International Journal of Technology and Applied Science (IJTAS)

E-ISSN: 2230-9004 ● Website: www.ijtas.com ● Email: editor@ijtas.com

Energy, Efficiency, and Sustainability in LLMs, RAG, and Agent Architectures

Yash Agrawal

yash.agr96@gmail.com

Abstract:

Artificial Intelligence now underpins consumer applications, enterprise systems, and national infrastructure through Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and AI Agents. Their rapid adoption, however, raises concerns over energy use, carbon emissions, and environmental impact. This review synthesizes scattered research on the sustainability challenges of these three paradigms and proposes a comparative framework that highlights both inefficiencies and opportunities for greener design. We examine (i) the compute and carbon costs of training and inference, (ii) RAG's potential as a lower-impact alternative to retraining, (iii) the energy overhead of agent orchestration, and (iv) emerging eco-efficiency benchmarks. We conclude with design patterns, policy directions, and future research priorities for aligning AI innovation with sustainable computing.

Keywords: Retrieval-Augmented Generation (RAG), sustainable AI, carbon footprint, green computing, eco-benchmarks, hybrid inference, energy-aware orchestration, hardware-software co-design, carbon-aware scheduling, federated RAG, responsible AI, sustainable architecture, AI lifecycle emissions.

1. INTRODUCTION

1.1 The Rise of Large-Scale AI

Over the past five years, Large Language Models (LLMs) have evolved from research curiosities into core infrastructure. Systems like GPT-4, Claude, and LLaMA now support search engines, productivity tools, customer service platforms, and coding assistants used by millions each day. Their scale has unlocked new opportunities for automation but also exposed the resource demands of this technology. Training a frontier model requires petaflop-years of computation and produces hundreds to thousands of tons of carbon emissions[1]. Once deployed, the greater challenge is inference, which has become the main source of long-term energy use as models respond to billions of queries.

1.2 Retrieval-Augmented Generation as a Partial Solution

To reduce the inefficiencies of retraining, many systems now use retrieval-augmented generation (RAG). Rather than repeatedly retraining models, RAG pipelines draw information from external databases and feed it into the model at runtime[2]. This grounds responses in evidence and lessens the need for constant retraining. However, RAG is not without cost. Creating embeddings for millions of documents, maintaining vector databases, and refreshing them as knowledge changes all require significant resources. Its sustainability depends on whether the savings from reduced retraining outweigh the ongoing costs of retrieval and storage, a balance that has not yet been fully studied.

1.3 The Emergence of AI Agents

The latest development is the rise of **AI agents**, systems that plan, act, and reason across multiple steps. Agents can decompose tasks, call APIs, coordinate with other agents, and reflect on their outputs. This flexibility enables new applications but also introduces higher energy demand. A task that once needed one model call may now trigger dozens of LLM queries, retrieval operations, and memory updates. Early

evidence shows naïve workflows can increase per-task energy use by an order of magnitude, underscoring the need for more efficient designs.

1.4 The Missing System-Level Perspective

Research on these three paradigms has developed largely in silos. Studies on LLMs emphasize scale and accuracy, work on RAG focuses on retrieval quality, and agent benchmarks measure task success. What is missing is a broader perspective on environmental impact. Do RAG pipelines actually reduce emissions once storage and update costs are considered? Can agents be organized in ways that cut redundancy rather than increase it [3]? And what metrics, beyond accuracy, should guide evaluation [5]? This review explores these questions by drawing on recent findings and case studies. We argue that sustainability should be treated as a core design principle for LLMs, RAG, and agents alike. By bringing together scattered evidence and highlighting unresolved challenges, our aim is to outline a path toward AI systems that are not only powerful and efficient but also environmentally responsible.

2. BACKGROUND & EXISTING WORK

2.1 Large Language Models (LLMs)

The conversation about sustainability in AI often begins with Large Language Models. Training advanced systems such as GPT-3 has been estimated to release more than 550 metric tons of CO_2, roughly equal to the yearly emissions of 120 cars. Models at the scale of GPT-4 almost certainly consume more, although no official figures have been disclosed.

Training, however, is only part of the story. Once deployed, LLMs process billions of queries each day. At this scale, inference becomes the main driver of lifetime emissions, with Google researchers estimating it accounts for more than 80 percent. This has shifted attention toward making inference more efficient through methods such as distillation, quantization, and hardware optimization [4]. Benchmarking studies show striking variation: the most optimized models can be tens of thousands of times more efficient than the least, demonstrating that greener AI depends as much on design choices as on raw model size.

2.2 Retrieval-Augmented Generation (RAG)

LLMs are powerful but static. Updating them requires costly retraining, which has led to the rise of retrieval-augmented generation (RAG). Rather than embedding new knowledge directly into the model, RAG systems query external databases at runtime. This grounds outputs in evidence and reduces the need for repeated retraining, but it also introduces new costs. Creating embeddings for millions of documents, refreshing them as knowledge evolves, and maintaining large vector databases all require significant resources.

When managed carefully, RAG can be more sustainable than retraining. A 2025 medical QA study found that a local RAG system using LLaMA-3.1 8B was both more accurate and less energy-intensive than domain-specific mini-LLMs, largely because embeddings were updated monthly rather than models retrained. The long-term sustainability of RAG depends on balancing the overhead of retrieval with the savings from reduced retraining.

2.3 AI Agents

The newest paradigm is the rise of AI agents, systems that can plan, act, and coordinate across multiple steps. Agents can break down tasks, call APIs, and even collaborate, enabling applications that range from research assistants to workflow managers. This flexibility, however, comes with a cost. A task that once required a single LLM call can expand into a dozen or more model queries, retrievals, and memory updates. Early evaluations suggest that naive workflows may increase per-task energy use by an order of magnitude.

At the same time, agents could also become part of the solution. With smarter orchestration, caching, and pruning, they have the potential to cut redundancy and optimize resource use [3]. The open question is whether agents can evolve into energy-aware decision makers rather than energy amplifiers, an area of research that is still in its early stages.

2.4 The Knowledge Gap

LLMs, RAG, and agents each have their own body of research, but most studies remain siloed, focused on model accuracy, retrieval quality, or task success. What is missing is a holistic view of sustainability that compares their footprints, shows where they complement or amplify one another, and clarifies trade-offs at scale.

This review builds on that gap. By examining these paradigms side by side, we aim to move beyond isolated efficiency improvements toward a system-level understanding of green AI, one that recognizes how models, retrieval systems, and agents interact in shaping environmental impact.

3. COMPARATIVE ENERGY PROFILES

3.1 Large Language Models (LLMs)

The energy footprint of LLMs has two main phases: training and inference. Training frontier systems such as GPT-3 or GPT-4 requires weeks of GPU computation, with GPT-3 alone estimated to have produced 552 metric tons of CO_2 [1]. These costs are substantial but occur only once for each model release. Inference, in contrast, runs continuously and now accounts for most of the lifetime emissions. Efficient models like Claude-3.7 Sonnet can process a query using about 0.4 Wh, while larger models such as LLaMA-70B may require ten times more. Choices in model size, quantization, and batching directly influence this footprint. Cooling adds another hidden cost. Mistral reported that a 400-token prompt consumed about 45 milliliters of water, which scales quickly when multiplied across millions of queries each day.

3.2 Retrieval-Augmented Generation (RAG)

RAG lowers retraining emissions by shifting updates to retrieval, but it brings both upfront and ongoing costs. Creating embeddings for millions of documents is compute-intensive, and storing and refreshing them in vector databases requires continuous energy.

At runtime, retrieval is cheaper than inference, but overall efficiency depends on database size and how often embeddings are updated. In a medical QA study, a LLaMA-3.1 8B RAG pipeline was both more accurate and less energy-intensive than fine-tuned mini-LLMs, cutting emissions by 50 to 100 times through monthly embedding updates. The sustainability of RAG ultimately depends on careful refresh management, since poorly scaled systems can see storage demands outweigh the savings.

3.3 AI Agents

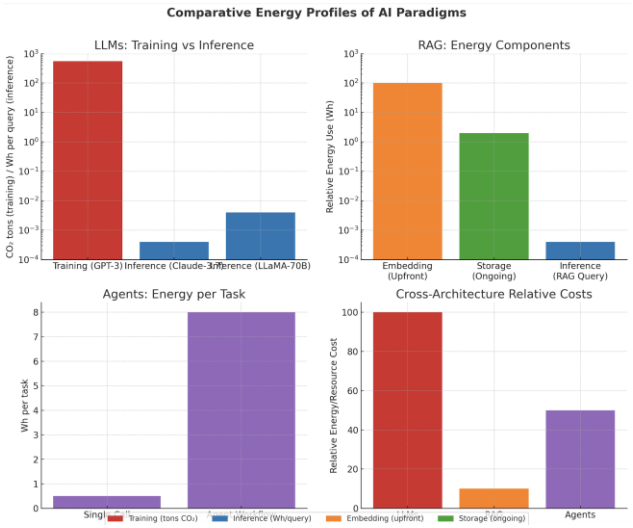
Agents add flexibility but can significantly increase energy use. Instead of a single model call, workflows may involve 10 to 20 queries along with multiple retrievals and memory updates, raising energy costs from about 0.5 Wh to nearly 8 Wh per task.

However, agents do not have to be wasteful. With caching, pruning, and adaptive model selection, they can reduce redundancy and make their calls more efficient. Prototypes such as CarbonCall (2025) have shown that agents can shift between smaller and larger models depending on task complexity, cutting carbon impact by as much as half [3]. This suggests that agents could evolve into energy-aware orchestrators if sustainability is built into their design.

3.4 Cross-Architecture Comparisons

Viewed together, the three paradigms reveal contrasting dynamics. LLMs are expensive to train but relatively predictable at inference. RAG avoids retraining but depends heavily on storage efficiency. Agents expand capabilities but often multiply inference costs. The challenge is not choosing one over the others, but integrating them wisely. Lightweight models paired with lean RAG and coordinated by efficient agents can balance performance with sustainability, while poorly designed combinations risk compounding waste instead of reducing it.

Figure 1. Comparative energy profiles of LLMs, RAG, and agents. The figure illustrates how training and inference dominate LLM costs, RAG shifts emissions to embedding and storage, and agents amplify inference demands unless optimized for efficiency.



4. METRICS & BENCHMARKING

4.1 Energy Use per Query

The clearest measure of efficiency is the energy required to serve a single query, usually expressed in watt-hours (Wh). Benchmarks show large variation. Claude-3.7 Sonnet can handle a 400-token query at about 0.4 Wh, while LLaMA-70B may require 4 to 5 Wh. At billions of queries each day, these differences add up to the energy demand of entire data centers. Hardware also plays a major role, since GPUs, TPUs, and other accelerators differ in performance per watt. Software optimizations such as quantization or batch serving can further reduce energy use [4].

4.2 Carbon Intensity

Energy use must also be considered alongside carbon intensity. The same 0.4 Wh query produces about 0.4 grams of CO_2 equivalent on a hydro-powered grid but as much as 6 grams in regions that rely heavily on coal. Without location-specific disclosures, comparisons across models can be misleading.

4.3 Water Usage

Cooling also adds a hidden footprint. Mistral reported that a 400-token query on its L2 model required about 45 milliliters of water. While this may seem minor, the impact becomes significant when multiplied across millions of daily requests, particularly in regions facing water scarcity.

4.4 Storage and Retrieval Costs in RAG

For RAG, sustainability also includes the cost of creating and maintaining databases. Generating embeddings for millions of documents is highly compute-intensive, while indexes such as FAISS or Milvus consume about 1 to 3 Wh per gigabyte each month for storage and replication. These are ongoing rather than one-time costs, and they grow steadily with the size of the database.

4.5 Beyond Accuracy: Toward Eco-Benchmarks

Traditional benchmarks focus only on accuracy and overlook efficiency. A model that produces the right answer at ten times the energy cost cannot be considered sustainable. Proposed eco-benchmarks, such as measuring joules per correct answer or using composite metrics that integrate energy, carbon, and water, would allow fairer comparisons and encourage designs that balance accuracy with efficiency [5].

4.6 The Challenge of Transparency

Transparency remains a major barrier. Although Hugging Face and Mistral have started publishing carbon and water figures, most providers still do not. A 2024 survey found that more than 80 percent of models lacked any pubic emissions data, which limits accountability and prevents informed decision-making.

4.7 Summary
Metrics are not just technical details but the foundation of accountability. Without them, sustainability remains invisible, and with them, it becomes possible to distinguish responsible systems from wasteful ones. Establishing consistent eco-benchmarks will be essential for aligning AI with long-term ecological goals.

5. GREEN DESIGN PATTERNS
5.1 Hybrid Inference Architectures
A practical way to reduce AI's footprint is through hybrid deployment. Smaller on-device models can handle routine tasks, while more complex queries are sent to cloud models or RAG pipelines. This approach lowers inference and transmission energy, reduces latency, and improves privacy, showing that efficiency and user experience can go hand in hand.
5.2 Semantic Caching and Pruning
AI systems often waste energy by repeating work. Semantic caching helps by storing frequent outputs or embeddings, while pruning reduces unnecessary retrieval or reasoning. In RAG, this can mean limiting retrieval to only a few top documents, and in agents, stopping unproductive loops early. Both approaches cut down the total number of operations rather than making individual steps faster.
5.3 Energy-Aware Orchestration for Agents
Agents consume the most energy but also offer the greatest control. Adding cost awareness to their planning allows them to rely on smaller models for simple subtasks, call larger ones only when needed, and avoid excessively deep retrieval. Prototypes such as CarbonCall have shown that adaptive strategies can cut emissions by up to 50 percent without reducing accuracy.
5.4 Carbon-Aware Scheduling
Not all energy carries the same carbon cost. Carbon-aware scheduling aligns heavy tasks with times or regions where energy grids are cleaner, for example running embedding refreshes in wind or hydro-powered data centers. This approach requires coordination but reduces emissions without changing the models themselves.
5.5 Hardware–Software Co-Design
Sustainability also depends on aligning software with the right hardware. General-purpose GPUs are versatile but not efficient for tasks such as vector search. Specialized accelerators, neuromorphic chips, and optical processors can reduce energy use significantly. A truly green design must extend beyond software and include the underlying infrastructure.
5.6 Efficiency as a First-Class Goal
Together, these patterns reflect an important shift. Efficiency is no longer secondary to accuracy. With billions of users, sustainability must be treated as a core design principle. Many of these efficiency practices also improve latency, reliability, and trust, showing that greener AI can also deliver better performance.

6. POLICY, TRANSPARENCY & INDUSTRY TRENDS
6.1 The Transparency Gap
The biggest barrier to sustainable AI is the lack of consistent disclosure. Most companies provide little or no information on energy use, emissions, or water consumption. A 2024 survey found that more than 80 percent of deployed models had no public sustainability data, leaving comparisons to rely on estimates rather than evidence.
6.2 Early Leaders in Disclosure
A few firms are beginning to change this trend. Mistral AI published a lifecycle audit of its L2 model that included training emissions and water use. Hugging Face introduced an emissions tracker to log energy during training runs. These examples show that transparency is possible and can create pressure for others to follow.

6.3 Regulatory Momentum
Governments are beginning to take action. The EU AI Act may require reporting on resource use for high-impact systems [6], and states such as Virginia and California have started regulating energy and water consumption in data centers. These efforts echo earlier environmental regulations in other industries, where voluntary standards often developed into legal requirements.

6.4 Market Pressure
Enterprises are also driving change by asking vendors to provide AI sustainability scores along with accuracy and latency measures. Some contracts now require carbon reporting as a condition for procurement, turning efficiency into a competitive advantage rather than an afterthought.

6.5 Toward Accountability
Together, these forces point to a future where the environmental impact of AI can no longer remain hidden. Transparency and reporting are likely to become standard, providing the foundation for greener design choices and stronger regulatory accountability.

7. FUTURE RESEARCH DIRECTIONS

7.1 Energy-Aware Agents
Agents are highly flexible but often energy intensive, optimizing for task success rather than efficiency. A key direction for research is to embed energy and carbon budgets into their planning, encouraging them to select smaller models or cached results when possible. With eco-reward functions, agents could shift from being heavy consumers to active stewards of sustainability.

7.2 Eco-Benchmarks and Standardized Metrics
Unlike accuracy leaderboards, there are very few eco-benchmarks that allow systems to be compared fairly. Metrics such as joules per correct answer or carbon per successful task, including water use and RAG storage costs, would make hidden impacts more visible. Over time, eco-efficiency could become as central to evaluation as accuracy itself.

7.3 Federated and Decentralized RAG
Most RAG pipelines rely on centralized vector databases, which create storage and transmission costs. A greener alternative is federated or decentralized retrieval, where devices or local servers store embeddings and use the cloud only when necessary. This approach reduces energy use, strengthens privacy, and aligns with the broader move toward edge computing.

7.4 Hardware Innovation and Co-Design
Software efficiency must be supported by hardware designed for AI workloads. Specialized accelerators for retrieval, neuromorphic chips for reasoning, and optical processors for inference could significantly lower energy use per query. Hardware and software co-design, a practice long used in high-performance computing, offers a path to achieving the same capabilities with only a fraction of today's footprint.

7.5 Integrating Sustainability with Broader AI Goals
Sustainability should be considered alongside fairness, safety, and transparency. Energy-aware agents could also reduce harmful outputs by pruning unnecessary steps, while carbon-aware scheduling may raise fairness concerns across regions. This makes sustainability a cross-cutting priority, central to AI alignment rather than an afterthough.

7.6 Summary
The next decade of research must treat efficiency as a primary objective. Energy-aware agents, eco-benchmarks, decentralized RAG, and sustainable hardware all point toward a future where AI is scaled responsibly, not only for performance but also for the health of the planet.

8. CONCLUSION
AI's rapid growth has delivered remarkable capabilities but also rising costs in energy, carbon, and water. LLMs are expensive to train and operate, RAG reduces retraining but adds storage and refresh demands, and agents expand capabilities while often multiplying inference costs. Each paradigm comes with trade-

offs, and when combined, they make sustainability a system-wide challenge rather than a problem of individual components.

Efficiency can no longer be an afterthought. With billions of daily interactions, even small savings per query scales into massive global impact. Eco-benchmarks, transparent reporting, and policy frameworks will be essential for holding AI accountable, while market demand is already making sustainability a priority in adoption decisions.

The path forward is to design for efficiency from the start: lightweight models paired with lean RAG, coordinated by energy-aware agents, supported by specialized hardware, and scheduled with carbon awareness in mind. If the last decade of AI was about scaling capabilities, the next must be about scaling responsibly, aligning innovation not only with market needs but also with the long-term health of the planet.

REFERENCES:

1. Patterson, David, et al. "Carbon emissions and large neural network training." *arXiv preprint arXiv:2104.10350* (2021).
2. Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in neural information processing systems* 33 (2020): 9459-9474.
3. Paramanayakam, Varatheepan, et al. "CarbonCall: Sustainability-Aware Function Calling for Large Language Models on Edge Devices." *arXiv preprint arXiv:2504.20348* (2025).
4. NVIDIA. (2023). *Inference Optimization Techniques for Large Language Models*. Technical Report.
5. Henderson, Peter, et al. "Towards the systematic reporting of the energy and carbon footprints of machine learning." *Journal of Machine Learning Research* 21.248 (2020): 1-43.
6. European Union. (2024). *EU Artificial Intelligence Act*. Official Journal of the EU.