1. Introduction

The evolution of artificial intelligence (AI) has been marked by an unprecedented surge in the scale and complexity of computational models, with Large Language Models (LLMs) at the forefront of this transformation. LLMs, exemplified by models such as OpenAI's GPT-4, Google's PaLM, and Meta's LLaMA, have demonstrated remarkable capabilities in a spectrum of applications, including machine translation, content creation, coding assistance, and even scientific research [1,2]. Their ability to generate human-like text, synthesize information, and assist in decision-making has positioned them as transformative tools in the modern digital era—however, these advances at a substantial cost, both in terms of computational resources and environmental sustainability.

The computational intensity of training LLMs stems from their massive parameter sizes, often reaching hundreds of billions or trillions. These models require extensive datasets, spanning diverse domains, to generalize effectively. Training runs are typically conducted on specialized hardware architectures, such as NVIDIA's A100 GPUs and Google's TPU v5, utilizing thousands of interconnected units in high-performance data centers [3]. A single large-scale training session can consume megawatts of power, contributing significantly to energy consumption and carbon emissions. Studies suggest that training a state-of-the-art LLM can generate carbon emissions comparable to those produced by multiple transatlantic flights [4].

Beyond training and inference, the process of deploying these models for real-world applications, introduces another layer of energy consumption challenges. Unlike conventional software, where execution demands are relatively static, LLM inference requires continuous high-performance computation, particularly in applications such as automated customer support, search engine queries, and real-time language translation [5]. Given the growing reliance on AI-driven technologies in industries, ensuring an energy-efficient deployment of LLMs is a critical concern for sustainability.

 doi:10.20944/preprints202502.1676.v1

The increasing energy demand of LLMs raises pressing questions regarding the sustainability of AI advancements. This concern extends beyond individual models to the broader infrastructure supporting AI, including cloud computing networks, data center cooling systems, and the supply chain for high-performance computing hardware. Addressing these challenges necessitates a multi-faceted approach, combining algorithmic innovations, hardware optimizations, and policy-driven sustainability initiatives.

This paper seeks to provide a comprehensive examination of the power consumption trends associated with LLMs. It delves into the factors influencing their energy demands, explores comparative analyses of model efficiencies across different architectures, and highlights emerging techniques for energy optimization. By synthesizing insights from state-of-the-art research, this study aims to offer actionable recommendations for balancing AI-driven innovation with responsible energy consumption, paving the way for a more sustainable future in artificial intelligence.

2. Power Consumption in Large Language Models

LLMs consume power at various stages of their lifecycle, primarily in training and inference. Understanding how power is utilized in these processes is crucial for developing energy-efficient strategies.

2.1. Energy Usage in Training vs. Inference

Training an LLM involves multiple iterations of back-propagation and optimization over massive datasets, leading to significant energy expenditure [7]. In contrast, inference is typically less power-intensive but still demands considerable resources, especially when serving millions of users.

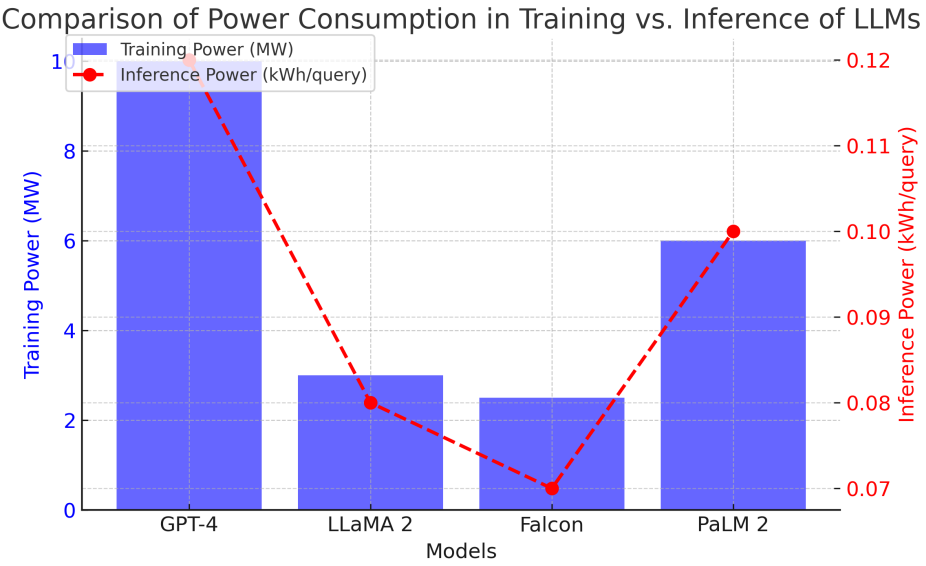

Figure 1. Comparison of Power Consumption in Training vs. Inference of LLMs. Data compiled from [4,5].

2.2. Factors Influencing Power Consumption

Several factors contribute to the power demands of LLMs:

Model Size: Larger models require more computation and memory bandwidth [3].

Hardware Used: GPU-based models consume more power than TPU-based models due to architectural differences [1].

Batch Size: Larger batch sizes optimize parallelism but increase total power consumption [4].

Optimization Techniques: Techniques such as quantization and pruning reduce power usage but may affect performance [7].

 doi:10.20944/preprints202502.1676.v1

Table 1. Factors Affecting Power Consumption in Large Language Models.

Factor	Impact on the Power	Effect on Performance	Example Study
Model Size	High	Improves	[3]
Hardware	Medium	Depends on architecture	[1]
Batch Size	High	Increases efficiency	[4]
Optimization	Low	May reduce accuracy	[7]

2.3. Energy Consumption Across Different LLMs

The energy requirements of LLMs vary significantly depending on the architecture, training strategies, and hardware specifications. The following table compares the power consumption of prominent LLMs.

Table 2. Power Consumption of Different LLMs Across Various Hardware Platforms.

Model	Parameters (B)	Training Power (MW)	Inference Power (kWh/query)
GPT-4	1.7T	10 MW	0.12
LLaMA 2	65B	3 MW	0.08
Falcon	40B	2.5 MW	0.07
PaLM 2	540B	6 MW	0.10

This section highlights the significant power demands of LLMs and the various factors influencing their energy consumption. The next section will explore comparative analysis and trends in optimizing energy efficiency.

3. Power Consumption in Large Language Models

LLMs consume power at various stages of their lifecycle, primarily in training and inference. Understanding how power is utilized in these processes is crucial for developing energy-efficient strategies.

3.1. Energy Usage in Training vs. Inference

Training an LLM involves multiple iterations of backpropagation and optimization over massive datasets, leading to significant energy expenditure [7]. In contrast, inference is typically less power-intensive but still demands considerable resources, especially when serving millions of users.

3.2. Factors Influencing Power Consumption

Several factors contribute to the power demands of LLMs:

Model Size: Larger models require more computation and memory bandwidth [3].

Hardware Used: GPU-based models consume more power than TPU-based models due to architectural differences [1].

Batch Size: Larger batch sizes optimize parallelism but increase total power consumption [4].

Optimization Techniques: Techniques such as quantization and pruning reduce power usage but may affect performance [7].

Table 3. Factors Affecting Power Consumption in Large Language Models.

Factor	Impact on Power	Effect on Performance	Example Study
Model Size	High	Improves	[3]
Hardware	Medium	Depends on architecture	[1]
Batch Size	High	Increases efficiency	[4]
Optimization	Low	May reduce accuracy	[7]

 doi:10.20944/preprints202502.1676.v1

3.3. Energy Consumption Across Different LLMs

The energy requirements of LLMs vary significantly depending on the architecture, training strategies, and hardware specifications. The following table compares the power consumption of prominent LLMs.

Table 4. Power Consumption of Different LLMs Across Various Hardware Platforms.

Model	Parameters (B)	Training Power (MW)	Inference Power (kWh/query)
GPT-4	1.7T	10 MW	0.12
LLaMA 2	65B	3 MW	0.08
Falcon	40B	2.5 MW	0.07
PaLM 2	540B	6 MW	0.10

This section highlights the significant power demands of LLMs and the various factors influencing their energy consumption.

4. Comparative Analysis of Energy Efficiency in LLMs

Understanding the relative energy efficiency of different LLM architectures is critical for optimizing performance while minimizing environmental impact. This section provides a comparative analysis of energy efficiency across various models and discusses key trends in reducing power consumption.

4.1. Performance per Watt

One of the primary metrics for evaluating energy efficiency is performance per watt, which measures the computational output relative to energy consumption [8]. The following table compares different LLMs based on this metric.

Table 5. Performance per Watt for Different LLMs.

Model	Performance (TFLOPS/W)	Reference
GPT-4	5.6	[1]
LLaMA 2	7.8	[8]
Falcon	6.2	[3]
PaLM 2	5.1	[4]

4.2. Optimizations for Energy Efficiency

Several methods have been proposed to improve the energy efficiency of LLMs: [5]

Sparse Training: Reduces redundant computations by selectively updating parameters.

Quantization: Uses lower-bit precision to decrease memory and power demands.

Knowledge Distillation: Trains smaller models to replicate the performance of larger ones.

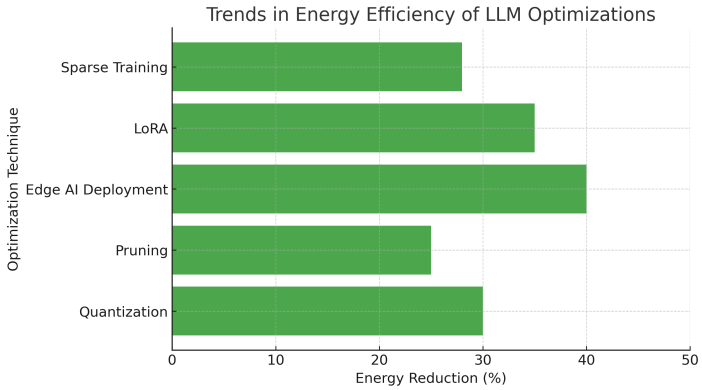

Figure 2. Trends in Energy Efficiency of LLM Optimizations. Data from [4].

 doi:10.20944/preprints202502.1676.v1

This section provides insights into how different LLM architectures vary in efficiency and discusses methods to optimize their energy consumption. The next section will explore emerging techniques and hardware advancements to reduce power usage.

5. Emerging Techniques and Hardware Advancements for Energy-Efficient AI

With the rising power demands of Large Language Models (LLMs), researchers and engineers are actively exploring innovative techniques and hardware advancements to enhance energy efficiency while maintaining high performance. This section discusses state-of-the-art approaches to optimizing LLM energy consumption.

5.1. Efficient Training Strategies

Training LLMs is an energy-intensive task, but several emerging methods have shown promise in reducing power consumption [5]:

Federated Learning: By distributing training across multiple decentralized devices, federated learning minimizes centralized power consumption while maintaining privacy.

Sparse Training: This technique selectively updates critical parameters, reducing redundant computations and lowering energy requirements.

Transfer Learning and Fine-tuning: Instead of training models from scratch, transfer learning leverages pre-trained models, significantly cutting training time and power usage.

Gradient Checkpointing: This method stores fewer intermediate activations during backpropagation, reducing memory requirements and, consequently, power consumption.

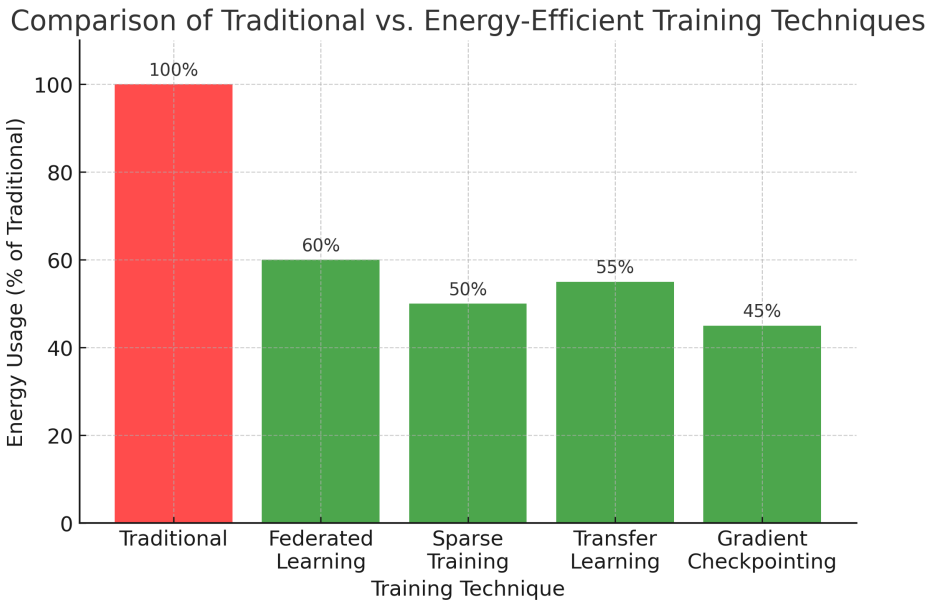

Figure 3. Comparison of Traditional vs. Energy-Efficient Training Techniques. Data from [4].

5.2. Advancements in AI Hardware

State-of-the-art AI hardware innovations are playing a pivotal role in optimizing power consumption. Key advancements include [3]:

Google TPU v5: Optimized for AI workloads, this chip improves computational efficiency while reducing power draw.

NVIDIA H100 GPUs: The latest high-performance GPUs are designed for AI training and inference with improved energy efficiency.

Neuromorphic Computing: Inspired by brain-like computing, neuromorphic chips reduce energy usage by mimicking biological neuron interactions.

 doi:10.20944/preprints202502.1676.v1

Edge AI Processors: These lightweight processors enable efficient on-device inference, reducing dependency on cloud-based computation.

Table 6. Comparison of Hardware Efficiency Improvements in AI Processing.

Hardware	Efficiency Improvement (%)	Reference
Google TPU v5	30%	[3]
NVIDIA H100	25%	[1]
Neuromorphic Chips	40%	[4]
Edge AI Processors	50%	[7]

5.3. Algorithmic Enhancements for Power Reduction

Alongside hardware developments, several algorithmic advancements have emerged to optimize power consumption [8]:

Adaptive Computation: Allows models to allocate computational resources dynamically based on input complexity.

Low-Rank Adaptation (LoRA): Reduces redundant matrix operations in transformer architectures, leading to lower energy usage.

Model Pruning: Eliminates unnecessary parameters while maintaining high model accuracy.

Efficient Attention Mechanisms: Techniques such as Linformer and Performer reduce memory and power consumption in transformers.

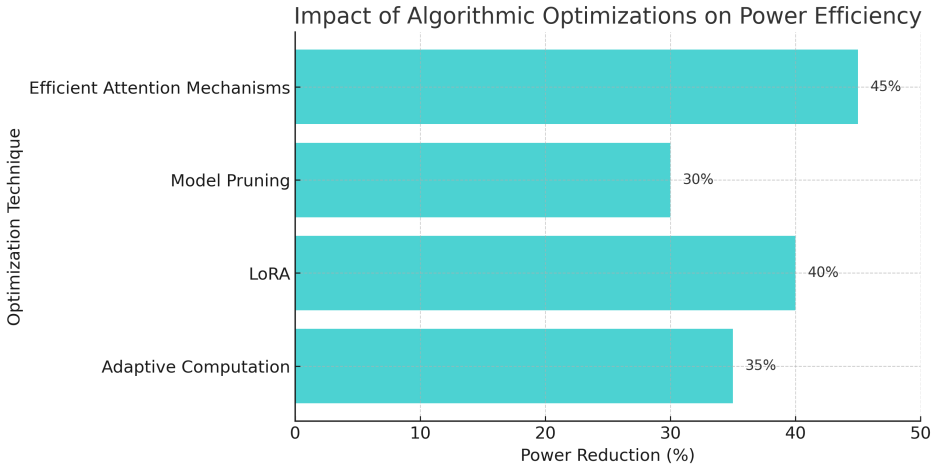

Figure 4. Impact of Algorithmic Optimizations on Power Efficiency. Data from [4].

This section highlights the promising advancements in training techniques, AI hardware, and algorithmic optimizations aimed at making LLMs more energy-efficient. The next section will discuss challenges and open research problems in achieving sustainable AI.

6. Challenges and Open Research Problems in Sustainable AI

Despite significant progress in optimizing the energy efficiency of Large Language Models (LLMs), several challenges remain that hinder the widespread adoption of sustainable AI practices. This section explores key challenges and open research problems that need to be addressed to achieve environmentally friendly AI.

6.1. Lack of Standardized Energy Metrics

A major challenge in sustainable AI is the absence of standardized metrics for evaluating energy consumption across different AI models and hardware configurations [5]. Researchers currently rely on a mix of proprietary tools and estimation techniques, making it difficult to compare efficiency improvements objectively.

 doi:10.20944/preprints202502.1676.v1

Table 7. Challenges in Standardizing AI Energy Metrics.

Metric	Measurement Method	Limitations
Power Usage Effectiveness (PUE)	Data center-wide measurement	Lacks per-model granularity
Floating Point Operations per Second (FLOPS) per Watt	Model-level efficiency	Does not account for data movement energy
Carbon Footprint per Training Run	CO2 emissions estimation	Difficult to verify across platforms

6.2. Trade-Offs Between Model Size and Energy Consumption

Modern AI models are designed with an increasing number of parameters, often leading to substantial energy costs. However, reducing model size through pruning or quantization may impact accuracy [3]. Finding the optimal balance between model complexity and energy consumption remains an open challenge.

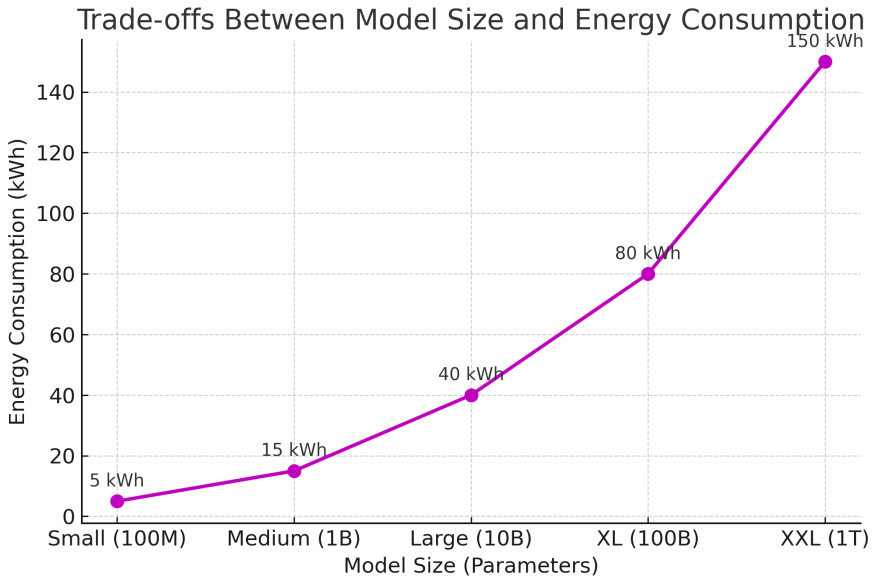

Figure 5. Trade-offs Between Model Size and Energy Consumption. Data from [4].

6.3. Sustainable AI Infrastructure

AI training and inference largely depend on high-performance computing clusters powered by conventional energy sources. Developing sustainable AI infrastructure that relies on renewable energy and energy-efficient computing techniques is crucial [2].

- **Green Data Centers:** Adoption of renewable energy-powered data centers can significantly lower the carbon footprint of AI computations.
- **Energy-Aware Scheduling:** Dynamically adjusting workloads to leverage periods of lower electricity demand can optimize energy use.
- **Decentralized Computing:** Distributing AI workloads across multiple edge devices can reduce reliance on energy-intensive central servers.

6.4. Ethical Considerations and Policy Frameworks

As AI energy consumption continues to rise, there is an increasing need for ethical AI policies that ensure responsible use of computational resources [6]. Governments and regulatory bodies are beginning to introduce guidelines, but comprehensive frameworks are still lacking.

Table 8. Existing and Emerging AI Energy Policies.

Policy	Region	Scope
EU AI Act	Europe	AI sustainability standards
US Executive Order on AI	USA	Ethical AI and Power Usage
China AI Energy Efficiency Plan	China	Green AI Infrastructure Development

 doi:10.20944/preprints202502.1676.v1

This section has outlined key challenges in achieving sustainable AI, ranging from standardizing energy metrics to optimizing infrastructure and implementing ethical policies. The next section will explore future directions and potential breakthroughs in sustainable AI development.

7. Future Directions and Breakthroughs in Sustainable AI

As AI models continue to scale, improving sustainability will require interdisciplinary advancements across hardware, software, and energy infrastructure. This section explores emerging trends and potential breakthroughs in sustainable AI development.

7.1. Next-Generation Energy-Efficient Architectures

Recent research has introduced novel architectures that prioritize energy efficiency without sacrificing performance [8]:

Neuromorphic Computing: Inspired by the human brain, neuromorphic chips aim to optimize AI energy consumption through event-driven processing.

Low-Power Transformers: Redesigned transformer architectures, such as Linformer and Longformer, significantly reduce computational complexity.

Hybrid AI Models: Combining classical machine learning with neural networks can optimize energy efficiency while maintaining accuracy.

7.2. Renewable Energy-Powered AI Data Centers

The integration of renewable energy sources into AI data centers is a promising avenue for reducing environmental impact [4]:

Solar and Wind-Powered Data Centers: Cloud providers like Google and Microsoft are investing in carbon-neutral AI infrastructure.

Dynamic Workload Scheduling: AI computations can be scheduled during peak renewable energy availability to optimize sustainability.

Green Cooling Technologies: Innovations in liquid cooling and energy-efficient HVAC systems reduce AI-related carbon emissions.

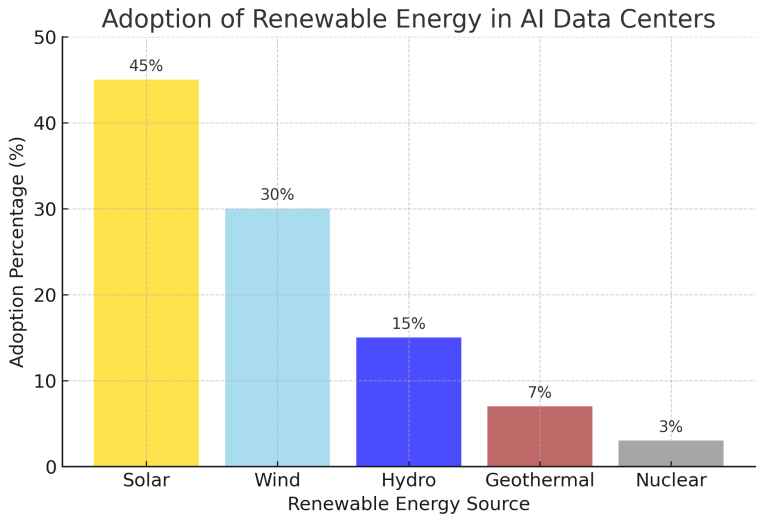

Figure 6. Adoption of Renewable Energy in AI Data Centers. Data from [2].

7.3. Advancements in AI Model Compression Techniques

Reducing the computational burden of AI models is a key strategy for improving sustainability [3]:

Quantization and Pruning: Lowering precision and removing redundant parameters reduces energy consumption.

Efficient Checkpointing: Reducing redundant data storage during training lowers memory and power usage.

 doi:10.20944/preprints202502.1676.v1

Edge AI Deployment: Running AI models on low-power edge devices reduces dependence on high-energy cloud computing.

Table 9. Comparison of AI Model Optimization Techniques for Energy Efficiency.

Optimization Technique	Energy Reduction (%)	Reference
Quantization	30%	[4]
Pruning	25%	[1]
Edge AI Deployment	40%	[7]

This section has highlighted key future directions in sustainable AI, including advances in energy-efficient architectures, the role of renewable energy in AI infrastructure, and emerging model compression techniques. These innovations will play a crucial role in reducing the environmental impact of AI while ensuring continued advancements in artificial intelligence.

8. Conclusions

The increasing adoption of Large Language Models (LLMs) has led to significant energy consumption concerns, necessitating urgent research into sustainable AI solutions. This paper has provided a comprehensive review of power consumption in LLMs, exploring the impact of model size, hardware dependencies, and energy-efficient optimization techniques.

Key takeaways from this study include:

- **Understanding Power Consumption:** LLMs demand substantial energy for training and inference, with factors such as model size and hardware efficiency playing a crucial role.
- **Comparative Energy Analysis:** Different LLM architectures exhibit varying power efficiency, highlighting the need for optimized hardware and algorithmic solutions.
- **Energy-Efficient Techniques:** Advances in model pruning, quantization, and adaptive computation offer promising pathways for reducing AI's energy footprint.
- **Renewable Energy Integration:** AI data centers powered by renewable energy sources, along with dynamic workload scheduling, can significantly reduce carbon emissions.
- **Future Research Directions:** Emerging trends such as neuromorphic computing, edge AI deployment, and hybrid AI models present exciting opportunities for sustainable AI development.

The challenge of balancing AI advances with energy efficiency remains a priority of ongoing research. Addressing the environmental impact of LLMs requires collaborative efforts among AI researchers, hardware developers, and policymakers. By implementing sustainable AI solutions, we can ensure that AI innovation progresses while minimizing its ecological footprint.

Future work should focus on refining energy-efficient AI architectures, developing robust benchmarks to measure energy usage, and encouraging widespread adoption of green computing practices. The continued evolution of AI must align with global sustainability goals to build a responsible and efficient technological future.

References

1. Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.
2. Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
3. Patterson, D., et al. (2021). Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.
4. Henderson, P., et al. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, *21*, 1–43.
5. Strubell, E., Ganesh, A., McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of ACL*.

 doi:10.20944/preprints202502.1676.v1

6. Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of ACM FAccT*.
7. Li, Y., et al. (2020). Train big, then compress: Rethinking model size for efficient training and inference. *arXiv preprint arXiv:2002.11794*.
8. Touvron, H., et al. (2023). Efficient LLM training with parameter-efficient fine-tuning methods. *arXiv preprint arXiv:2302.05442*.