Rightsizing LLMs: the *pathway* to more sustainable AI

Why optimized models with a smaller footprint will define the next phase of enterprise AI

Capgemini

Is the race toward AI at scale putting **sustainability at risk?**

As more leaders strive to make the switch from discrete pilots to organization-wide AI adoption, concerns about its impact on sustainability initiatives are growing.

Executives are increasingly aware that their AI tools, especially Generative AI and Agentic AI, are driving up energy and water consumption and emissions. Unchecked, this translates to significantly higher costs and regression in net zero progress. Of 2,000 senior executives surveyed by Capgemini, 48% believe that their AI initiatives are increasing their enterprise's greenhouse gas emissions, and 42% are re-examining their climate goals as a result.[1]

There are pressures from elsewhere, too. Regulations and data sovereignty requirements, for example, are putting more scrutiny on how AI workloads are managed. Protecting operational and customer data and ensuring it's processed, stored, and used in approved settings are being driven up the priority list.

If organizations want to avoid their AI ambitions grinding to a halt or their costs and resource use spiraling out of control, something needs to change. Many organizations are currently using large language models (LLMs) that are far larger than needed for their use cases. Simple chatbots, for example, often use models that include parameters for much more complex operations.

Rightsizing these LLMs will be necessary for successful scaled deployments of generative AI. This guide explores how reducing model size will help organizations build future-ready, sustainable generative AI workflows period.

31% of organizations are now incorporating sustainability measures into their generative AI projects[2]

As AI value explodes, so do hidden **resource demands**

The truth is that AI economics break at scale. What works as a small, self-contained pilot suddenly becomes unmanageable and impractical when teams turn their attention to full deployment.

There are well-documented funding and energy requirements for building and training a model; OpenAI's GPT-4, for example, cost close to $100 million to train and used enough electricity to power 5,000 US homes for a year.[4]

But the AI lifecycle is much broader; there's also the hardware manufacturing, ongoing operations, and end-of-life footprint to consider. Rather than flattening or reducing as AI deployments mature, the costs and impact of generative AI workloads continue to mount as long as they're running.

The cost

- Many inference workloads depend on expensive and power-hungry high-end GPUs, increasing both CapEx and OpEx
- Over-parameterized models and lengthy reasoning steps require more memory and compute than the active use cases need
- Agents can consume 1,000 times more tokens than chatbots, increasing workload costs[5]
- As cloud workloads grow, hosts continue to expand their data centers and increase service costs
- Greater demands for compute, power, and cooling all lead to higher utilities spend
- Rapid developments in AI could lead to faster obsolescence for chips, GPUs and other equipment

The impact

- GPUs are built using rare earth materials, a finite resource that requires pollution and waste-heavy mining and processing[6]
- Each unnecessary parameter and reasoning token uses extra resources every time the model runs
- Resource use scales with the complexity of the prompt; an image demands 1,450x the energy of basic text classification[8]
- Cloud providers' data centers contribute to organizations' scope three emissions
- Servicing generative AI's energy and water requirements puts growing pressure on utility networks
- Researchers say e-waste from generative AI could reach 2.5 million tons per year by 2030[9]

1. https://www.capgemini.com/insights/research-library/sustainable-gen-ai/

Costs and resource demands keep growing as organizations add new workloads and use cases, expand access to more of their people, and automate processes to run constantly. Without more effective methods for measuring and containing the impact of AI deployments, organizations will struggle to achieve their AI goals.

74% of executives say a lack of transparency from their generative AI providers makes measurement challenging period[10]

Model compression: smaller footprints and **higher performance for lower cost**

Many organizations focus on optimizing the infrastructure that supports the model, but there's a much more effective route to models that deliver high performance with a smaller footprint. Instead of changing how workloads are distributed or tweaking hardware to boost cooling efficiency, compression and optimization techniques allow teams to reduce the complexity of the model itself.

Methods such as sparsity, pruning, quantization, low-rank adaptation (LoRA), and knowledge distillation compress, optimize, and finetune models in various ways. Combined, they can help organizations run models with fewer parameters, smaller memory requirements, faster inference, and lower energy consumption. And with the right tuning, teams can achieve minimal or even zero accuracy loss, delivering the same capabilities at a lower cost and with a lower environmental impact.

Currently, only 20% of organizations rank a model's footprint as one of the top five considerations in their decision making.[11] But selecting the right initial model can improve energy usage by as much as 70% with only 1% accuracy loss, even before optimization.[12] A mixture of experts architecture, for example, enforces sparsity by only activating parts of the neural network for each input. For many, shifting their perspective to focus on model efficiency could mean the difference between operationalizing AI and abandoning it at the pilot stage.

2. https://www.capgemini.com/insights/research-library/sustainable-gen-ai/
3. https://www.techradar.com/pro/openai-spent-usd80m-to-usd100m-training-gpt-4-chinese-firm-claims-it-trained-its-rival-ai-model-for-usd3-million-using-just-2-000-gpus
4. https://hbr.org/2023/07/how-to-make-generative-ai-greener
5. https://qz.com/agentic-ai-compute-demands-data-center-chip-demand-051126
6. https://theconversation.com/the-race-to-mine-critical-minerals-for-ai-and-clean-energy-is-creating-sacrifice-zones-that-harm-water-and-health-of-worlds-poor-281524
8. https://unu.edu/inweh/news/environmental-cost-of-AIs-Enrgy-use-carbon-water-and-land-footprints

9. https://www.dw.com/en/e-waste-from-ai-computers-could-escalate-beyond-control/a-70619724
10. https://www.capgemini.com/insights/research-library/sustainable-gen-ai/
11. https://www.capgemini.com/insights/research-library/sustainable-gen-ai/
12. https://arxiv.org/pdf/2601.19311

Sustainable by design: Capgemini and **Multiverse Computing**

At Capgemini, we're committed to designing and implementing AI projects that are guided by sustainability principles from the outset. From the very first engagement, we consider leaders' ambitions alongside their operational realities to identify and resolve the tension between digital advancement and sustainability. Our team compares in-demand use cases to available models to match capabilities to capacity.

This approach means our clients can balance performance, security, sovereignty, and impact from pilot to full deployment, with a more comprehensive view of the costs, resource requirements, and emissions involved.

A key part of this commitment is providing clients with the best combination of solutions and partnerships for their needs. And that means connecting organizations with innovative companies like Multiverse Computing to create a tightly integrated AI roadmap.

This collaboration, enabled by Capgemini Ventures, reflects a shared commitment to shaping the future of sustainable enterprise AI - helping organizations right-size their models, reduce computational complexity, and scale innovation within real-world infrastructure constraints.

Multiverse Computing is pioneering efficient and secure AI for organizations around the world. The team delivers tailored solutions for production-ready AI, so leaders can reduce compute costs, resource use, and performance loss while retaining full control across cloud, data centers, and edge devices.

Explore the full model catalog here

CompactifAI: quantum-inspired model compression

Multiverse's CompactifAI technology uses quantum-inspired tensor networks to reduce a model's footprint by up to 80%, cutting its memory and power requirements. Where other techniques target neurons or precision, this versatile method focuses on the model's correlation space for more controlled compression. This means:

- Smaller models with fewer unnecessary parameters
- 50-80% lower demands for compute, energy, and water
- Up to 2x faster inference and model retraining
- Improved performance with up to 4x more requests per second
- Up to 70% lower inference costs[13]

By reducing the hardware footprint required to run large AI models, organizations can accelerate AI adoption while keeping infrastructure costs under control. These compressed and optimized AI models help enterprises extend the life of existing compute resources, reduce dependence on scarce GPU capacity, and scale AI services more economically across cloud, data centers, and edge environments.

Layering multiple optimization approaches together helps organizations right-size their models for specific use cases, enabling more pilots to move successfully into production and delivering a clearer path to AI ROI.

For example, CompactifAI compression, combined with quantization, reduced the memory size of Meta's LlaMA-2 7B by 80%, resulting in **a 75% reduction in energy consumption in a deployment with Telefónica,**

13. https://docs.compactif.ai/models/

Spain's leading telecommunications company. The smaller model delivered comparable performance while dramatically reducing the infrastructure required to serve AI workloads. **Learn more.**

With partners like Multiverse providing model-level efficiency for our clients, Capgemini teams can focus on delivering the enterprise-grade tooling, integration, governance, and scale needed to build a practical strategy for launching AI. Together, we help organizations deploy AI that is not only powerful and secure, but also cost-efficient and sustainable at scale.

A right-sized **strategy for LLMs**

By shrinking models instead of endlessly growing infrastructure, organizations can scale AI in a way that's economically viable and environmentally responsible, with:

More manageable AI economics, with lower hardware and compute demands

By reducing complexity at the source, organizations can serve more users and bring down inference costs to make AI more affordable at scale, without expanding their data center or hardware footprint.

Sustainability and traceability, without performance tradeoffs

When energy and water consumption drop alongside costs, leaders can simultaneously prioritize their environmental commitments and support organization-wide AI initiatives. More comprehensive measurement and reporting can help address regulatory burdens and guide decision-making as leaders target net zero.

Broader deployment options, with smaller, more efficient models

With a more practically sized model at its heart, AI becomes a flexible component of enterprise infrastructure, with options to deploy on premises, and in private clouds, as well as using hyperscaler or colocated data centers. Multiverse's solution even allows organizations to securely deploy fully offline AI on edge devices for lightweight tasks, intelligently routing to API-based models for more complex thinking. These capabilities directly support digital sovereignty and expand the range of viable AI architectures, especially in highly regulated sectors.

The push for AI adoption spans industries, and the speed of change has driven many organizations to expensive, environmentally unsustainable AI operations. Long-term success will reward organizations that take a considered, strategic, and right-sized approach to which models they use and how they put those parameters to work.

Ready to right-size your AI strategy?

Explore how Capgemini's sustainable AI solutions help you reduce model footprint, optimize performance, and scale responsibly.

Authors

Maik Schwalm
Global Sustainability Lead | Cloud Infrastructure Services | Capgemini

Franco Amalfi
Global Sustainability AI Partner Ecosystem Lead | Capgemini

Philippe Cordier
Vice President & Global Chief AI Scientist | Capgemini Invent

For more details, contact:

Infra.global@capgemini.com

About Capgemini

Capgemini is an AI-powered global business and technology transformation partner, delivering tangible business value. We imagine the future of organizations and make it real with AI, technology and people. With our strong heritage of nearly 60 years, we are a responsible and diverse group of 420,000 team members in more than 50 countries. We deliver end-to-end services and solutions with our deep industry expertise and strong partner ecosystem, leveraging our capabilities across strategy, technology, design, engineering and business operations. The Group reported 2024 global revenues of €22.1 billion.

Make it real.
www.capgemini.com

GIS_28.04.26_Soham