

(mailto:info@technative.io)

(<https://x.com/TechNative>)
(<https://www.linkedin.com/company/technative>)
(https://www.youtube.com/channel/UCSDa1_54M05QQxVf-PIQtwQ)

(<https://technative.io>)

[Latest](#) [Articles](#) [Video](#) [Contribute](#) [About](#)

SCALING AI RESPONSIBLY: THE OPEN-SOURCE SOLUTION

The capabilities of AI are accelerating at a lightning pace.

Comparing what ChatGPT could do when it first rose to prominence in 2022 to now is like night and day – the technology has evolved beyond a simple chatbot to being able to generate comprehensive research reports and beyond. However, as its capabilities grow, so do its computing requirements. The US Department of Energy reported that data centre power demands have increased from 30 TWh to nearly 100 TWh over the past decade (<https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf>), and this will barely scratch the sides of what's needed to power multimodal models of the future.

At the core of this challenge lies the infrastructure powering AI. Many of the existing data centre concepts were designed for traditional computing rather than the energy-intensive workloads demanded by AI. As models grow increasingly complex, these facilities are being pushed to their operational limits.

However, the recent release of the AI model DeepSeek R1 demonstrates that there is another way. Although DeepSeek claims to have been trained for under \$6 million – a claim that has faced substantial scrutiny, it still highlights a potential path to reducing model training costs. It borrows significantly from other AI firms' work by distilling prompt data from ChatGPT (<https://hls.harvard.edu/today/deepseek-chatgpt-and-the-global-fight-for-technological-supremacy/>), but in doing so, it manages to lower the cost of model training and inference. The entire source code for DeepSeek is freely available (<https://github.com/deepseek-ai/DeepSeek-V3>) online with a fully open-source software license, allowing other developers to learn about and reuse techniques that drive computation requirements down.

Though it wasn't the first to do this – Meta's Llama models have been open-source since their creation – it is a sign that developers are waking up to a different approach to creating AI. If we want to prevent AI use from impacting sustainability efforts, developers need to work in a way that brings them together.

Open-source provides a crucial framework to solve these global challenges – it promotes more efficient use of compute, reduces inefficiencies through open collaboration, and helps AI scale sustainably.

The power of open-source collaboration

Unlike traditional proprietary systems, open-source software allows developers to share tools and techniques. This, in turn, leads to better quality software that avoids wasting resources by making source code publicly available for scrutiny. Open-source also ensures long-term resilience, as anyone can step in to fix, improve, or adapt source code, which is invaluable in fast-moving sectors like AI.

The proof is in how ubiquitous some open-source projects are. Python, for example, is one of the most widely used programming languages in the world, used everywhere from AI development to financial institutions using it in trading algorithms. It has stood the test of time because collaborative, open-source development practices ensure its adaptability across diverse applications. [The strength of open-source lies in the community](https://www.nscale.com/blog/why-open-source-matters) (<https://www.nscale.com/blog/why-open-source-matters>), where users are contributors and improvements are continuously driven by real-world needs rather than commercial strategy.

Applying this approach to AI can only serve sustainability efforts well. The [OpenUK AI Openness Report 2025](https://openuk.uk/wp-content/uploads/2025/02/AI-Report-FINAL.pdf) (<https://openuk.uk/wp-content/uploads/2025/02/AI-Report-FINAL.pdf>) highlights how open models like DeepSeek R1 and Meta's Llama not only reduce model training costs but also promote transparency by allowing others to verify and reuse efficient techniques, helping to lower overall compute demands. Open-source demands transparency above all. Claims of performance improvements that impact energy efficiency can actually be verified – provided they work, these can then be freely borrowed by other developers for their own AI models. Rather than limiting the keys to making AI sustainable to proprietary secrets owned by a select few companies, a different approach could transform the way the world builds AI. Breakthroughs don't happen in isolation – they rely on incremental advancements from past research.

It's important to recognise how much AI models today depend on open-source projects. Underpinning every AI model, open-source machine learning frameworks like PyTorch are integral for providing GPU acceleration and enable the fundamental mathematical techniques used to build neural networks. When training models, workload orchestration tools shepherd GPUs in data centres into clusters that are fundamental for attaining the necessary compute requirements. The most popular of these tools –

Kubernetes – is open-source. Its widespread use shows how open platforms are trusted for mission-critical infrastructure. It only makes sense to continue the collaborative process that has led to AI models being possible in the first place.

Reinventing data centres for AI

At the level of infrastructure, open-source software helps to streamline compute resource requirements. There is less pressure to optimise for different AI models in terms of hardware when they all share the same practices. Already, the workloads for AI require bespoke infrastructure beyond what can be offered through a traditional data centre. The process to manufacture AI data centres is not inherently sustainable – the energy costs, the extraction of materials for parts, and the processes involved in constructing a facility tend to produce emissions.

However, it is important to recognise that sustainable data centres are not only possible – they're already being built. Data centres have already been evolving to make the best use of resources possible. GPUs designed for AI typically run hotter than traditional data centre compute and significant strides have been made towards housing these in a sustainable way. Advanced cooling technologies like direct-to-chip liquid cooling, where the cooling element is applied to individual GPUs instead of relying on airflow-based methods, help to reduce the energy and water footprint of data centres. Some facilities are even using closed-loop systems to ensure that no water is lost at all.

Data centre energy consumption can be made sustainable by selecting locations that use renewables. As an example, our Glomfjord site is situated in a region with an overabundance of hydroelectric power.

This allows us to directly contribute to sustainability efforts – renewables don't generate power in line with usage peaks, but a data centre runs 24 hours a day, which helps renewable energy projects guarantee that excess clean energy is consistently utilised.

The challenge is made more complex by closed-source AI models. The optimal way to configure hardware for a particular AI model varies significantly between different models. However, when it comes to making

software more efficient, some optimisations are universal. Efficiency improvements are measurable, but when AI models are developed in isolation, it becomes significantly more challenging to set up infrastructure optimised for efficiency. Without open-source practices to establish a degree of uniformity between AI models, optimising hardware performance becomes increasingly difficult. For sustainability, AI models must be compatible with both data centre hardware and inference engines.

Poor optimisation increases emissions and costs for end users.

Building a sustainable future

AI holds immense promise to change the way our industries operate, but its impact on the environment can't be ignored. To shape a sustainable AI future, we need to focus on more than just what we build – how we build it matters. By embracing open-source practices and the genuine collaboration these bring, we can continue to make progress with AI without harming our planet on the way.

About the Author

Nick Jones is VP of Engineering at Nscale (<https://www.nscale.com/>). Nscale is the Hyperscaler engineered for AI, offering high-performance compute optimised for training, fine-tuning, and intensive workloads. From our data centres to software stack, we are vertically integrated in Europe to provide unparalleled performance, efficiency and sustainability.

Featured image: seventyfourimages



MORE INSIGHTS



(<https://technative.io/navigating-cyber-threats-in-the-retail-sector/>)

Navigating Cyber Threats in the Retail Sector
(<https://technative.io/navigating-cyber-threats-in-the-retail-sector/>)



(<https://technative.io/it-strategies-to-navigate-the-ever-changing-digital-workspace/>)

IT Strategies to Navigate the Ever-Changing Digital Workspace
(<https://technative.io/it-strategies-to-navigate-the-ever-changing-digital-workspace/>)



(<https://technative.io/why-its-a-smart-move-to-grow-a-tech-company-in-the-middle-east-amid-economic-volatility/>)

Why it's a smart move to grow a tech company in the Middle East amid economic volatility
(<https://technative.io/why-its-a-smart-move-to-grow-a-tech-company-in-the-middle-east-amid-economic-volatility/>)

(<https://technative.io>)

Technology at Work

CONTACT US

info@technative.io(<mailto:info@technative.io>)

(<https://x.com/TechNative>)

(<https://www.linkedin.com/company/technative>)

JOIN THE LIST

Get updates on special events and upcoming content

Email

SUBSCRIBE