Chintan Research Foundation

The Dialogue®
INFORM ENGAGE IDEATE

GREENING THE CODE

A Roadmap for Environmentally Responsible AI Innovation in India

DISCUSSION PAPER

ABOUT CHINTAN RESEARCH FOUNDATION

Chintan Research Foundation (CRF) is an emerging independent think tank dedicated to shaping policy through rigorous research and thought leadership. With a strong focus on fostering collaboration between policymakers and industry, CRF integrates practical insights into its research and advocacy efforts. It conducts comprehensive research to support informed decision-making and engages with stakeholders through discussions, events, and publications. CRF's research is focused on three core areas – Climate Change & Energy Transition, Economy & Trade, and Geopolitics & Strategic Studies. The Centre for Climate Change & Energy Transition aims to be a research and expertise hub, focusing on sustainable and clean energy, environmental stewardship, and climate actions.

For further information about CRF, please visit: crfindia.org

ABOUT THE DIALOGUE

The Dialogue is a public-policy think tank with a vision to drive a progressive narrative in India's policy discourse. Founded in 2017, we believe in facilitating well-researched policy debates at various levels to help develop a more informed citizenry, in areas around technology and development issues. With its ninth-year running, The Dialogue has focused its attention on building a credible and constructive voice in India's technology policy discourse. We are the country's leading think tank, having been recognised as one of the top 10 think tanks to watch out for by the University of Pennsylvania. The Dialogue is now recognised as a leading change maker in India's policy discourse, working and engaging closely with policy-makers, parliamentarians, government officials, and stakeholders such as the media, research community, corporates, and academia.

For further information about The Dialogue, please visit: thedialogue.org.in

Discussion Paper

GREENING THE CODE

A Roadmap for Environmentally Responsible AI Innovation in India

Chintan Research Foundation

The Dialogue®
INFORM ENGAGE IDEATE

Authors

Meemansa Agarwal
Program Manager, The Dialogue
Saharsh Panjwani
Senior Research Associate, The Dialogue
Mandar Suhasrao Joshi
Research Consultant, Centre for Climate Change and Energy Transition, Chintan Research Foundation
Nishit Abhay Patil
Research Analyst, Centre for Economy and Trade, Chintan Research Foundation
Kamesh Shekar
Associate Director, Strategy & Research, Lead, Privacy and Data Governance Vertical & AI Vertical, The Dialogue

Reviewer

Karn Kedar
Advisor, Research and Outreach, Chintan Research Foundation

CONTENTS

1. Introduction – AI as a System, Not Merely as Infrastructure

Artificial Intelligence (AI) is often spoken in terms of abstractions, algorithms, models, and intelligence at scale. It is frequently reduced further to visible infrastructure: data centres, GPU clusters, and cloud capacity. This framing, while convenient, is incomplete. AI is not a single layer of infrastructure; it is a system, deeply interconnected, materially grounded, and shaped as much by physical constraints as by computational ambition.

To view AI merely as infrastructure is to see the engine but ignore the vehicle. A systems view reveals something closer to a nervous system, spanning hardware, software, data, energy, and institutional arrangements. This perspective matters because it shifts the lens: from isolated investments in compute capacity to the broader ecosystem required to sustain and scale intelligence. It also foregrounds a critical but often underexplored reality: that the expansion of AI is inseparable from the expansion of physical infrastructure and the resources that sustain it.

This system can be understood across three interdependent layers:

- **Upstream: semiconductor design, fabrication, and advanced packaging**
  This layer underpins the entire AI stack. It includes firms designing AI chips (e.g., GPUs and TPUs), fabrication by advanced foundries, and increasingly complex packaging technologies required for high-performance computing.

- **Midstream: cloud infrastructure and data centres**
  This layer enables access to compute at scale. Hyperscale cloud providers host the data centres where AI models are trained and deployed.

- **Downstream: applications across sectors**
  This is where AI interacts with the real economy. Applications range from AI-enabled diagnostics in healthcare, to credit scoring in fintech, to precision agriculture and logistics optimisation.

Each layer is defined by distinct actors, cost structures, and challenges. Constraints in semiconductor supply chains can limit compute availability; gaps in data infrastructure can restrict access; and weak downstream adoption can blunt economic impact. The value chain operates as an integrated system, its efficiency determined by the strength of its weakest link.[1] For countries like India, this requires prioritisation across layers, rather than a singular focus on any one segment.

Despite its digital character, AI is fundamentally grounded in physical resources. Training large models and supporting real-time inference requires a continuous energy supply. Hyperscale data centres demand land, often near urban clusters. Cooling systems rely heavily on water. Underpinning all of this is specialised compute hardware, whose production involves complex and resource-intensive global supply chains. These dependencies are not incidental; they are foundational. As AI systems grow in scale and complexity, their demands on energy, land, and water intensify, creating new pressures on already constrained infrastructure systems. This is particularly relevant for emerging economies, where competing development priorities and resource limitations make these trade-offs more pronounced.

Data centres sit at the centre of this material reality. They are not simply passive repositories of data, but active sites of computation that anchor the AI ecosystem. More importantly, they function as the connective tissue of the AI value chain, linking upstream hardware capabilities with downstream applications. Data infrastructure, comprising storage systems, data pipelines, and high-speed connectivity networks, enables the collection, processing, and movement of the vast datasets that AI systems depend on.

In an era where data is often described as the new oil, data centres can be understood as the refineries of the digital economy. Their strategic importance is growing rapidly as AI workloads become more compute-intensive and latency-sensitive. Countries and companies are increasingly competing not just on access to data, but on their ability to store, process, and derive value from it efficiently.

This expansion, however, comes with significant environmental implications. Data centres accounted for approximately 1.5 per cent of global electricity consumption in 2024[2], and this demand is projected to nearly double by the end of the decade. In absolute terms, this would place their energy consumption on par with that of large industrialised economies. Water use presents an equally pressing challenge. The global data centre industry is estimated to consume hundreds of billions of litres of water annually[3], comparable to the domestic water needs of a large metropolitan population . Land use, often overlooked, is also emerging as a constraint, particularly in dense urban regions where competing demands for space drive up costs and limit expansion.

These pressures are already visible in several advanced markets. In parts of the United States and Europe, rapid data centre expansion has begun to strain local power grids and water systems, prompting regulatory scrutiny and, in some cases, temporary moratoriums on new developments.[5] [6] These experiences offer an important lesson: the costs of scaling digital infrastructure are not evenly distributed, and without careful planning, can generate significant localised impacts.

India stands at a critical juncture in this development trajectory of AI. It generates nearly one-fifth of the world's data, yet accounts for only a small share (153 out of 12,000+) of global data centre capacity.[7] This gap is likely to narrow quickly. As digital adoption deepens, AI use cases expand, and demand for domestic data infrastructure is expected to accelerate significantly. This presents a clear economic opportunity, to build capacity, attract investment, and strengthen India's position within the AI value chain. At the same time, it introduces a set of complex challenges linked to resource use, infrastructure readiness, and spatial planning.

Governments globally are deploying fiscal incentives, tax concessions, subsidised land, preferential power tariffs, and infrastructure support, to attract data centre investments. India is no exception. While a comprehensive national policy framework for data centres has yet to be formalised, a series of enabling measures have been introduced. These include granting infrastructure status to large data centre projects[8] and announcing long-term tax incentives[9] for investments in data centre and cloud infrastructure. In the absence of a binding national framework, state governments have taken the lead. Currently, 15 states support data centres through dedicated policies or through IT/ITeS and industrial policies.[10]

While these measures have catalysed growth, they also raise important questions about alignment and sustainability. Incentive structures, if not carefully designed, risk prioritising short-term capacity expansion over long-term resilience. They may not fully account for local resource constraints, particularly in regions already facing water stress or energy shortages. Moreover, they can reinforce existing patterns of geographic concentration, leading to congestion and compounding environmental pressures in a limited number of urban hubs.

The scale and strategic importance of data centre infrastructure has, therefore, transformed it into a core policy concern. It now sits at the intersection of industrial policy, energy planning, urban development, and environmental governance. Decisions taken today on where to locate facilities, how to power them, what cooling technologies to adopt, and how to integrate them with local infrastructure will have long-term consequences. These are path-dependent choices that will shape resource use patterns, optimal life cycle of AI related infrasturcture, and environmental outcomes for decades.

For India, the challenge is particularly acute. The country faces existing constraints in energy availability, water resources, and urban infrastructure. At the same time, it is poised for rapid growth in data generation and AI adoption. This combination makes it imperative to integrate sustainability and resilience into infrastructure planning from the outset, rather than as an afterthought.

This discussion paper is situated within this broader context. It seeks to move beyond a narrow

focus on infrastructure expansion to its associated environmental and policy implications. Specifically, the paper aims to: assess India's current trajectory in infrastructure expansion; identify emerging stress points and structural constraints; and explore alternative approaches to scaling AI infrastructure in a more resource-efficient and sustainable manner.

In doing so, the paper brings together supply-side considerations such as energy, land, water, and connectivity requirements with their implications for downstream AI adoption. It also engages with the evolving policy landscape, highlighting areas where greater coordination, clearer frameworks, and forward-looking planning will be required.

As India moves to scale its AI capabilities, the question is no longer whether infrastructure will expand, but how it will do so. The choices made at this stage on design, location, and governance will determine whether this expansion is merely rapid or whether it is also sustainable, resilient, and aligned with broader development priorities.

2. The AI Value Chain in India – Structure, Gaps, and Strategic Priorities

2.1. Semiconductor Design, Fabrication, and Global Supply Dependencies

India's integration into the global AI value chain depends on its capacity to transition from a consumer of microelectronics to domestic architect of semiconductor technologies, a strategic pivot catalysed by the evolution of the India Semiconductor Mission (ISM).[11] Initially capitalised with a ₹76,000 crore outlay, ISM 1.0 laid the essential groundwork for a domestic ecosystem by approving ten major projects across six States, drawing cumulative investments exceeding ₹1.6 lakh crore by late 2025 to establish assembly, testing, marking, and packaging and outsourced semiconductor assembly and test facilities capable of producing millions of chips daily, including major ventures by Tata Electronics and Micron. However, the domestic market, which is projected to expand from $38 billion in 2023 to an estimated $100-$110 billion by 2030, currently consumes nearly 20% of global microprocessor output while remaining reliant on imports for over 90% of critical inputs such as high purity chemicals, neon, specialty gases, photoresists, and silicon wafers from concentrated geopolitical nodes like China, Taiwan, and South Korea.[12] Recognising that this supply chain poses an acute risk to strategic sectors such as 5G, defence, and high-performance computing, the Union Budget 2026-27 formally introduced ISM 2.0, pivoting the national strategy towards full-stack creation and upstream material autonomy. This second phase focuses explicitly on establishing domestic capabilities in semiconductor equipment and materials, fortifying global supply chain linkages, and accelerating the development of indigenous intellectual property, aiming to establish a roadmap for 3-nanometre and 2-nanometre technology nodes (lower power consumption chips with faster processing times) while advancing proprietary architectures for indigenous microprocessor (such as DHRUV64). Concurrently, geographical concentration of firms is materialising through highly specialised regional corridors, notably the Chennai-Hyderabad belt for operational equipment manufacturing and the Noida-Delhi route for advanced design, which are complemented by the Design Linked Incentive scheme that currently supports 24 semiconductor design startups advancing to foundry nodes as low as 12 nm.[13] Ultimately, this trajectory indicates a sophisticated structural recognition that authentic digital autonomy in the AI era cannot be achieved through software scale alone, but requires an upstream command over the physical materials, fabrication equipment, and proprietary silicon architectures that constitute the bedrock of computational power, positioning India to manufacture chips for 75% of its domestic applications by 2029.[14]

2.2. Compute Infrastructure – Hyperscale, Enterprise, Public, and Edge

The physical foundation of India's AI economy is undergoing an exponential and dual-tracked structural change, characterised by the simultaneous concentration of hyperscale gigawatt facilities and the rapid decentralisation of edge computing networks to support an installed data centre capacity that has tripled since 2020 to 1.5 GW, with projections reaching 6.5 GW by 2030.[15] At the centralised nexus, hyperscale compute infrastructure is being scaled through strategic partnerships with global silicon leaders. For instance, the integration of over 20,000 NVIDIA Blackwell Ultra GPUs into Yotta's Shakti Cloud across Navi Mumbai and Greater Noida, alongside Larsen & Toubro's development of gigawatt scale, sovereign AI factories under the IndiaAI Mission, exemplifies the massive capital deployment aimed at retaining highly complex, dense foundational model training within sovereign borders.[16] However, the centralised cloud model is increasingly strained by power constraints, land limitations, and inherent network latency, slowing the critical pivot towards distributed edge infrastructure, a market segment in India projected to grow at a robust 14.6% compound annual growth rate from $10.48 billion in 2025 to $27.2 billion by 2032.[17] But, this is also a functional necessity driven by the proliferation of 5G networks and the demands of agentic AI applications (autonomous mobility, industrial robotics, etc.) which require sub-millisecond responsiveness (often <10ms) that distant metropolitan data centres cannot physically provide due to data transit times.[18] However, by pushing computational nodes to Tier-2 and Tier-3 cities, edge architecture significantly reduces wide area network bandwidth loads, and accelerates digital inclusion across India's vast demographic landscape, enabling localised compute for the hundreds of millions of users driving digital payments, and e-commerce beyond traditional IT hubs. This bifurcation suggests an infrastructural evolution of centralised gigawatt factories with them acting as the heavy-duty neural engines for model training and complex enterprise workloads, while decentralised web of edge micro-datacentres serves as the real time, low latency sensory network, effectively blanketing the subcontinent in a high performance computing fabric that is essential to unlocking the full economic multiplier of applied Artificial Intelligence.[19]

2.3. Cloud Platforms and AI Model Development Ecosystems

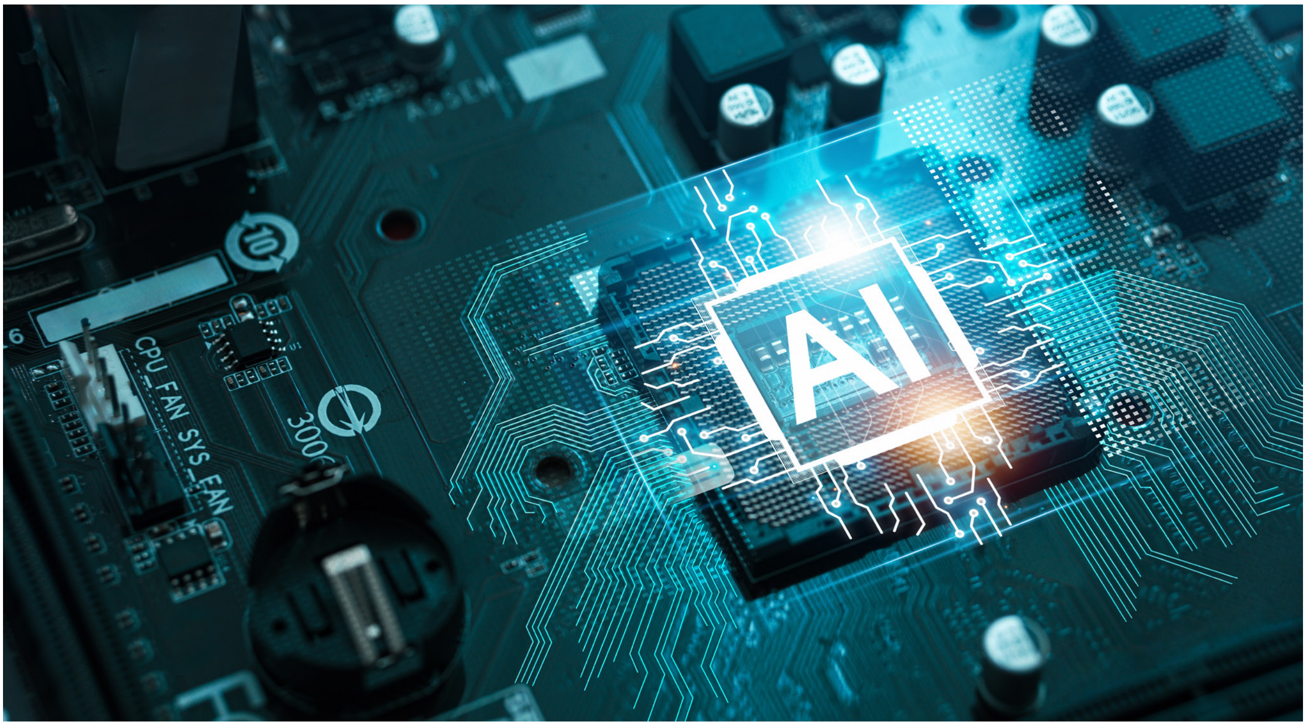
India's AI model development ecosystem has rapidly evolved from an experimental phase into a structurally robust sovereign software layer, evidenced by a 3.6x growth in the generative AI startup base to over 240 enterprises between the first half of 2023 and the first half of 2024, alongside the deployment of more than 17 indigenous vertical and Indic language models.[20] This proliferation is anchored by strategic, vernacular focused initiatives such as Sarvam AI's OpenHathi (recognised as the first publicly available Hindi large language model), Krutrim, and the government backed Bhashini project, which collectively aim to dismantle linguistic barriers and ensure that AI capabilities are deeply contextualised for India's diverse digital populace rather than remaining reliant on culturally homogenous Western algorithmic architectures.[21] To physically anchor the global cloud ecosystems that host these models, the Union Budget 2026-27 introduced a tax holiday until 2047 for global cloud service providers (discussed in detail in further chapters), Yet, despite these infrastructural and policy advances, a critical structural gap persists within the capital markets in terms of a modest 1.25x increase in GenAI startup funding, which remains largely restricted to early stage seed rounds (Also seen as a deficit of mid-to-late stage venture capital). Consequently, while India possesses unparalleled data diversity and a strategic fiscal framework that successfully anchors global cloud infrastructure, the long-term competitiveness of its AI model ecosystem and its ability to transition from pilot programs to globally scalable, high-impact enterprise solutions will depend intrinsically on cultivating domestic financial mechanisms capable of sustaining the protracted, capital-heavy lifecycle of foundational model training against international competition.

2.4. Startups, MSMEs, and Applied AI Use Cases

The application layer of India's AI value chain shows that while consumer adoption is globally leading (evidenced by over 100 million weekly users of AI chatbots), the integration of these technologies into the Micro, Small, and Medium Enterprises sector, which contributes roughly 30% to the national GDP and 45% of exports, remains constrained. Despite the transformative potential of AI to elevate productivity and export readiness, recent industry surveys indicate that merely 7% of Indian MSMEs are actively exploring AI-powered tools, with 36% expressing resistance due to a lack of digital readiness, data infrastructure deficits, and the prohibitive operational costs associated with implementing high-end AI solutions.[22] To bridge this systemic gap, the national strategy relies heavily on the paradigm of "frugal AI", meaning the development of highly contextual, low-cost, and impact-driven applications that bypass the need for massive on-premise compute by leveraging India's Digital Public Infrastructure.[23] Strategic policy frameworks, such as those outlined by NITI Aayog, advocate for the establishment of shared "landing zones" for GPUs and the deployment of pre-trained "utility agents" accessible via pay-per-use models, which would democratise access to enterprise-grade AI for smaller manufacturing and financial entities without requiring prohibitive capital expenditures.[24] Cultivating specialised AI service startups to provide predictive maintenance, quality analytics, and supply chain optimisation to traditional manufacturing units is also identified as a pivotal mechanism to modernise the industrial base. Ultimately, moving the MSME sector from being underprepared to scalable adoption relies on this targeted, frugal approach, transforming AI into an accessible, utility driver of grassroots industrial competitiveness across the Global South.

2.5. Public Sector AI and National Compute Goals

The Indian government has positioned itself as both the primary architect of a democratised AI ecosystem through the comprehensive IndiaAI Mission. Backed by a ₹10,371 crore financial outlay spanning five years, the mission's cornerstone is the Compute Capacity pillar, which has the target of deployment of a scalable infrastructure comprising over 38,000 graphics processing units, which is a substantial elevation from the initial 10,000 unit goal facilitated through robust Public-Private Partnerships. This massive hardware expansion is intrinsically linked to the creation of comprehensive digital public goods, notably the AIRAWAT compute platform and the AIKosh datasets platform, the latter serving as an interoperable, consent-based national repository that provides developers with critical non-personal data, sandbox environments, and sovereign foundational models required to train culturally relevant AI systems.[25] Furthermore, the State is heavily driving AI adoption for large-scale socioeconomic transformation through initiatives like the IndiaAI Innovation Challenge 2026. By partnering with state governments, this programme solicits scalable, deployment-ready AI solutions to resolve structural bottlenecks in urban infrastructure, healthcare, education, and last-mile service delivery.[26] To ensure these technological leaps do not exacerbate the digital divide, the FutureSkills pillar is concurrently financing the establishment of Data and AI Labs across Tier-2 and Tier-3 cities,



systematically aiming to upskill a projected 50 lakh students and professionals. By heavily subsidising compute access by offering up to a 40% reduction in operational costs for verified startups and academia, and structuring datasets as digital public goods, the public sector is effectively pre-empting the monopolisation of AI infrastructure, ensuring that national compute goals are uniformly aligned with digital sovereignty, grassroots innovation, and equitable societal advancement.

2.6. Implications for India's Energy Security and Decarbonisation Pathway

In India, data centre demand is expected to rise from the current 0.8% of total electricity use to about 2.6% by 2030[27] and further to around 6% by 2050. At the edge, energy use is more distributed but rising rapidly due to real-time inference across devices, telecom networks, and industrial applications.[28] While per-query energy consumption remains small, aggregate demand is significant, particularly as AI-enabled services scale across sectors.[29] For India, the expansion of AI infrastructure presents both risks and opportunities for energy security and decarbonisation. On the risk side, concentrated growth in data centres could intensify peak electricity demand, strain urban grids, and increase reliance on fossil fuel imports if not aligned with clean energy expansion. However, India's growing renewable energy capacity, particularly solar, offers an opportunity to align AI-driven demand with low-carbon supply. Strategic siting of data centres in renewable-rich states such as Gujarat, Rajasthan, and Tamil Nadu can reduce transmission pressures and facilitate direct integration with clean energy sources. There is also scope for system innovation. Waste heat recovery in industrial clusters, integration with green hydrogen production, and demand-side management (including time-of-day pricing for AI workloads) could enhance overall system efficiency. Policy design will play a critical role. Moving beyond conventional renewable purchase obligations toward real-time clean energy matching, grid flexibility incentives, and efficiency standards for AI infrastructure could position India to harness AI as a driver of clean energy investment rather than a source of additional carbon-intensive demand.

The feasibility of such models is reinforced by global practice, where larger cloud operators have increasingly co-located infrastructure with renewable energy assets and optimised workloads to align with energy availability. These approaches suggest that compute demand can, to an extent, be shaped in response to energy system conditions rather than treated as a purely fixed load.

3. Way Forward

3.1. From Centralised Hubs to Distributed AI Infrastructure

The rapid expansion of AI has brought renewed attention to the architecture of compute infrastructure. While recent discourse has often centred on the scaling of graphics processing units ('GPUs'), especially for training large models, the broader trajectory of computing suggests a shift toward heterogeneous environments that combine multiple types of processors optimised for different tasks.

Historically, general-purpose central processing units ('CPUs') formed the backbone of computing systems. The rise of GPUs has marked a significant shift in computing, enabling parallel processing at scale and accelerating machine learning workloads. More recently, specialised accelerators such as neural processing units ('NPUs'), tensor processing units ('TPUs'), and field-programmable gate arrays ('FPGAs') have emerged, each designed to optimise specific computational patterns.

This evolution reflects a fundamental principle in that no single architecture is optimal for all types of workloads. Training large foundation models requires high-throughput, parallel computing, often delivered by GPUs or specialised accelerators. Inference workloads, particularly those deployed at scale or at the edge, may be more efficiently handled by CPUs or NPUs. Similarly, workloads in sectors such as telecommunications, finance, or industrial automation may require customised configurations that balance performance, cost, and energy consumption. To buttress this further, we see literature which notes that improvements in compute efficiency increasingly depend on both hardware specialisation and software optimisation, rather than uniform scaling of a single architecture.[30] Industry analyses similarly indicate that heterogeneous systems can deliver significant gains in performance per watt, a critical metric as energy constraints become more binding.[31]

From the Indian context, this shift is of great relevance. India's AI ecosystem is characterised by diverse use cases, ranging from large-scale model development to deployment in resource-constrained environments such as rural healthcare, agriculture, and public service delivery. A homogeneous compute strategy in all likelihood risks over-provisioning for some applications while under-serving others. Heterogeneous compute, by contrast, offers a pathway to align infrastructure more closely with the specific needs of different sectors and regions.

3.1.1. Edge Data Centres and Latency-Sensitive AI Applications

Edge data centres are a core part of this emerging architecture. Unlike hyperscale facilities, which are designed for large-scale aggregation of compute and storage, edge data centres are relatively smaller and geographically distributed facilities located closer to end users or data sources.

The primary value proposition of edge infrastructure lies in latency reduction. Certain applications, such as autonomous systems, real-time video analytics, industrial automation, and telemedicine, require almost immediate response times, which can be measured in milliseconds. Routing data to distant centralised facilities adds to delays that can compromise functionality and safety. However, by processing data closer to the source, edge data centres enable faster response times and reduce dependence on long-haul network connectivity. In addition to latency, edge deployment can reduce network congestion. As data volumes increase, transmitting all data to centralised facilities becomes inefficient and costly, and edge processing allows for pre-processing, filtering, and local decision-making, reducing the volume of data that needs to be transmitted upstream.

Support for this is evident from a global standpoint. According to the International Data Corporation ('IDC'), spending on edge computing is expected to reach over USD 300 billion globally by 2026, driven by AI-enabled applications and IoT deployment.[32] While India's edge ecosystem remains nascent, the expansion of 5G networks and digital public infrastructure is likely to accelerate adoption.

3.1.2. Energy and Sustainability: Trade-Offs of Decentralisation

While distributed infrastructure offers benefits in terms of latency and access, it also introduces issues concerning energy and sustainability trade-offs. Larger hub-centric data centres attract the benefits from economies of scale in energy efficiency, cooling optimisation, and renewable energy procurement. Large facilities can invest in advanced cooling systems, negotiate long-term power purchase agreements, and optimise operations across large workloads. In contrast, smaller and distributed facilities may face higher per-unit costs and lower efficiency. Achieving low power usage effectiveness ('PUE') in small-scale edge environments can be challenging due to constraints in design and scale. Moreover, access to renewable energy may be more limited in certain regions, increasing reliance on grid electricity with higher carbon intensity. At the same time, decentralisation can reduce transmission losses and enable better alignment between compute demand and local energy resources. In regions with high renewable energy potential, distributed data centres could be co-located with generation assets, supporting local grid balancing and reducing reliance on long-distance transmission.

The overall sustainability impact of distributed infrastructure, therefore, depends on design choices, location decisions, and integration with energy systems. This has found support in international literature as well, stating that the energy footprint of digital infrastructure is shaped not only by efficiency improvements but also by structural shifts in demand and architecture.[33] In the Indian context, where energy systems are undergoing rapid transition, these trade-offs are particularly salient. Integrating data centre expansion with renewable energy planning, grid upgrades, and water resource management will be critical to ensuring that AI growth does not exacerbate environmental pressures.

3.1.3. Workload-Appropriate Optimisation Across AI Tasks

The concept of workload-appropriate optimisation lies at the heart of heterogeneous computing. Instead of treating computation as a uniform resource, this approach recognises that different AI tasks impose distinct computational requirements and should therefore be matched with appropriate architectures.

At a higher level, AI workloads can be divided into two primary parts i.e., training and inference. Training involves processing large datasets to develop models, often requiring high-performance clusters with specialised accelerators. In contrast, inference involves applying trained models to new data, often at scale and in real time. While training is compute-intensive, inference can dominate total compute demand over time, particularly in applications with large user bases.

Recent estimates suggest that inference workloads may account for a growing share of AI-related compute demand, driven by the proliferation of AI-enabled applications.[34] This shift has important implications for infrastructure design. Deploying GPUs for all inference tasks may be inefficient, both in terms of cost and energy consumption. Alternative architectures, including CPUs and NPUs, may offer more efficient solutions for many use cases. In India, where cost sensitivity is a critical factor for startups, MSMEs, and public institutions, workload-appropriate optimisation has significant economic implications. Over-reliance on high-cost hardware can increase barriers to entry and limit access to AI capabilities. By aligning compute choices with actual workload requirements, it is possible to reduce total cost of ownership and expand access.

3.1.4 Interoperability and Portability Across Processing Architectures

As compute environments become more heterogeneous, interoperability and portability emerge as critical design considerations. Without the ability to be flexible and move workloads seamlessly across different architectures, the benefits of heterogeneity may be offset by fragmentation and inefficiency. Interoperability refers to the ability of different systems and components to work together, while portability refers to the ease with which applications can be transferred across environments. In the context of AI infrastructure, these concepts are closely linked to software frameworks, programming models, and standardised interfaces.

At present, the AI ecosystem uses a combination of open and proprietary frameworks. We have platforms such as TensorFlow and PyTorch that provide a certain degree of abstraction, allowing developers to deploy models across different hardware environments. At the same time, hardware-specific optimisations and proprietary toolchains can create dependencies that limit portability. The risk of fragmentation is particularly stark in heterogeneous environments. If each type of processor requires distinct programming models or optimisation techniques, developers may face increased complexity and higher switching costs. This can slow innovation and limit the ability of smaller firms to adopt advanced architectures.

From a policy perspective, promoting interoperability does not necessarily require mandating specific technologies. Instead, it would involve encouraging the development and adoption of open standards, supporting ecosystem collaboration, and ensuring that public investments in compute infrastructure prioritise flexibility and compatibility. In India, where the AI ecosystem includes a wide range of actors with varying levels of technical capacity, interoperability is essential to ensure that infrastructure investments translate into usable capabilities. Public compute initiatives, in particular, must be designed to support diverse workloads and avoid creating siloed environments.

At the same time, neutrality does not necessarily imply the absence of strategy. Governments and authorities ought to still make choices about where to invest and how to prioritise their resources. The core challenge lies in balancing strategic direction with flexibility, ensuring that infrastructure decisions remain adaptable to technological change.

3.2. Sustainability and Multi-Dimensional Benchmarking Across the AI Stack

The rapid expansion of data centre capacity has historically been accompanied by a strong focus on efficiency metrics, most notably power usage effectiveness ('PUE'). PUE measures the ratio of total facility energy consumption to the energy used by IT equipment. Over the last decade, it has become the dominant benchmark for assessing data centre efficiency, driving significant improvements in infrastructure design and operations. Current data shows that average PUE values have declined from approximately 2.0 in the early 2000s to around 1.5 or lower in leading facilities today.[35] Larger operators have reported even lower values

in optimised environments, reflecting advances in cooling systems, power management, and facility design. However, the continued reliance on PUE as a primary metric presents important limitations.

Firstly, PUE focuses exclusively on energy efficiency within the data centre boundary and does not account for the carbon intensity of the electricity consumed. This limitation has led leading operators to move beyond facility-level efficiency metrics toward energy sourcing and temporal alignment indicators. For instance, larger cloud providers have begun adopting carbon-aware strategies that track when and where electricity is consumed relative to clean energy availability, reflecting a shift from static efficiency measurement to dynamic carbon accounting.

A facility with a low PUE but powered by fossil fuels may have a higher overall environmental impact than a less efficient facility powered by renewable energy. Secondly, PUE does not capture water usage, which is increasingly relevant in regions facing water stress. Cooling systems, particularly those based on evaporation principles, can consume significant volumes of water. Studies indicate that large data centres can use millions of litres of water annually, depending on cooling technology and local climate conditions.[36] Thirdly, PUE does not reflect resilience under extreme conditions. As climate change increases the frequency of heatwaves and extreme weather events, data centre performance under stress conditions becomes a critical consideration. Facilities designed for optimal performance under average conditions may experience degradation or increased energy consumption during peak stress periods. Finally, PUE does not capture system-level impacts, such as strain on local grids or the interaction between data centre loads and broader energy systems. As data centres scale, their role as major electricity consumers requires a more integrated assessment of their impact on energy infrastructure.

These limitations have led to growing recognition that single-metric benchmarking is insufficient for guiding sustainable data centre development, particularly in diverse and rapidly evolving contexts such as India.

3.2.1. Designing Multi-Dimensional Sustainability Benchmarks

A multi-dimensional approach to sustainability benchmarking seeks to address these limitations by incorporating a broader set of parameters that reflect the full environmental and operational footprint of data centres. At a conceptual level, such a framework moves from measuring efficiency in isolation to assessing performance across multiple dimensions, including energy, water, carbon, resilience, and system integration. The objective is not to replace existing metrics such as PUE, but to situate them within a more comprehensive evaluation structure.

The European Commission has proposed a range of indicators for data centre sustainability, including energy reuse factor ('ERF'), water usage effectiveness ('WUE'), and carbon usage effectiveness ('CUE').[37] Similarly, industry certification bodies such as the Uptime Institute have emphasised the need for metrics that capture operational resilience and risk management alongside efficiency.[38] In addition, advances in software and hardware optimisation are enabling more granular measurement of performance. AI-driven monitoring systems can track energy consumption, cooling efficiency, and workload distribution in real time, allowing for dynamic optimisation. Industry practice has also been evolving in this regard. Some firms have now moved toward integrated measurement frameworks that combine infrastructure efficiency with carbon, water, and workload-level optimisation. Larger cloud platforms increasingly report metrics on renewable energy matching, water stewardship, and hardware-level efficiency gains, signalling a transition toward multi-layered sustainability assessment rather than single-point metrics.

These capabilities support the development of more sophisticated benchmarking frameworks that go beyond static metrics. In the Indian context, the design of such benchmarks must account for local conditions, including variations in climate, grid composition, and resource availability. A framework that is effective in temperate regions with abundant renewable energy may not be directly applicable in regions characterised by high temperatures, water scarcity, or carbon-intensive grids.

3.2.2. Accounting for Climatic, Geographic, and Grid Diversity

India's geographic and climatic diversity presents a fundamental challenge for uniform sustainability benchmarking. Temperature, humidity, water availability, and grid characteristics vary significantly across regions, influencing both the feasibility and performance of different data centre designs. For instance, air-based cooling systems that are effective in cooler climates may be less efficient in regions with high ambient temperatures. Similarly, water-intensive cooling solutions may be unsuitable in areas facing water scarcity. Grid composition also varies, with some regions having higher shares of renewable energy than others, affecting the carbon footprint of data centre operations. These variations suggest that a one-size-fits-all approach to benchmarking may be both impractical and counterproductive. Instead, benchmarks may need to be calibrated to regional conditions, allowing for flexibility while maintaining overall sustainability objectives.

At the same time, excessive fragmentation of standards can create complexity and reduce comparability. The challenge lies in designing a framework that balances standardisation with contextual adaptation. This may involve defining core metrics that apply nationally, supplemented by region-specific adjustments or thresholds. Additionally, geographic diversity has implications for infrastructure planning. Locating data centres in regions with favourable climatic conditions or renewable energy potential can improve sustainability outcomes. However, such decisions must be balanced against other considerations, including latency, connectivity, and regional development objectives.

3.2.3. A Proposed Framework for Approving Authorities

Building on the above discussion, a multi-dimensional benchmarking framework for data centre approvals in India can be conceptualised around five core dimensions, which are as follows:

- **Energy Efficiency and Performance:** This dimension includes traditional metrics such as PUE, supplemented by measures of IT equipment efficiency and workload optimisation. It emphasises both facility design and operational practices, including the use of advanced cooling technologies and AI-driven optimisation systems.
- **Carbon Intensity and Energy Sources:** This dimension assesses the carbon footprint of data centre operations, including both the share of renewable energy procurement and the temporal alignment between energy consumption and clean energy availability, reflecting emerging approaches such as hourly carbon-free energy matching.
- **Water Usage and Resource Management:** This dimension captures water consumption and efficiency, using metrics such as water usage effectiveness. It also considers the sustainability of water sourcing and the adoption of water-efficient cooling technologies, particularly in water-stressed regions.
- **Resilience and Climate Adaptation:** This dimension evaluates the ability of data centres to operate under extreme conditions, including heatwaves, flooding, and power disruptions. It includes design standards, redundancy measures, and stress-testing protocols to ensure operational continuity.
- **System Integration and Grid Impact:** This dimension assesses the interaction between data centres and the broader energy system. It includes considerations such as grid load management, participation in demand response programs, and alignment with regional energy planning.

3.3. Governance, Incentives, and Institutional Coordination

3.3.1. Centre-State Coordination in Data Centre Siting

The physical infrastructure of India's AI infrastructure is constrained by geographical concentration and friction in Centre-State administrative coordination, primarily because physical inputs (land, water, and grid power) fall under the jurisdictional purview of State governments. Currently, the spatial distribution of the nation's 271 data centres is

highly skewed, with Tier-1 hubs occupying ~23 million square metres of land overall and Mumbai monopolising nearly 25% of the national footprint, which leads to acute infrastructural congestion, localised power deficits, and water wars analogous to those witnessed in Western hyperscale hubs. Consequently, state-level policy responses have become highly fragmented and competitive. While states like Uttar Pradesh and Telangana actively court hyperscalers with incentives such as 50% land subsidies and complete stamp duty exemptions, other regions like Karnataka are actively reviewing their data centre policies specifically due to the unmanageable energy and freshwater burdens these facilities have on municipal resources. To resolve this competitive race among states, regulatory frameworks such as the Data Centre Policy 2020 advocate for the establishment of an Inter-Ministerial Empowered Committee and an independent Data Centre Industry Council to act as a unified, consultative interface between hyperscalers and State bureaucracies. By operationalising a National Single Window System equipped with deemed approval mechanisms for non-critical permissions, and strategically demarcating at least four Data Centre Economic Zones in coastal or resource-rich Tier-2 areas, the central government aims to enforce a harmonised, pan-India spatial plan that equitably distributes fiscal transfers while internalising local ecological carrying capacities into national siting approvals.[39]

3.3.2. Aligning Digital, Industrial, Energy, and Climate Policy

The exponential expansion of India's data centre capacity, which is projected to consume up to 6.5 GW of power and roughly 150 billion litres of water annually by 2030, exposes a critical regulatory dissonance between the nation's rapid digital acceleration ambitions and its legally binding climate and energy commitments. At present, institutional friction outweighs technological limitations in the pursuit of green computing. Data centre operators seeking to decarbonise their supply chains face hurdles in grid connection approvals, cross-subsidy surcharges, and inconsistent state-level banking rules that affect renewable energy open access.[40] To rectify this fragmentation, urgent institutional coordination is required between the relevant Ministries to embed digital infrastructure seamlessly within macroenvironmental frameworks, such as the National Green Hydrogen Mission, which outlays the decentralised production of clean energy. Furthermore, modern grid codes must evolve to legally classify AI data centres as flexible, dynamic grid assets capable of participating in demand response markets and executing load shifting during peak renewable generation windows. Ultimately, enabling standardised, long-term renewable power purchase agreements backed by utility-scale battery storage, alongside the waiving of interstate transmission charges for green energy utilised by compute facilities, will be vital to synthesising industrial growth with ecological preservation, ensuring that the hardware foundation of India's AI economy actively propels the national trajectory toward net-zero emissions.

3.3.3 Managing Trade-Offs Between Growth, Sustainability, and Resilience

As India relentlessly pursues an 8% annual GDP growth trajectory and seeks to inject $1.7 trillion into its economy via AI by 2035, policymakers face an acute strategic trilemma – balancing the infrastructural demands of digital sovereignty against ecological constraints and the necessity of long term climate resilience. Unlike the European Union, which has proactively enforced stringent sustainability governance through the revised Energy Efficiency Directive mandating strict energy and water reporting for data centres above 500 kW, zero-carbon power integration, and obligatory waste heat recovery for facilities exceeding 1 MW, India's regulatory apparatus currently relies heavily on voluntary compliance, prioritising rapid capital expenditure and capacity expansion over definitive ecological limits. However, the physical reality of modern AI workloads, which push server rack power densities to nearly three times those of conventional cloud computing, necessitates regulatory intervention to avert (extreme) resource depletion, particularly given that a typical 100 MW data centre requires approximately 2 million gallons of daily freshwater for cooling, placing immense strain on water stressed metropolises. To manage these profound trade-offs, governance structures must urgently transition from singular metrics like Power Usage Effectiveness to multidimensional sustainability benchmarking incorporating Water Usage Effectiveness and Energy Reuse Effectiveness (as discussed in previous chapter) whilst simultaneously mandating the use of treated wastewater and advanced direct-to-chip liquid cooling systems in all new hyperscale constructions. By fundamentally embedding frugal AI principles (optimising algorithmic efficiency to demand less compute power) and enforcing circular industrial symbiosis, India can uniquely architect a resilient digital ecosystem that pioneers environmentally sustainable digital sovereignty, proving that technological ascendancy in the Global South need not be predicated on the exhaustion of its critical natural capital.[41] The questions now is – how do we clarify trade-offs, institutional constraints, and identify areas where policy design must balance growth, sustainability, and resilience. •

End Notes

1. CarloRossotto, Artificial Intelligence Infrastructure: A System-of-Systems Approach. https://ssrn.com/abstract=6448014 or http://dx.doi.org/10.2139/ssrn.6448014
2. International Energy Agency, Energy and AI (Paris: International Energy Agency, 2025), https://iea.blob.core.windows.net/assets/de9dea13-b07d-42c5-a398-d1b3ae17d866/EnergyandAI.pdf.
3. Jon Gorey, “Data Drain: The Land and Water Impacts of the AI Boom,” Lincoln Institute of Land Policy, Land Lines Magazine, October 17, 2025, https://www.lincolninst.edu/publications/land-lines-magazine/articles/land-water-impacts-data-centers/.
4. International Energy Agency, Energy and AI. https://iea.blob.core.windows.net/assets/de9dea13-b07d-42c5-a398-d1b3ae17d866/EnergyandAI.pdf.
5. Thomas Le Goff, “Where to Put Data Centres: A Complex Problem for Europe,” Centre on Regulation in Europe, November 19, 2025, https://cerre.eu/news/where-to-put-data-centres-a-complex-problem-for-europe/.
6. Data Center Watch, “$64 Billion of Data Center Projects Have Been Blocked or Delayed amid Local Opposition,” 2025, https://www.datacenterwatch.org/report.
7. Rasmus Leichter, “Number of Data Centers by Country (November 2025),” Cargoson, last updated January 4, 2026, https://www.cargoson.com/en/blog/number-of-data-centers-by-country.
8. “Govt Accords Infrastructure Status to Data Centres with More Than 5 MW Capacity,” The Hindu, October 13, 2022, https://www.thehindu.com/business/govt-accords-infrastructure-status-to-data-centres-with-more-than-5-mw-capacity/article66006320.ece.
9. Press Information Bureau, “Capacity Building for Data Centre Operations,” Government of India, February 24, 2026, https://www.pib.gov.in/PressReleasePage.aspx?PRID=2221455&reg=3&lang=2.
10. Profile IT Solutions, “Data Center Policies, Subsidies & Approval Process in India,” https://www.profileits.com/data-center-policies-subsidies-approva a al/.
11. India Semiconductor Mission. "India Semiconductor Mission: Home." https://ism.gov.in/.
12. Press Information Bureau. "India Semiconductor Mission 2.0 - Press Release." Government of India. https://www.pib.gov.in/PressReleasePage.aspx?PRID=2224839&Reg=3&lang=1.
13. Press Information Bureau. "India Semiconductor Mission 2.0 - Press Release." Government of India. https://www.pib.gov.in/PressReleasePage.aspx?PRID=2224839&Reg=3&lang=1.
14. Press Information Bureau. "India Semiconductor Mission 2.0 - Press Release." Government of India. https://www.pib.gov.in/PressReleasePage.aspx?PRID=2224839&Reg=3&lang=1.
15. Council on Energy, Environment and Water (CEEW). "How is Data Centre Infrastructure in India Shaping Power and Water Use." https://www.ceew.in/publications/how-is-data-centre-infrastructure-in-india-shaping-power-and-water-use.
16. NVIDIA. "India Fuels Its AI Mission With NVIDIA." NVIDIA Blog. https://blogs.nvidia.com/blog/india-ai-mission-infrastructure-models/.
17. PR Newswire. "India Data Center Market to Grow at 14.60% During 2026-32, Driven by AI Integration, Hyperscale Expansion, and Digital Infrastructure Investments - MarkNtel Advisors." https://www.prnewswire.com/news-releases/india-data-center-market-to-grow-at-14-60-during-202632--driven-by-ai-integration-hyperscale-expansion-and-digital-infrastructure-investments--markntel-advisors-302603836.html.
18. pv magazine. "Why India's Digital Future Will Be Built at the Edge." https://www.pv-magazine-india.com/2026/01/21/why-indias-digital-future-will-be-built-at-the-edge/.
19. Data Centre Magazine. "L&T and NVIDIA's Data Centre Deal to Hyperscale India's AI." https://datacentremagazine.com/news/l-t-and-nvidias-data-centre-deal-to-hyperscale-indias-ai.
20. nasscom. "India's Generative AI Startup Landscape 2024." https://nasscom.in/knowledge-center/publications/indias-generative-ai-startup-landscape-2024-0.
21. Press Information Bureau. "Sarvam AI Powering a Made-in-India AI Revolution." Government of India. https://www.pib.gov.in/PressReleasePage.aspx?PRID=2231169&Reg=3&lang=1.

22. Businessworld. "AI-led Cluster Strategy Recasts MSME Competitiveness In Global South." https://www.businessworld.in/article/ai-led-cluster-strategy-recasts-msme-competitiveness-in-global-south-593598.

23. PwC India. "Unlocking the AI Edge for MSMEs." https://www.pwc.in/assets/pdfs/unlocking-the-ai-edge-for-msmes.pdf.

24. NITI Aayog. "AI for Viksit Bharat: The Opportunity for Accelerated Economic Growth." Government of India. https://niti.gov.in/sites/default/files/2025-09/AI-for-Viksit-Bharat-the-opportunity-for-accelerated-economic-growth.pdf.

25. NITI Aayog. "AI for Viksit Bharat: The Opportunity for Accelerated Economic Growth." Government of India. https://niti.gov.in/sites/default/files/2025-09/AI-for-Viksit-Bharat-the-opportunity-for-accelerated-economic-growth.pdf.

26. IndiaAI. "IndiaAI Compute Capacity." Ministry of Electronics and Information Technology, Government of India. https://indiaai.gov.in/hub/indiaai-compute-capacity.

27. “Will Datacenter Growth in India Propel It to Global Hub Status? | S&P Global,” accessed March 19, 2026, https://www.spglobal.com/en/research-insights/special-reports/india-forward/shifting-horizons/will-data-center-growth-india-propel-country-global-hub-status.

28. Nguyen Anh Tuan et al., “Edge AI for Smart Energy Systems: A Comprehensive Review,” Edge Computing - Latest Advancements, Challenges, and Applications [Working Title], ahead of print, September 29, 2025, https://doi.org/10.5772/intechopen.1011425.

29. Laure De Roucy-Rochegonde and Adrien Buffard, AI, Data Centers and Energy Demand: Reassessing and Exploring the Trends, 2025.

30. International Energy Agency, “Data Centres and Data Transmission Networks,” last modified July 11, 2023, https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks.

31. Jesse Noffsinger, Mark Patel, and Pankaj Sachdeva, “The Cost of Compute: A $7 Trillion Race to Scale Data Centers,” McKinsey & Company, April 28, 2025, https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-cost-of-compute-a-7-trillion-dollar-race-to-scale-data-centers.

32. International Data Corporation, “Worldwide Spending on Edge Computing Forecast to Reach $378 Billion by 2028,” press release, 2024, https://my.idc.com/getdoc.jsp?containerId=prUS52587424.

33. International Energy Agency, “Data Centres and Data Transmission Networks,” last updated July 11, 2023, https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks.

34. Stanford Institute for Human-Centered Artificial Intelligence (HAI), The 2024 AI Index Report, April 15, 2024, https://hai.stanford.edu/ai-index/2024-ai-index-report.

35. Uptime Institute, Uptime Institute Global Data Center Survey Results 2023, Uptime Institute Intelligence, 2023, https://uptimeinstitute.com/resources/research-and-reports/uptime-institute-global-data-center-survey-results-2023.

36. Arman Shehabi et al., United States Data Center Energy Usage Report, LBNL-1005775 v2 (Berkeley, CA: Lawrence Berkeley National Laboratory, June 2016), https://eta-publications.lbl.gov/sites/default/files/lbnl-1005775_v2.pdf.

37. John Booth and Mark Acton, 2023 Best Practice Guidelines for the EU Code of Conduct on Data Centre Energy Efficiency, JRC Technical Notes (Ispra: European Commission Joint Research Centre, February 9, 2023), https://e3p.jrc.ec.europa.eu/en/publications/2023-best-practice-guidelines-eu-code-conduct-data-centre-energy-efficiency.

38. Uptime Institute, “Uptime Institute’s 2022 Outage Analysis Finds Downtime Costs and Consequences Worsening as Industry Efforts to Curb Outage Frequency Fall Short,” press release, June 8, 2022, https://uptimeinstitute.com/about-ui/press-releases/2022-outage-analysis-finds-downtime-costs-and-consequences-worsening.

39. NITI for States. "Data Centre Policy 2020." Government of India. https://www.nitiforstates.gov.in/public-assets/Policy/policy_files/PNC510C000384.pdf.

40. Council on Energy, Environment and Water (CEEW). "How is Data Centre Infrastructure in India Shaping Power and Water Use." https://www.ceew.in/publications/how-is-data-centre-infrastructure-in-india-shaping-power-and-water-use.

41. Ministry of Finance. "2025-26." India Budget. Government of India. https://www.indiabudget.gov.in/economicsurvey/doc/eschapter/epreface.pdf.

Recommended Citation

Agarwal. M, Panjwani. S, Joshi. M, Patil. N and Shekar. K (2026). Greening the Code: A Roadmap for Environmentally Responsible AI Innovation in India, Chintan Research Foundation (CRF), New Delhi

Chintan Research Foundation

info@crfindia.org

www.crfindia.org

The Dialogue®

INFORM ENGAGE IDEATE

info@thedialogue.org.in

www.thedialogue.org.in