PRIMUS PARTNERS®
Solutions for Tomorrow

From

Data Centers to AI Factories:

9 GW Capacity, 3% Power, 358 Billion Litres of Water

India’s Sustainable Leadership Ambition

Foreword

Mr. Nilaya Varma

Group Founder and CEO
Primus Partners

In a world increasingly shaped by unprecedented technological change, the emergence of Generative AI represents a pivotal moment. This technology promises to unlock new frontiers of productivity and innovation, but its ascent also brings a critical and complex challenge to the forefront: the escalating environmental footprint of our digital infrastructure. This is not merely a technical problem; it is a strategic imperative that requires a new paradigm of thinking.

India's climate journey is anchored in the ambitious Panchamrit commitments, which aim to achieve 500 GW of non-fossil capacity by 2030 and a net-zero future by 2070. These pledges are more than just climate targets; they are a signal to businesses, investors, and citizens that India is determined to chart a sustainable growth pathway. This context makes the energy demands of Generative AI particularly urgent. With India's data center capacity projected to rise from ~1 GW to 9 GW by 2030, the energy required to power this growth places immense pressure on our ability to meet national climate goals.

This report, **"From Data Centers to AI Factories: 9 GW Capacity, 3% Power, 358 Billion Litres of Water- India’s Sustainable Leadership Ambition"** is designed to serve as a comprehensive roadmap for navigating the critical intersection of this digital ambition with our national climate goals. At Primus Partners, we are dedicated to bridging the gap between national policy and on-the-ground implementation. Our work focuses on helping leaders across government and the private sector translate ambitious visions into actionable strategies. We believe this report is a testament to that mission, providing a blueprint for how we can leverage policy, technology, and public-private partnerships to build a sustainable digital economy.

As India marches towards the Viksit Bharat 2047 vision, our success will depend on how effectively we can embed sustainability into the very core of our economic and technological growth. This is the moment to reframe the climate cost of AI from a burden to an opportunity.

By embedding efficient technologies and integrating climate data into strategic decision-making, businesses can turn the Panchamrit vision into a corporate reality. The choices we make today in infrastructure, policy, and innovation will determine whether India’s growth story is also the world’s sustainability story.

01 The Challenge to Ambition

1.1 The Unseen Price of Intelligence

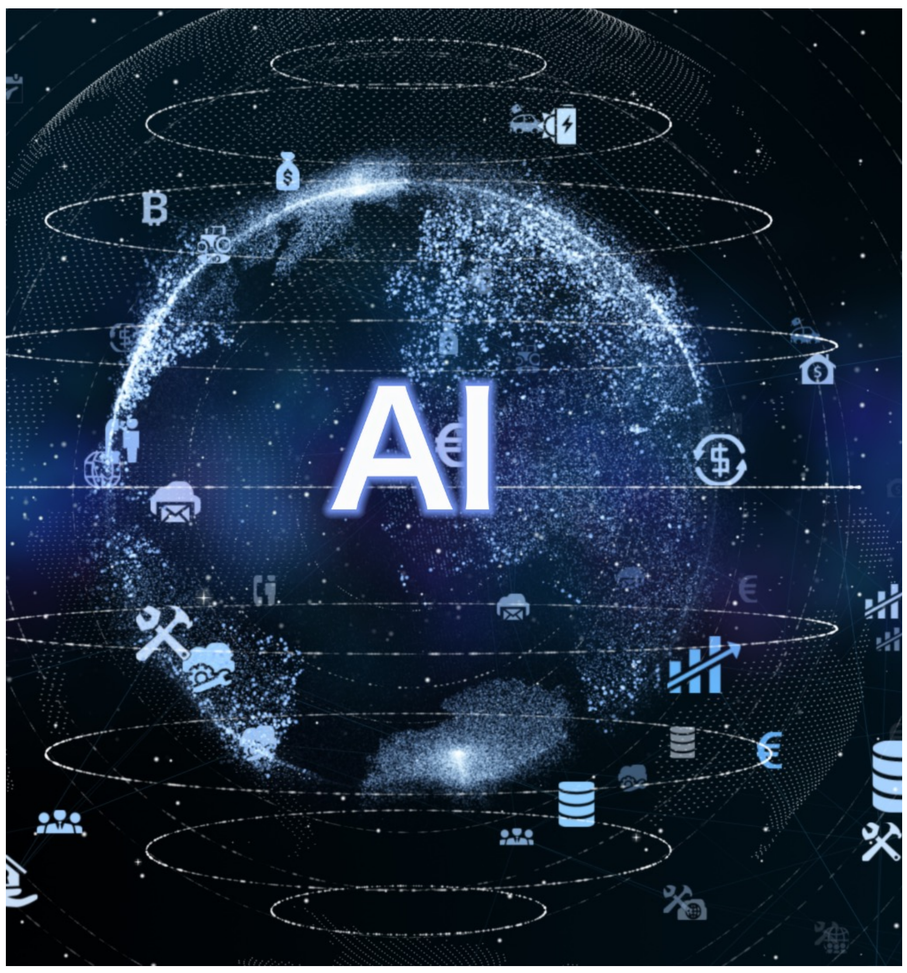
Generative AI is a powerful tool that has changed how we create, interact with, and analyze information. In just a few short years, generative AI systems such as ChatGPT, Gemini, Claude, and Stable Diffusion have become a part of our everyday lives. Chatbots draft legal briefs, image models generate lifelike art, and language models assist doctors (of course, with discretion) and researchers in real time.

AI is also pushing the boundaries of human imagination and capability in ways once thought impossible. Scientists use AI to sift through terabytes of telescope data, identifying new exoplanets and mapping galaxies light-years away something no human team could do alone. In medicine, AI has accelerated the discovery of new drugs, such as antibiotics effective against resistant bacteria, and is now helping radiologists detect cancers at early, more treatable stages. During the COVID-19 pandemic, AI models were critical in tracking the spread of the virus and accelerating vaccine research. In disaster response, AI-driven early warning systems are saving lives by predicting floods, wildfires, and hurricanes with increasing accuracy. These breakthroughs highlight AI's extraordinary potential to serve humanity.

This revolution, however, is not powered by "intelligence" in the abstract. It is powered by an immense physical infrastructure of chips, servers, data centers, and electricity and water. The sleek, seemingly weightless experience of asking a chatbot a question masks the reality that each response draws from global supply chains of rare earth metals, coal-fired electricity, and millions of litres of water.

Unlike earlier waves of software, generative AI is not a "lightweight" technology. It is instead powered by one of the most resource-intensive infrastructures ever built, and its environmental costs are only beginning to be understood.

This chapter examines that hidden infrastructure and the planetary price tag attached to it. It begins by unpacking what exactly powers generative AI and how those systems consume energy and natural resources. It then turns to the four major challenges that define AI's environmental burden: exploding carbon emissions, an insatiable appetite for electricity, enormous amount of e-waste generated, and a growing, but often overlooked, demand for freshwater. Together, these costs form a challenge that will shape not only the sustainability of AI but also its equity and fairness across regions of the world.



1.2 How Generative AI Works – and Why It Needs So Much Power

To understand why Generative AI is so resource intensive, it is important to first understand how it differs from earlier AI.

- **Traditional AI and machine learning** were built to recognize patterns in data and make predictions. For example, it might look at X-rays and say whether a scan shows an abnormality or not. The goal was classification and decision-making.
- **Generative AI,** by contrast, doesn't just recognize patterns, it creates new content from them. Instead of only labelling an X-ray, it can generate a doctor's note, answer follow-up questions, or draft a treatment summary. That shift, from recognizing to generating, is what makes it so powerful, but also far more demanding.

Generative AI models learn the relationships between words, images, or sounds in massive detail. That scale of learning requires enormous computing power, which in turn depends on three things, which are not mutually exclusive and operate like nested layers.

GPUs/ TPUs:

Generative AI learns by analysing billions of data points to identify patterns. To process such enormous amounts of data, it requires specialized high-performance chips like GPUs and TPUs. These chips perform thousands of calculations in parallel, powering both the training phase, where the model learns, and the inference phase, where it generates new outputs. The GPUs/ TPUs, sit inside data centers and are a critical component. They are being treated as a distinct driver, because they are the primary source of power draw within a data center and directly linked to AI's compute intensity.

These chips are designed for speed and scale, but they also draw extraordinary amounts of power. Each chip uses a lot of electricity, and tens of thousands of them are often linked together in clusters running continuously, making specialized hardware the engine that drives Generative AI. Today's most powerful new data center GPUs for AI workloads can consume as much as 700 watts apiece. With a 61% annual utilization, that would account for about 3,740,520 Wh or 3.74 MWh per year per GPU.[1]

Data centers –

These are vast facilities that serve as the physical homes of AI. Data centers are the housing and the infrastructure layer: physical buildings, racks, servers, networking equipment, HVAC, etc. Their job is to integrate thousands of GPUs/TPUs, keep them powered, cooled, and connected. This is where many AI models are trained and deployed. These facilities sprawl across the U.S., Europe, and increasingly Asia. As of March 2024, there were approximately 11,800 data centers worldwide. Data center construction reached an all-time high in 2023, with 3,077.8 Megawatts under construction. This is a 46% year-over-year increase.

Country	No. of Data Centers
United States	5426
Germany	529
United Kingdom	523
China	449
France	322
Australia	314
Netherlands	298
India	270
Russia	251
Japan	222
Brazil	196
Mexico	173
Italy	168
Poland	144
Spain	143
Hong Kong	122
Switzerland	121
Singapore	99
Sweden	95
Indonesia	84
New Zealand	83
Belgium	80
Austria	68
Malaysia	62
Chile	59
Ukraine	58
Ireland	55
Denmark	50
Finland	48
Norway	47
South Korea	43

Table 1: Number of Data Centers by Country

Each data center hosts racks of servers running 24/7. These servers generate enormous amounts of heat while operating thousands of high-performance chips continuously. To manage this, data centers rely on cooling systems, which is essential to prevent overheating, maintain performance, and avoid hardware damage.

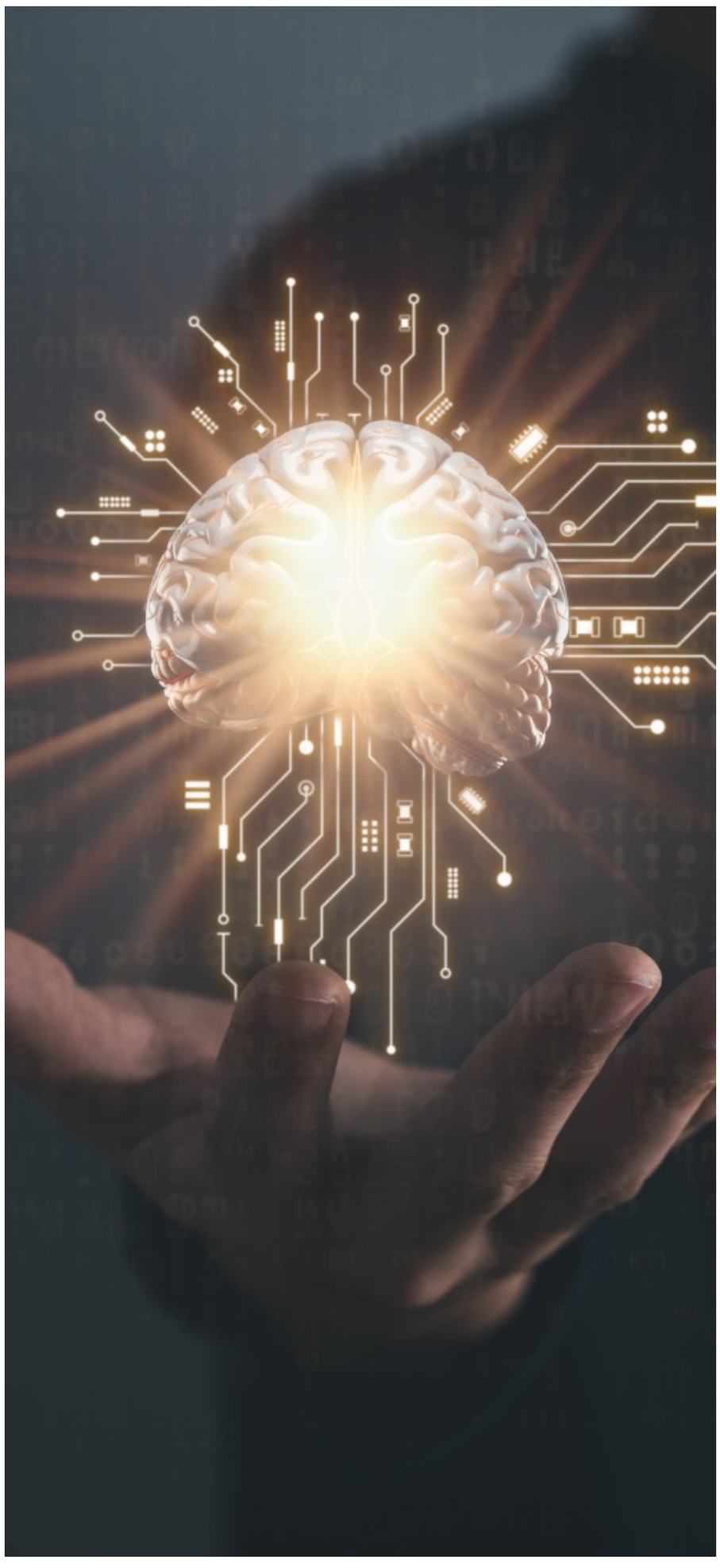
Data centers provide the space, continuous power, and infrastructure needed to run these chips non-stop. Unlike traditional IT workloads, Generative AI requires constant retraining, fine-tuning, and large-scale inference, which makes these facilities far more resource-intensive than typical cloud services.

Energy and water inputs –

The real fuels of AI are electricity and water that keep both the chips and data centers running. Running thousands of chips around the clock consumes enormous electricity. In September 2024, Microsoft announced an agreement to re-open the Three Mile Island nuclear power plant to provide 100% of its electric power for 20 years, aiming to support its AI data centers with clean energy. Globally, data centres consumed around 1.5% of electricity consumption in 2024. AI is only one of a range of workloads that data centres perform, but in anticipation of growing demand for AI-related services, investment in data centres is growing rapidly and the size of the largest data centres is increasing. A hyperscale, AI-focused data centre can have a capacity of 100 MW or more, consuming as much electricity annually as 100 000 households. AI-focused data centres are increasing in size to accommodate larger and larger models and growing demand for AI services.

In short, Generative AI is powered by specialized hardware working non-stop in large data centers, with electricity to run the chips and water to keep them cool. The more data the model learns from and the larger it grows, the greater the demand on these resources.

Now that we know what powers Generative AI, we turn to the costs this infrastructure imposes on the planet.



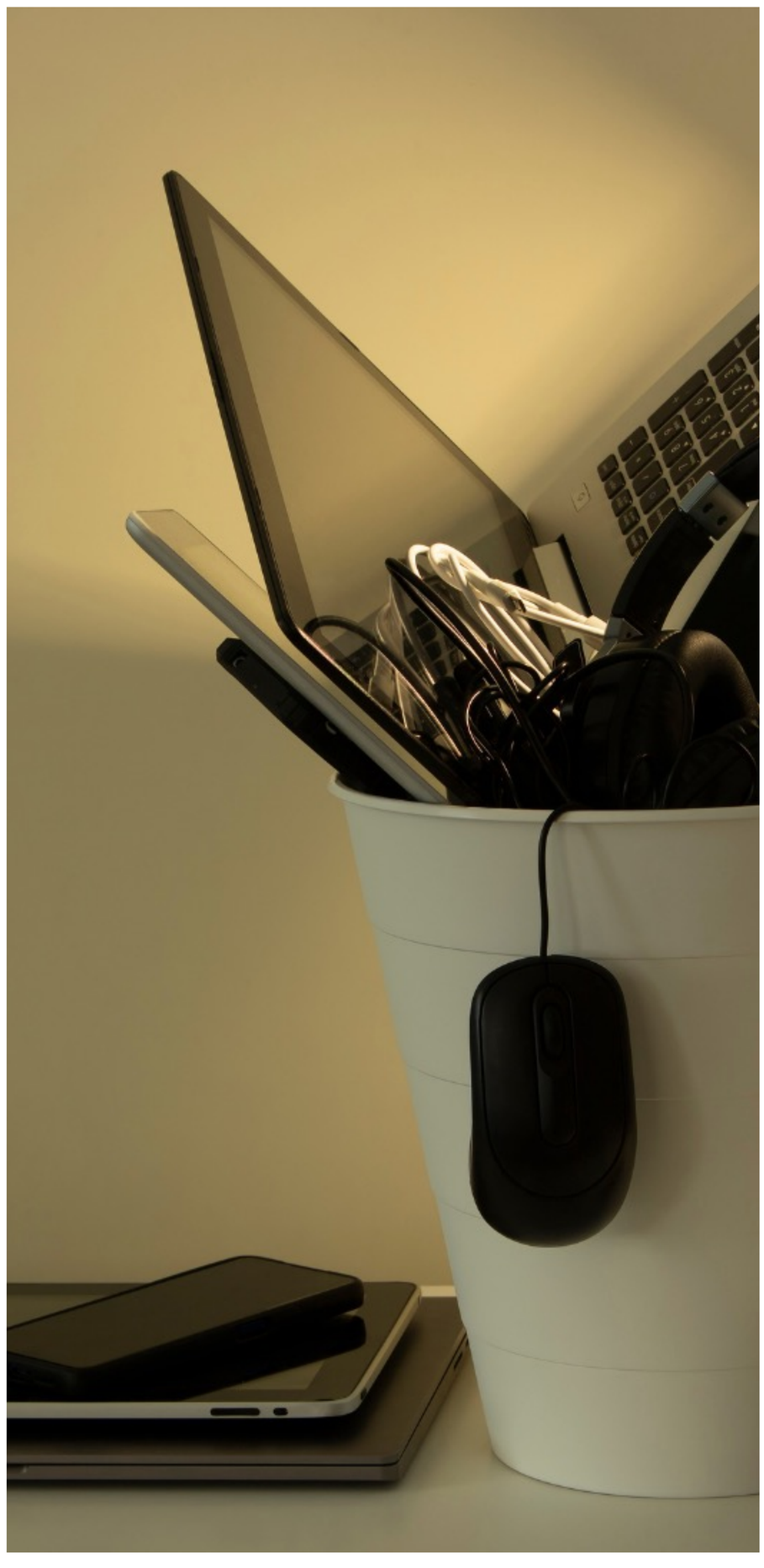
1.3 The Planetary Price Tag

Artificial intelligence is transforming our world, but not without a cost. From energy and water consumption to emissions and e-waste, AI's environmental footprint is large, growing, and often hidden. Below, we explore the primary challenges posed by AI to the planet.

1.3.1. The First Challenge: Energy Consumption

The AI lifecycle, from development to deployment, requires a lot of computing power. Generative AI is reshaping global electricity consumption in ways that may promote the rise of entire new industries. The International Energy Agency projects that electricity demand from data centers, cryptocurrencies, and AI combined will more than double between 2022 and 2026, growing from around 460 terawatt hours to over 1,000 TWh. AI is the fastest-growing driver of this expansion. To put that in perspective, this is equivalent to the yearly electricity use of Japan, one of the world's largest economies.

Training and inference together have created a continuous demand cycle, making AI one of the fastest-growing consumers of electricity worldwide.



Scandinavia, creating both economic opportunities and significant local pressures on grids, water resources, and land. In Northern Virginia, the largest data center hub in the world, utilities have already warned of transmission bottlenecks and potential delays in meeting demand. In Ireland, where data centers consume more than 18% of national electricity, authorities have paused new grid connections to avoid blackouts. These examples illustrate that AI's energy demand is not only a global concern but also a local stressor, affecting infrastructure reliability and communities.

Grid Strains and Fossil Dependency

This demand surge is not evenly spread across the globe but is concentrated in specific "AI clusters," regions where data centers, cloud providers, and connectivity converge, such as Northern Virginia in the U.S., Dublin in Ireland, Singapore, and parts of

Beyond local strain, AI's energy requirement is closely tied to the type of power feeding the grid. In many regions, electricity grids remain heavily reliant on fossil fuels. In the United States, a rapid expansion of AI-driven data centers has slowed the retirement of natural gas plants. In parts of Europe, new AI facilities have been approved only with conditional fossil backup to prevent shortfalls. Even when companies purchase renewable energy, contracts often fail to match AI's continuous, 24/7 demand. The result is a structural dependency on coal and gas that risks embedding AI growth in high-carbon grids for decades to come.

AI Energy Lifecycle

- **Training Phase:** Training a model involves exposing it to large datasets to learn patterns. This process is extremely resource intensive. For instance, GPT-3 training consumed approximately 1.287 GWh of electricity. To put that in perspective, this amount is roughly equal to the annual energy use of over 100 U.S. households. The computational complexity increases exponentially with the model size and the dataset, making this an important element.

- **Inference Phase:** The training phase happens only once for each model version, but the inference phase, where the trained model makes predictions or generates outputs, occurs continuously. A single query to an AI model, like a search request or a language translation, uses a small amount of energy. However, basis billions of daily interactions with AI services, the total energy consumption adds up to a huge amount. Google in August 2025, released a technical report detailing how much energy its Gemini apps use for each query. In total, the median

What makes an AI cluster?

AI clusters are regions where computing, connectivity, and capital converge to create the world's digital powerhouses. Their rise is not accidental; they form around a unique mix of factors:

- **Reliable, inexpensive Power –** Abundant electricity (often from fossil sources, but increasingly renewables) to sustain 24/7 demand.

- **Cooler Climates –** Regions like Scandinavia attract data centers because lower ambient temperatures reduce cooling needs.

- **High-Speed Connectivity –** Dense fiber optic networks and submarine cables allow seamless data transfer.

- **Skilled Workforce –** Proximity to talent pools in engineering, AI research, and cloud operations.

- **Government Incentives –** Tax breaks, subsidies, and land availability encourage investment in hyperscale facilities.

- **Strategic Geography –** Hubs like Singapore or Dublin serve as gateways to regional markets.

prompt, one that falls in the middle of the range of energy demand, consumes 0.24 watt-hours of electricity, the equivalent of running a standard microwave for about one second. This report was strictly limited to text prompts, so it doesn't represent what's needed to generate an image or a video. The report also finds that the total energy used to field a Gemini query has fallen dramatically over time. The median Gemini prompt used 33 times more energy in May 2024 than it did in May 2025, according to Google. The company points to advancements in its models and other software optimizations for the improvements. The same is true for other organizations too. Thus, as AI becomes more common, this ongoing energy demand will grow to become the main source of its carbon footprint, if not checked.

- **Data Center Infrastructure:** Most AI workloads take place in large data centers, which consume energy not only for computation but also for cooling, ventilation, and support systems. A common measurement for this is Power Usage Effectiveness (PUE), which compares the total power used by the facility to the power used by IT equipment. While modern data centers are becoming more efficient, with PUEs approaching 1.1, the average PUE is still around 1.5. This means that 50% of the energy consumed is for non-computational needs.

1.3.2. The Second Challenge: The Water Crisis You Don't See

Perhaps the least visible impact of AI is its strain on freshwater resources. Cooling is central to data center operation, and most systems rely heavily on water. Training GPT-3 alone consumed about 700,000 liters of freshwater, while large hyperscale data centers can use millions of liters daily. The water use of AI is not just a direct cost; it also puts pressure on local water sources, especially in areas already dealing with water shortages.

Local and Ecological Impact

Water is often the cheapest cooling option for data centers, making evaporative cooling systems widespread, even in regions already struggling with water scarcity. Facilities in arid deserts compete with agriculture and mining for limited aquifer reserves, while in parts of Europe, proposed mega–data centers have faced community opposition over projected water withdrawals. The challenge doesn't end with consumption. Heated discharge water can destabilize aquatic ecosystems, while cooling systems themselves are vulnerable during extreme heat, precisely when water is scarcest. This creates a paradox: the hotter the planet gets, the more water AI requires.

1.3.3. The Third Challenge: E-Waste and Resource Depletion

The hardware at the core of AI, specifically GPUs, CPUs, and TPUs, have a limited lifespan and adds to the growing e-waste problem. Some major concerns are below:

- **Rapid Obsolescence:** The fast pace of innovation in the semiconductor industry means that high-performance chips used for training today can become outdated within 1-2 years as newer, more efficient options are released. This cycle of constant upgrades creates a large amount of electronic waste, which is many times improperly disposed of.

- **Resource Depletion:** Producing this hardware needs a significant amount of rare-earth minerals and other valuable materials like cobalt, lithium, and gold. Mining and processing these resources is not necessarily the best for environment, causing habitat destruction, soil erosion, and water contamination. The non-renewable nature of these materials and the challenges in recycling them from complex circuit boards presents a long-term sustainability issue.

1.3.4. The Resulting Challenge – Emissions Explosion

AI's rise is coming with an explosion in carbon emissions. Every stage of its lifecycle – from training massive models, to running them daily at scale, to manufacturing the chips that power them – leaves a growing climate footprint.

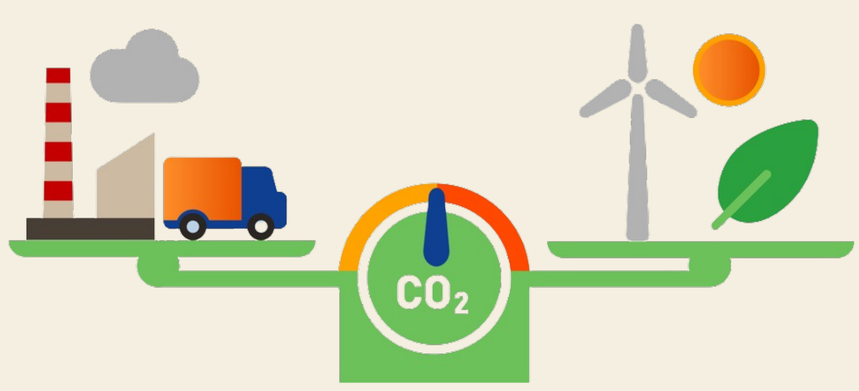
Training Emissions

Training is the most carbon-intensive stage of AI. To build a large model, tens of thousands of GPUs are kept running simultaneously for weeks or even months, often on fossil-fuel-heavy grids. The result is enormous one-time emissions that are "locked in" the moment the model is created.

Case Box:
The Carbon Cost of GPT-3

Training OpenAI's GPT-3 used about 1,287 MWh of electricity, and produced over 500 metric tons of CO_2e.

For perspective, the average Indian household emits about 6.5 metric tons of CO_2 per year, meaning GPT-3's training emissions are roughly equal to what more than 75–80 Indian households emit in a year. And this is just one model.



As model size grows, so do the emissions. GPT-4, for instance, is estimated to be several times more carbon-intensive than GPT-3 because of its vastly larger parameter count. The scale is now such that the emissions from training a state-of-the-art model can rival the annual footprint of smaller nations. At scale, the numbers are staggering. IEA estimates that if current growth trends continue, data centers could emit more than 1 gigaton of CO_2e annually by 2030, offsetting much of the progress made in renewable energy deployment.

Inference Emissions

After training, AI models continuously serve billions of queries each day. Every chatbot query, image generation, or translation task requires real-time computation. The emissions per use are smaller than training, but because billions of queries are processed daily, the total emissions from inference may overtime exceed those from training. This would make AI's climate impact continuous rather than one-off.

Hardware Emissions

But operational emissions are only part of the picture. AI hardware production itself is highly carbon-intensive. The manufacturing of GPUs and TPUs requires semiconductor fabrication processes, while the mining of rare earth minerals such as cobalt, lithium, and nickel adds embodied carbon costs. Every AI training cycle therefore carries not only the emissions from electricity consumed but also from the upstream production and supply chains of the hardware used.

Put simply, the emissions challenge is twofold:

1. **Operational emissions** from electricity use in data centers, heavily tied to fossil-heavy grids.

2. **Embodied emissions** from the production and replacement of AI hardware, including chips and servers.

Unless addressed, this dual footprint risks locking AI into a high-carbon growth trajectory, undermining global net-zero pathways and climate goals.

1.4 The Emerging Dilemma

As climate scientist ***Dr. Kate Marvel noted:***

> ***"AI reflects the choices we make in building it. If it is designed to be energy-hungry and resource-intensive, that is not a flaw but a decision. The question is not whether AI will solve the climate crisis, but whether we will design it to."***

This underscores the urgency of building AI systems that are not only powerful, but also sustainable. This emphasizes that finding a balance between reaping the benefits of Generative AI and reducing its environmental impact is crucial. The design choices made today will determine whether AI evolves into a high-carbon system or a more sustainable one.

In short, AI's environmental challenge is not just about energy use, water withdrawals, or carbon emissions in the abstract, it is about how these pressures are distributed and managed within specific national contexts. Nowhere is this balance more delicate than in India. As the country positions itself as a future hub for AI innovation and digital infrastructure, it must do so while navigating existing realities of coal-dependent power, water-stressed geographies, and deep socio-economic inequities. The following chapter turns to India's story: how its ambitions, constraints, and choices will shape not only the trajectory of its own AI ecosystem, but also its role in the global conversation on sustainable and equitable technology.

02

India’s case in leading AI: Power, Data, and the New Digital Divide

India’s ambition to become a leader in artificial intelligence is inseparable from the question of power. In the AI era, it’s not oil or gold that drives competitiveness – it’s data. Whoever controls data, controls the future. But data is not an intangible good that can be accessed without cost. Unlike oil, data can only be extracted through a vast infrastructure of servers, chips, and data centers. These facilities, in turn, consume enormous amounts of electricity and water, making AI into among the most resource-intensive industrial systems of the digital age.

Each new data center represents not just a digital asset but also a significant additional burden on already fragile electricity and water systems. For India, which has set its sights on becoming a global leader in artificial intelligence as part of its Viksit Bharat 2047 vision, this raises a fundamental question: how can the country expand its digital backbone without undermining equitable access to essential resources?

2.1 The Structural Challenge: Power and Capacity Lock-In

The expansion of AI-ready digital infrastructure represents a new and potentially disruptive new source of demand. India’s data center capacity, just over 1 gigawatt (GW) in 2024, is projected to increase to 9 GW by 2030. This appears modest compared to global hyperscale markets like the U.S. or China. Yet the scale of demand relative to India’s grid constraints makes the challenge acute.

A single hyperscale facility can consume as much electricity as 100,000 households. The clustering of multiple such facilities in a handful of urban regions magnifies the load. The result is a form of “capacity lock-in” where large fractions of local electricity capacity are tied up in serving digital infrastructure rather than broader development needs.

The risks are several. First, households and small enterprises may face either rising tariffs or unreliable supply. Second, essential services such as hospitals, schools, and public utilities may be forced to operate

under tighter constraints. Third, unless renewable generation is expanded rapidly, additional demand will likely be met through coal, deepening India's carbon dependency and jeopardizing its net-zero 2070 target.

In this context, efficiency is no longer just a technical target but an ethical imperative. The only sustainable path forward is to reduce the energy intensity of data centers themselves, so that innovation does not come at the cost of basic access.

2.2 Uneven distribution of AI's burden: Who bears the costs?

2.2.1. India Data Centre Landscape

India today hosts an estimated 270 data centres across 31 markets, but the vast majority are concentrated in just a few metropolitan hubs. Mumbai leads with 46 facilities, followed by Hyderabad (33), Chennai (31), and Bengaluru (31).

The Delhi-NCR region (New Delhi, Noida, Gurgaon combined) adds another 36, making it one of the largest clusters in the country. Mumbai accounts for nearly half of new capacity in 2024, as it serves as the landing point for multiple cables and houses the densest cluster of hyperscale and colocation facilities. Beyond these core hubs, secondary centres are emerging in Pune (18), Ahmedabad (10), and Kolkata (9).

India's data centre market valuation is projected to nearly double by 2030, requiring billions in new investment and significant growth in real estate space. Most new capacity is being built for cloud services and AI – especially by global tech firms and domestic players focused on meeting India's rapidly rising data localization and processing needs. Alongside big hyperscale facilities, smaller edge data centres are emerging in tier-II and tier-III cities to support low-latency applications and smart city initiatives.

2.2.2. Distributional Inequity across Regions

The geography of AI infrastructure in India reflects a double inequity. On one hand, wealthier states such as Maharashtra, Karnataka, Tamil Nadu – with stronger grids and better industrial policies- are already capturing the bulk of new data center investments. These hubs are becoming the backbone of India’s AI economy, positioning themselves as magnets for global cloud firms and domestic digital investment.

On the other hand, poorer or resource-stressed states face structural disadvantages that leave them either excluded from this digital transformation or vulnerable to risky forms of participation.

These limitations mean that while some regions advance rapidly, others risk being sidelined, deepening the prospect of a two-speed India: one integrated into the AI economy, the other left behind.

This imbalance is not unique to India – global patterns show similar inequities. As a result, the economic value is largely captured by global corporations headquartered in the U.S., Europe, and China, while the environmental costs are externalized to host regions.
What emerges is an inequitable distribution of costs and benefits.

- The **economic value** of AI is concentrated in global corporations and urban hubs.
- The **environmental costs** – high water consumption, energy demand, and e-waste – are pushed onto host regions and local communities.

2.2.3. Community-Level Burdens

If distributional inequity reflects the state-level divide, community-level burdens reveal how these inequities are felt on the ground. Farmers, peri-urban residents, and informal workers often shoulder the hidden costs of AI's expansion, while having little voice in the decisions that shape it.

- Water-scarce regions: In arid areas, irrigation water is already limited, yet new data centers draw on the same reserves to power cooling systems. This diversion can deepen rural distress for farmers. Similar tensions are evident elsewhere, where local farmers and Indigenous communities have raised concerns over water allocations to digital infrastructure

- Fossil-dependent regions: With over 70% of India's electricity still generated from coal, every additional unit of demand from AI amplifies emissions. Communities near coal plants bear the brunt through greater exposure to particulate matter, toxic fly ash, and land degradation. Comparable challenges are emerging in other developing regions where AI infrastructure remains tied to fossil-heavy grids, locking local populations into long-term health and environmental risks.

- Informal economies: Informal workers in e-waste recycling streams are exposed to toxic materials from discarded servers and GPUs. Feeding into the informal recycling economy, these workers absorb environmental and health costs while gaining little from the economic benefits of AI's growth.

2.3 Areas that need to be addressed: AI, Governance, Social Equity and the Environment in India

Despite the accelerating scale of AI, national and global policy frameworks remain far behind the curve. Most regulations focus on enabling digital growth, securing data sovereignty, or attracting foreign investment. Far less attention is given to the environmental and social safeguards that should govern how and where AI infrastructure is built. This mismatch has created three major blind spots.

2.3.1. Areas for policy attention and Recommendations

Energy and Carbon Accounting

- **Policy attention area:** AI's high computing demands significantly increase electricity use, but most policies do not yet measure or report its carbon footprint in a systematic way.

- **Indian Context:** With much of India's electricity still coal-based, additional demand from AI can add to emissions. While the country is rapidly expanding renewable energy, linking AI infrastructure growth more explicitly with clean energy targets would help balance innovation and sustainability.

Water Use and Local Resource Competition

- **Policy attention area:** Policies guiding where data centers are located rarely consider water availability, even though cooling systems may rely on large volumes of it.

- **Indian Context:** In water-stressed regions, new data centers could place additional pressure on resources already needed by farmers and local communities. India could benefit from guidelines that integrate water–energy trade-offs into digital infrastructure planning.

E-Waste and Recycling

- **Policy attention area:** Frequent hardware upgrades for GPUs, servers, and chips used in AI generate growing amounts of e-waste, which is not always captured under existing recycling frameworks.

- **Indian Context:** India has an Extended Producer Responsibility (EPR) policy for e-waste, but it is still evolving to handle the specialized components used in AI infrastructure. Strengthening formal recycling pathways could reduce risks for informal workers and improve recovery of valuable materials.

Equitable Access and Benefits

- **Policy attention area:** National AI policies tend to emphasize competitiveness and innovation, but less attention is given to who benefits and who may be left behind.

- **Indian Context:** Urban areas are early beneficiaries of AI adoption, while rural communities may face resource competition without equivalent digital gains. Linking AI applications to agriculture, healthcare, and rural development could help distribute benefits more evenly.

Integrated Governance

- **Policy attention area:** AI is often governed through innovation and IT-focused ministries, with limited integration of environmental or social equity perspectives.

- **Indian Context:** Greater coordination across ministries, including MeiTY, MOEF&CC, MoA&FW, MoRD, would allow AI policy to reflect multiple national priorities, from net zero to rural livelihoods.

Innovation and Sustainable AI Solutions

- **Policy attention area:** While AI policy has focused on scaling technology and boosting competitiveness, there is scope for guidance on fostering innovation that is environmentally sustainable and socially inclusive.

- **Indian Context:** Encouraging research and development in low-energy AI models, AI-enabled resource optimization (e.g., energy, water, waste management), and locally relevant solutions can ensure that innovation drives both economic growth and sustainability.

As India advances its AI ecosystem, governance frameworks are steadily evolving to keep pace. As much of the current focus has been on enabling innovation, strengthening competitiveness, and positioning the country as a global hub, there is also an opportunity to place greater emphasis on equity, ensuring that the benefits of AI are widely shared and that communities are supported in managing local impacts.

Policy attention area: Equity considerations are not yet fully integrated into AI governance. Issues such as how resources are allocated, how communities experience digital infrastructure, and how smaller enterprises can participate in the AI economy are still emerging areas of policy attention.

Indian Context: Urban centers and advanced industrial states are often the first to benefit from AI investments, while rural and resource-constrained regions may experience fewer direct gains. Proactively incorporating social and environmental equity into governance, for instance, through broader consultations, transparent reporting on resource use, and programs that link AI to rural development, can help bridge this gap. By doing so, India has the chance to set a global example of how AI growth can advance both innovation and inclusion.

03 Charting a Cleaner AI Future

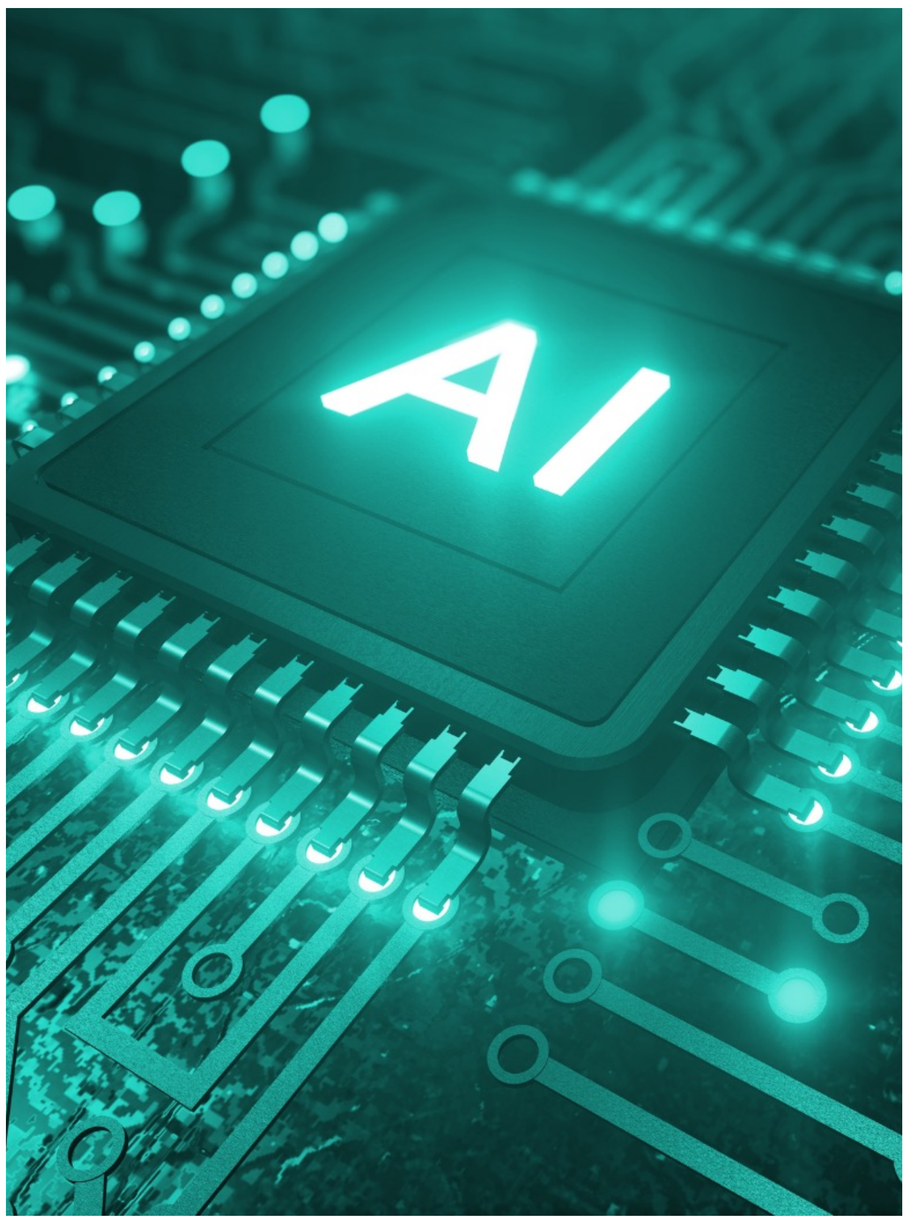
3.1 From Urgency to Opportunity

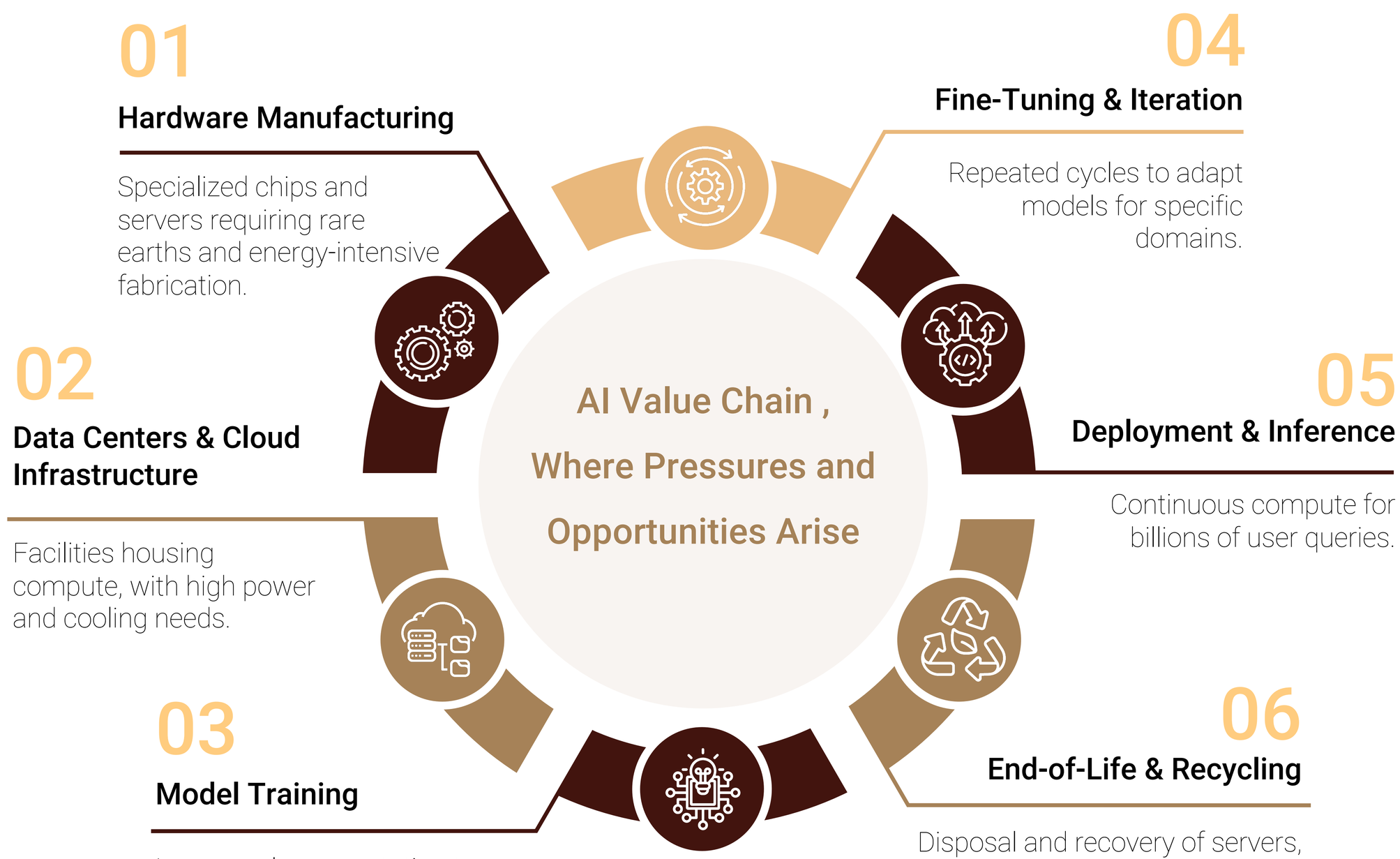
Sections 1 and 2 have discussed the stakes. We now know that AI's environmental footprint is not marginal, it is massive, increasing, and disproportionately felt in water-scarce, fossil-fuel reliant, and low-income regions. Gen AI - celebrated for its intelligence but powered by a resource appetite that our planet cannot sustain at current trajectories.

The question before us is therefore urgent and practical: Can AI scale without scaling its environmental harms? Experts across energy, climate, and technology argue that the answer is yes , but only if we fundamentally rethink how AI is designed, powered, and deployed.

This section sets out a roadmap for that rethinking. Instead of accepting AI's planetary cost as an unavoidable trade-off, we can take a path where the same way that we built the intelligence revolution, we power its sustainability revolution.

To chart a cleaner future, we must first understand the AI value chain i.e., the stages where environmental impacts accumulate and where interventions matter most. This includes from model design and hardware to the infrastructures that host them, to the energy systems that power them.





For this report, we focus on two intertwined levers on which the future of cleaner AI rests:

Technology innovation :

building leaner models and more efficient hardware.

Infrastructure reinvention :

reimagining the data centers where AI lives and grows.

3.2 Lever 1: Technology in Service of Sustainability

Why Technology Matters?

Every leap in AI capability, from GPT-2 to GPT-4, from narrow image recognition to multimodal reasoning, has been powered by increasing computational intensity. But higher is not always better, and it is rarely sustainable. With the right design choices, AI can become dramatically more efficient without sacrificing performance.

Energy-Efficient Model Architectures

Promising techniques are already reshaping how models are trained and deployed:

- **Pruning:** removing redundant parameters to shrink model size.
- **Quantization:** reducing precision to lower compute requirements.
- **Distillation:** training smaller models from larger ones.

Early results are encouraging: pruning and quantization together can reduce training intensity by up to 50%. One cloud provider found that quantization alone lowered inference costs by nearly 40% across production models, with no measurable loss in quality. Scaled globally, such innovations could save terawatt-hours of electricity each year.

Smarter Hardware: The Next Generation of Chips

Alongside model design, hardware innovation is redefining performance-per-watt. Smarter chips are not just faster; they are designed for sustainability:

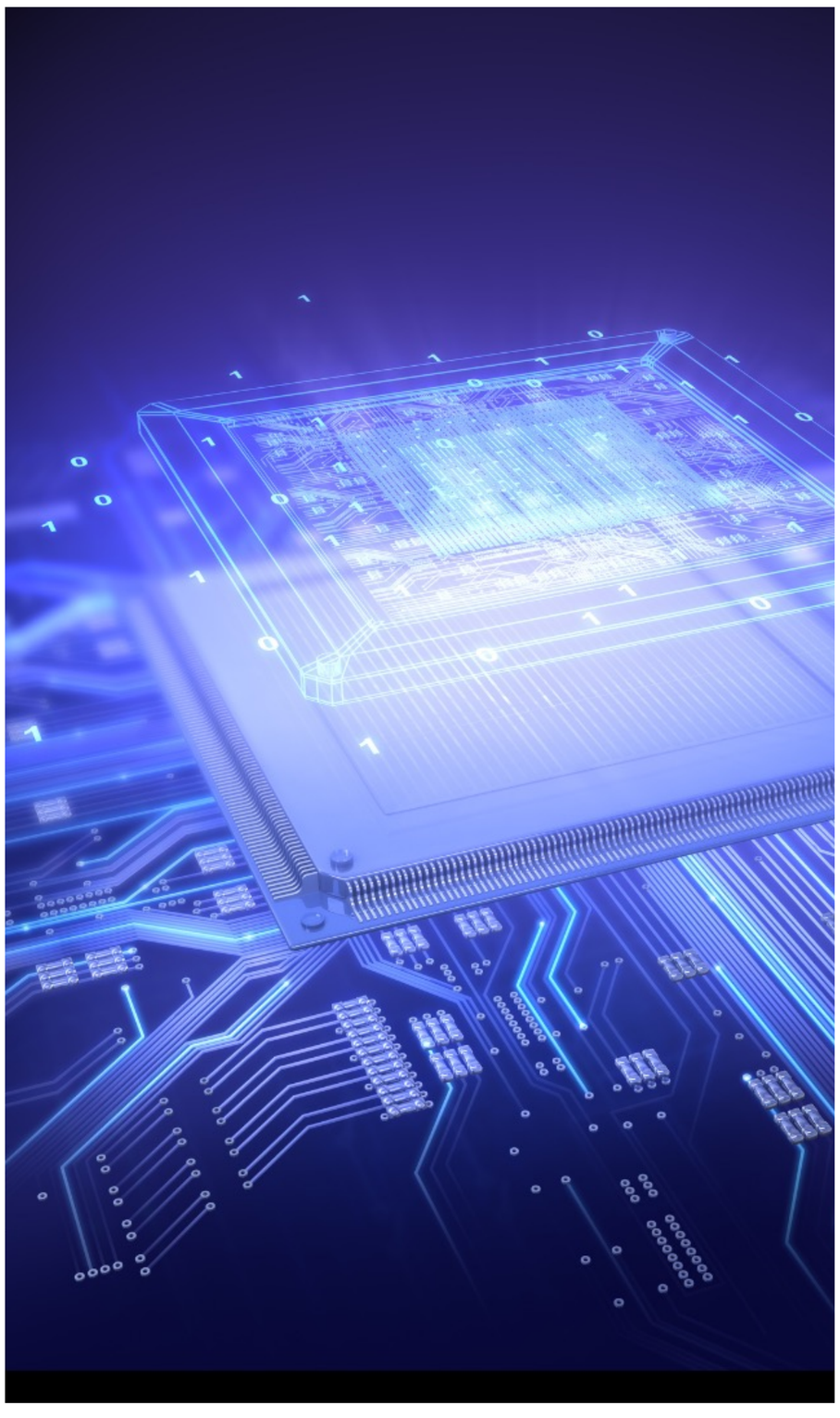
- Reducing the energy required per operation.
- Lowering waste heat, which shrinks cooling demands.
- Extending usable lifespans through modular design and recyclability.

Smarter hardware can lighten AI's load on the grid while aligning with circular economy principles, turning the backbone of AI into part of the sustainability solution.

Frontier Energy Horizons

Looking further ahead, technology may also reshape where AI's power comes from. Experiments in nuclear fusion, sometimes called "artificial suns", point to the possibility of virtually limitless clean energy. While still years away from commercialization, they signal the kind of long-term horizon worth planning toward: AI systems directly powered by abundant, carbon-free energy.

In the meantime, creative strategies can help: aligning model training with periods of renewable oversupply (such as midday solar peaks in sunny regions) could reduce fossil dependency without slowing innovation.

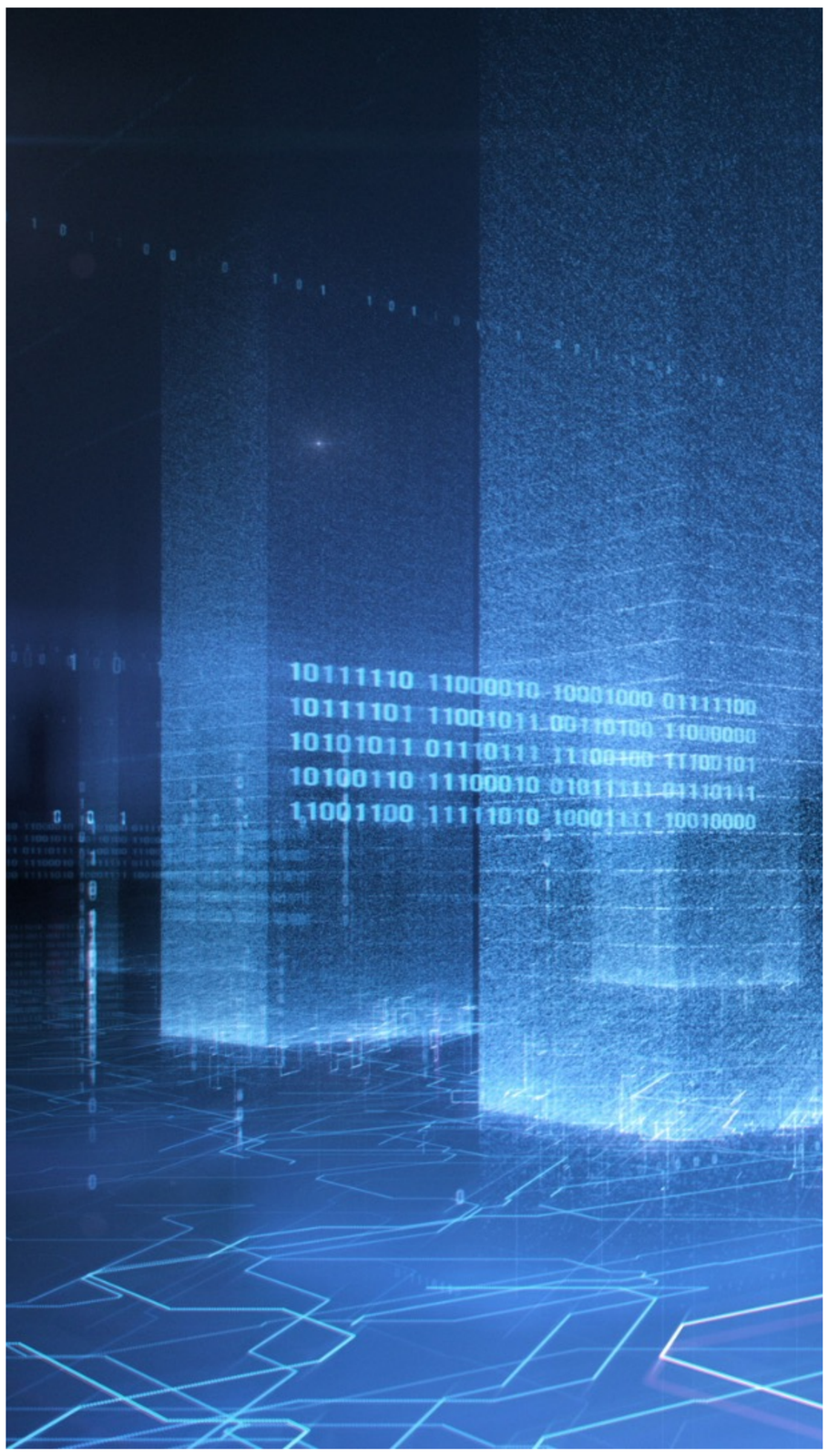


3.3 Lever 2: Rethinking Infrastructure: A New Standard for Data Centers

Why Infrastructure Matters

If AI is the brain of the digital revolution, then data centers are its body. They house the models, drive the computations, and consume a vast share of electricity and water. Yet the way we design, build, and operate these facilities often reflects assumptions from an earlier era , when efficiency meant simply keeping the servers running. In today's climate-constrained world, that standard is no longer enough. The next frontier lies in redesigning the spaces where AI lives so that growth does not come at the cost of planetary stability.

This transformation begins with rethinking four dimensions: cooling, energy reuse, siting, and integration with local ecosystems.



Liquid Cooling: Moving beyond Air and Water Cooling

The majority of today's data centers still rely on traditional cooling methods:

- Air cooling with fans and HVAC systems.
- Chilled water systems where water is pumped through cooling towers to absorb and dissipate heat.

While effective, these methods have clear limits:

- They require millions of liters of freshwater annually per large facility.
- They are energy-intensive, especially as AI workloads push computational density higher.
- They are increasingly unsustainable in water-stressed regions, where community needs must take priority.

These constraints mean that business-as-usual cooling cannot scale in parallel with AI's growth.

Liquid cooling, particularly immersion cooling, is emerging as a transformative solution. Instead of forcing air or water around hot components, servers are submerged directly in thermally conductive dielectric fluids that absorb and dissipate heat far more efficiently. Direct-to-chip cold-plate solutions are also maturing, offering alternative pathways for liquid cooling integration.

Energy Reuse: Turning Waste into Resource

The next generation of data centers will not only use less — they will do more with what they already produce. Servers generate huge amounts of heat, which is often wasted. With the right systems in place, this "waste heat" can become a resource:

- District heating for nearby residential and commercial buildings.
- Industrial processes such as drying, chemical production, or food processing.
- Greenhouse farming, where steady low-grade heat can support year-round food production.

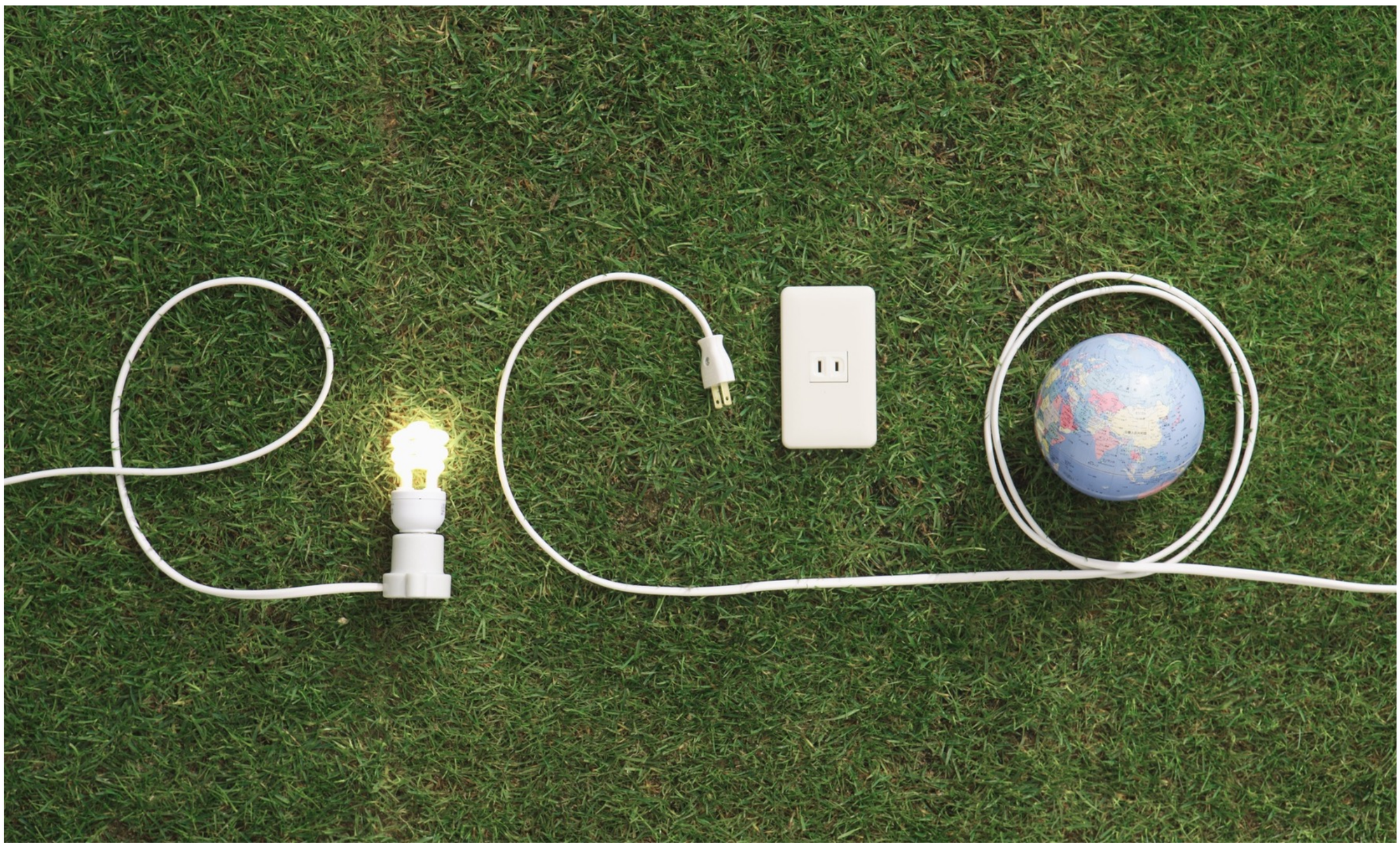
In Northern Europe, several facilities already pipe server heat into municipal grids. Extending such practices to AI data centers could dramatically improve lifecycle sustainability.

Smarter Siting and Load Balancing

A sustainable AI ecosystem requires rethinking not just how data centers operate, but where. Data centers do not need to be clustered in urban areas; they need to be close to abundant renewable energy sources.

- Siting near hydro, solar, or wind hubs can decouple growth from fossil fuels.
- Geographical load balancing allows workloads to "follow the sun" , running energy-intensive tasks in time zones where renewable supply is peaking.
- Microgrid integration can allow facilities to run partially independent of national grids, reducing stress on local utilities.

This approach could transform data centers from liabilities for grids and communities into enablers of renewable energy expansion.



From Burden to Backbone

Rethinking data centers is not just about reducing environmental harm. Done well, it positions AI infrastructure as a backbone of resilience: facilities that recycle heat, minimize water use, align with renewables, and integrate into local economies. Companies like Submer show that the technology is ready- what remains is scaling adoption and embedding these solutions into global standards.

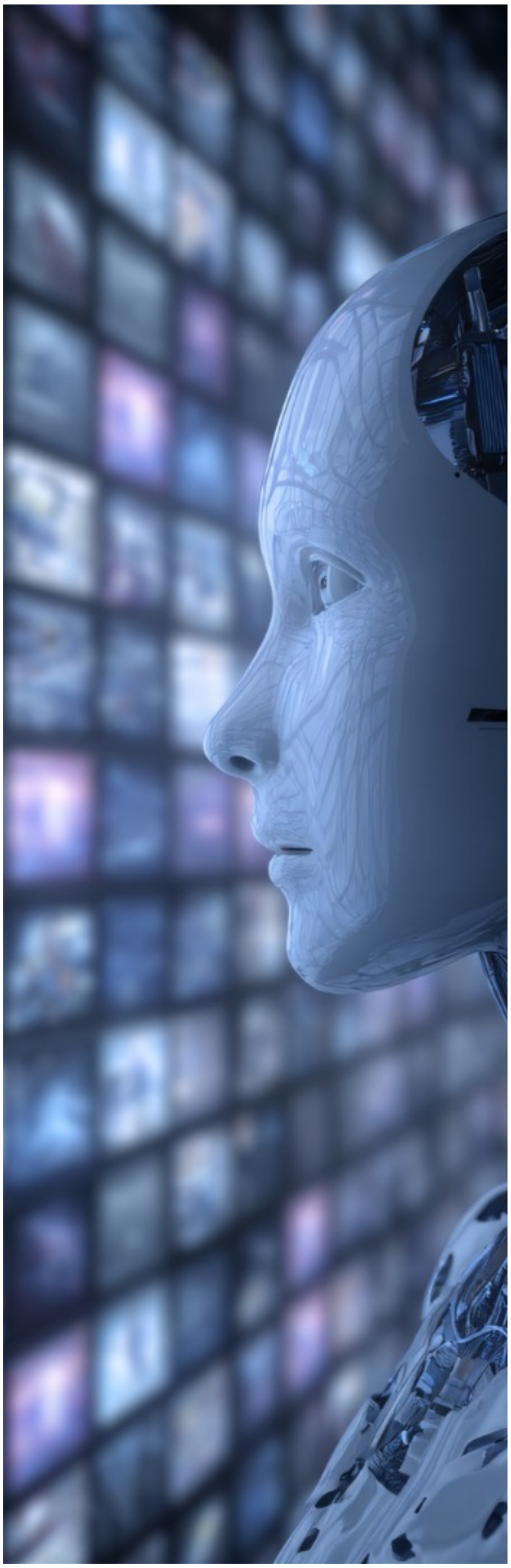
3.4 The Imperative of Strengthening the R&D Ecosystem for Data Center Innovation

AI's growing environmental impact threatens global sustainability but also offers opportunities for innovation. Addressing this challenge requires more than urgent technological and infrastructure interventions. A long-term, systemic solution is needed, with renewed emphasis on strengthening the R&D ecosystem. The future of AI and data centers depends on moving beyond off-the-shelf solutions toward a proactive, innovative strategy.

These pillars provide a focused framework that links technological, infrastructural, and collaborative strategies.

◉ Rethinking Data Centers as a Regenerative Asset

Next-generation data centers need to do more than just improve efficiency. The focus should be on creating facilities that are not only less resource-intensive but are also regenerative. This means developing technologies that reuse waste heat,



recycle water, and are designed for a circular economy of hardware. For instance, Submerged liquid cooling technology, is a perfect example and outcome of this R&D.

◉ From Idea to Scalable Solution

Successfully scaling these solutions requires more than just a good idea; it demands the ability to innovate across the entire value chain. Companies must take technologies from concept to deployment, providing end-to-end solutions. This involves offering OEM integration and Design & Build services that turn a conceptual solution into a tangible infrastructure asset. By offering these capabilities, they break down the technical and logistical barriers that often hinder the clean energy transition.

◉ Collaborative Ecosystems, Not Silos

No single player can solve AI's environmental footprint alone. A resilient R&D ecosystem depends on deep collaboration across disciplines, including:

- ✓ AI developers who align training with renewable energy cycles.
- ✓ Hardware manufacturers who design modular, recyclable systems.
- ✓ Energy providers who integrate data centers into renewable grids.
- ✓ Policymakers who enable pilot projects and set standards that reward sustainability.

By embedding sustainability into research agendas and building collaborative innovation pipelines, R&D becomes the engine of AI's sustainability revolution, rather than just a support function.

Case Study: Submer: From Cooling Technology to End-to-End Solutions

Submer is a European cleantech company that has emerged as a leader in rethinking data center design for the AI age. The company positions itself as a strategic partner, helping organizations make both the construction and operation of data centers more sustainable and efficient. Its mission is to transform data centers from resource-hungry facilities into regenerative, climate-aligned assets. It does so through solutions, including platforms, APIs, processes and facilities, enabling hyperscalers, colocation providers, and large industries to reach new heights of efficiency and innovation.

1. The Problem Submer is Solving

Data centers face a triple challenge:

- **Energy demand** from high-density AI clusters pushes cooling beyond traditional limits.
- **Water dependency** makes existing systems unsustainable, particularly in arid or water-stressed geographies.
- **Escalating costs** of electricity, cooling, and equipment replacement undermine long-term competitiveness.

Without intervention, these pressures risk locking the sector into a high-carbon, high-water trajectory that undermines both net-zero commitments and social license to operate.

2. Submer’s Technology Suite

At the core of Submer’s offering is a portfolio of liquid cooling technologies and scalable solutions that allow operators to radically improve efficiency while future-proofing infrastructure.

Immersion Cooling

- Servers are submerged in biodegradable dielectric fluids that transfer heat up to 1,000 times more efficiently than air.
- Power Usage Effectiveness (PUE), the industry’s key efficiency metric, can be reduced to as low as 1.03, which is far below the global average of ~1.6
- Massive water savings of up to 95% compared to conventional systems.
- Higher energy efficiency with 20- 40% reductions in power use depending on workload.
- Density gains i.e., enabling far more compute in less space, reducing land and building needs.
- Extends hardware lifespans as stable thermal conditions reduce wear on chips and servers.

Direct-to-Chip Cooling

- Liquid delivered directly to high-performance chips.
- Enables efficient heat extraction while maintaining compatibility with OEM standards.

SmartPod Solutions

- Compact, stackable, Tier III and Tier IV compatible modules that deliver 50–100 kW capacity per unit.
- Deployable in greenfield or retrofit projects, without the need for raised floors or cold aisles.
- Can be rapidly installed in raw spaces, accelerating deployment timelines.

Submer Cloud & Management Tools

- Remote monitoring and optimization interfaces.
- Integration with existing DCIM (Data Center Infrastructure Management) systems.
- Predictive analytics for efficiency gains and maintenance planning.

Together, these technologies form a flexible suite that can be tailored to the needs of hyperscale operators, colocation facilities, or edge deployments.

3. From Product to Platform: Submer as a Full-Stack Solution

Unlike many technology providers, Submer emphasizes that cooling is not a standalone product but part of a full-stack infrastructure transformation. Its advisory services extend beyond hardware into **design, build, retrofit, and lifecycle integration:**

Design & Build

- They provide end-to-end guidance from siting and architectural design to deployment. By offering comprehensive design and construction services, they transform their technology from a theoretical concept into a fully deployed infrastructure.
- This approach simplifies the path to adoption for clients and supports them in bypassing traditional cooling assumptions and embedding liquid-first approaches from the outset.

Retrofit Pathways

- Submer's modular pods allow existing facilities to be upgraded without massive structural changes.
- This lowers capex and provides a transitional pathway for operators not yet ready for full rebuilds.

Lifecycle & Circularity

- Hardware lasts longer under immersion conditions, reducing e-waste.
- Fluids are biodegradable and recyclable.
- Enables direct heat reuse, where warm fluid is cycled into secondary applications such as district heating, aquaculture, or industrial processes. Heat reuse potential creates revenue streams and community value.

By combining water savings, energy efficiency, hardware longevity and heat reuse, Submer illustrates how the AI industry can break free from the old cooling paradigm. In this sense, Submer acts as a strategic partner – helping operators future-proof infrastructure against both climate risks and regulatory pressures.

4. AI Advisory: Guiding Decisions in the AI Era

As AI reshapes data center requirements, operators face tough choices: how to handle exponential compute demand, where to locate new facilities, and how to align with sustainability goals.

Submer helps answer these questions by providing strategic guidance on:

- Planning workloads for high-density AI clusters.
- Choosing the right cooling approach for GPUs and AI-specific hardware.
- Siting decisions based on power, water, and connectivity.
- Building scalability roadmaps that move from pilot to hyperscale.
- Aligning with carbon, water, and circularity targets.

This advisory expertise works as a trusted guide for clients navigating the twin challenges of AI growth and sustainability.

04

From Awareness to Action

Artificial intelligence is no longer a niche technology. It is becoming central to economies, industries, and societies. Yet its environmental footprint is large and growing, from the energy used in training models, to the water consumed in data center cooling, to the waste created when hardware reaches the end of its life. If these impacts continue unchecked, AI will deepen existing climate and resource challenges.

This report has shown that the environmental cost of AI is not inevitable. With better design of models and hardware, with new approaches to data center cooling and siting, with integration of renewable energy and recycling systems, AI can be scaled in a more sustainable way. The tools and solutions are available; what is needed now is the will to implement them.

A cleaner AI future will depend on action at multiple levels: companies adopting efficiency and transparency as defaults; governments setting stronger environmental standards for digital infrastructure; and researchers advancing innovation in chips, cooling, and renewable integration. Collaboration across these groups is essential, because no single actor can solve the challenge alone.

The decisions made in the next few years will shape whether AI becomes another driver of environmental stress, or an example of how technology can grow within planetary limits. Building sustainability into AI is not a choice for later - it is an immediate priority.

Further, India's ambition to become a global digital hub under Viksit Bharat 2047 makes the redesign of data centers particularly important. While international best practices offer inspiration, the solutions must be tailored to India's geography, energy mix, and socio-economic priorities. Four pathways stand out:

1. Siting with Renewables in Mind

- India is expanding renewable capacity rapidly, especially in solar and wind. Data centers could be strategically located near mega solar parks in Rajasthan, Gujarat, and Madhya Pradesh, or wind corridors in Tamil Nadu and Karnataka.

- Linking new data centers with dedicated renewable purchase agreements (PPAs) would reduce carbon intensity while creating anchor demand for clean energy projects.

2. Integrating Waste Heat into Urban Systems

- India's growing smart city programs create opportunities to repurpose server heat for district cooling in IT hubs like Hyderabad, Bengaluru, and Pune, or for urban housing developments.
- Pilot projects could demonstrate how server-to-grid heat loops reduce both urban energy demand and data center emissions

3. Water-Conscious Cooling Technologies

- In water-stressed regions, immersion cooling technologies such as those pioneered by Submer could prevent conflict between digital infrastructure and local agriculture.
- Policy incentives could encourage operators to adopt non-water-intensive cooling as a default, especially in arid states.

4. Retrofitting Existing IT Clusters

- India's data ecosystem is concentrated in established hubs like Mumbai, Chennai, and Delhi NCR. Rather than only building new "green" facilities, retrofits using immersion cooling and modular energy-efficient designs could yield significant near-term savings.
- Retrofitting also supports India's circular economy agenda, by extending hardware lifespans and reducing e-waste flows.

The Opportunity for Leadership

For India, rethinking AI infrastructure is not just a defensive move to contain environmental costs. It is also a chance to demonstrate leadership by developing climate-aligned digital standards that others can emulate. With the right mix of technology adoption, renewable integration, and policy incentives, India could set a benchmark: AI infrastructure that drives innovation while reinforcing the country's commitments to sustainability, equity, and resilience.

References

1. https://www.tomshardware.com/desktops/servers/a-single-modern-ai-gpu-consumes-up-to-37-mwh-of-power-per-year-gpus-sold-last-year-alone-consume-more-power-than-13-million-households
2. https://brightlio.com/data-center-stats/
3. https://brightlio.com/data-center-stats/
4. https://www.statista.com/statistics/1228433/data-centers-worldwide-by-country/?srsltid=AfmBOopKxjFeW_ahZQHA0zZzkpZ-AP1XnQF1510D6STj73-SZrGba8X7
5. https://www.datacentermap.com/india/
6. https://www.iea.org/reports/energy-and-ai/understanding-the-energy-ai-nexus
7. https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117
8. https://www.technologyreview.com/2025/08/21/1122288/google-gemini-ai-energy/
9. https://www.carbonbrief.org/ai-five-charts-that-put-data-centre-energy-use-and-emissions-into-context/
10. https://ieefa.org/resources/indias-power-hungry-data-centre-sector-crossroads
11. https://www.datacentermap.com/india/

Contributors

Sameer Jain
Managing Director,
Primus Partners

Pooja Lahri
Vice President,
Primus Partners

Amit Purohit
Vice President,
Primus Partners

Tavishi Darbari
Manager,
Primus Partners

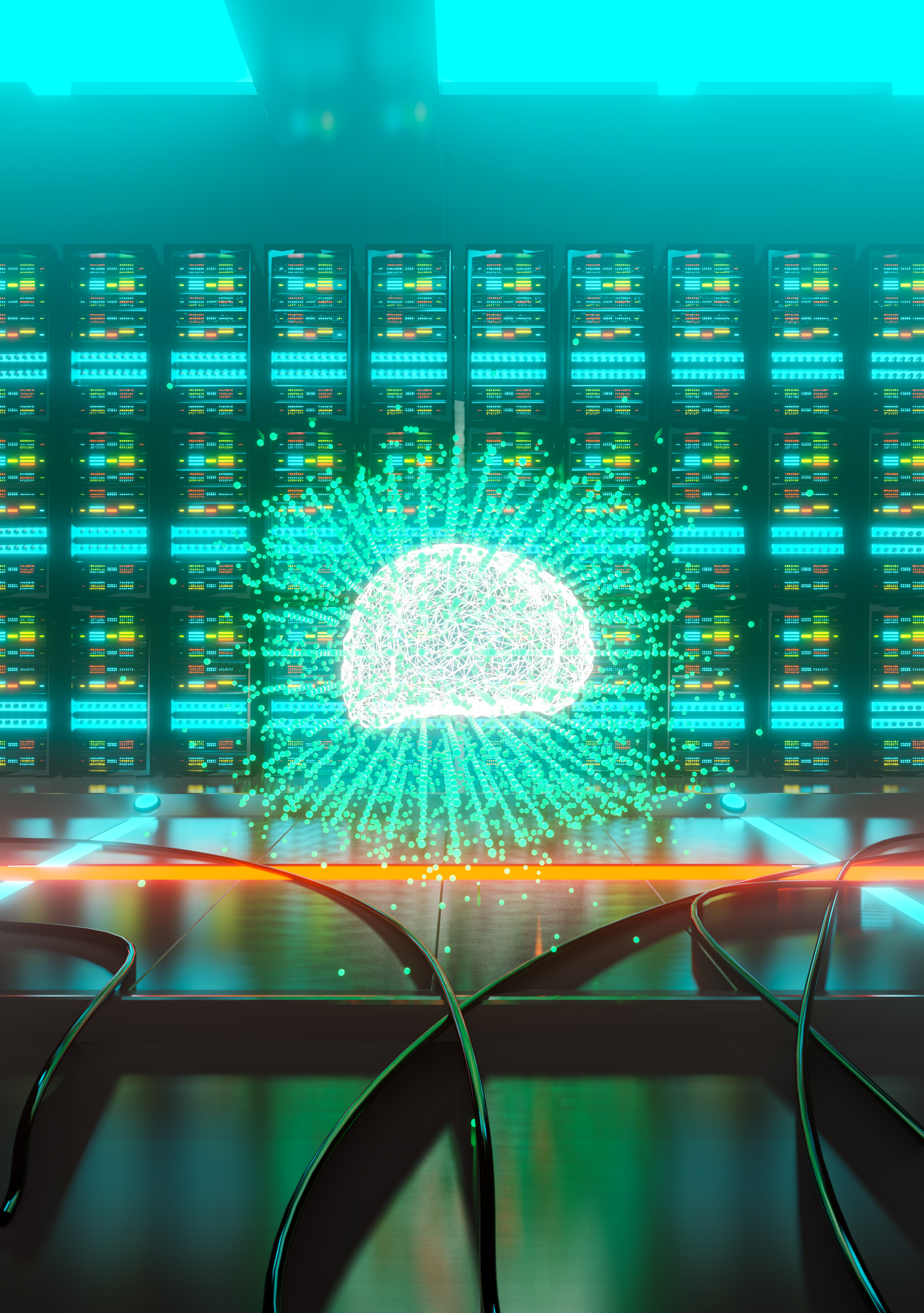
Harshpreet Kaur
Senior Consultant,
Primus Partners

Ankush Sharma
Creative Designer

Suyash Pandey
Creative Designer

Suchetta Singh
Research Assistant,
Primus Partners

Ishaan Mishra
Research Assistant,
Primus Partners



PASSION
for providing solutions to help clients achieve their goals

RESPECT
for all and alternate viewpoints

INTEGRITY
of thoughts and actions

MASTERY
of our chosen subject to drive innovative and insightful solutions

US
representing the Primus collective, where each individual matters

STEWARDSHIP
for building a better tomorrow

PRIMUS PARTNERS®

Solutions for Tomorrow

Primus Partners has been set up to partner with clients in 'navigating' India, by experts with decades of experience in doing so for large global firms. Set up on the principle of 'Idea Realization', it brings to bear 'experience in action'. 'Idea Realization'— a unique approach to examine futuristic ideas required for the growth of an organization or a sector or geography, from the perspective of assured on ground implementability.

Our core strength comes from our founding partners, who are goal-oriented, with extensive hands-on experience and subject-matter expertise, which is well recognized in the industry. Established by seasoned industry leaders with extensive experience in global organizations, Primus Partners boasts a team of over 250 consultants and additional advisors, showcasing some of the finest talent in the nation.

The firm has a presence across multiple cities in India, as well as Dubai, UAE. In addition, the firm has successfully executed projects across Africa, Asia Pacific and the Americas.

India Offices

Bengaluru
91 Springboard
Business Hub 175, 176
Bannerghatta Rd,
Dollars Colony,
Bengaluru – 560076

Chandigarh
2nd Floor, Netsmartz,
Plot No. 10, Rajiv
Gandhi Chandigarh
Technology Park,
Chandigarh – 160019

Chennai
147, Pathari Rd, Door #3,
WorkEz Hansa Building,
RK Swamy Centre,
Thousand Lights,
Chennai, TN - 600006

Delhi
1 to 7, UG Floor,
Tolstoy House,
Tolstoy Road,
Connaught Place
New Delhi - 110001

Kolkata
Siddhartha Apartments
4th Floor, 188/2,
Block J,
New Alipore,
Kolkata - 700053

Mumbai
156/157, 15th
Floor, Nariman
Bhavan, NCPA
Road, Nariman
Point,
Mumbai – 400021

International Offices

Dubai
United Arab Emirates
(UAE)

Dammam
Kingdom of Saudi Arabia
(KSA)

Washington D.C
United States of America
(USA)

www.primuspartners.in info@primuspartners.in Primus Partners India @partners_primus @primuspartners7128