rp.

IMA

Research Paper

One Long Thread: Computing, Intelligence and India

NVIDIA's India chief traces sixty years of computing history and what it means for India's AI moment

MAY 2026

IMA

RESEARCH PAPER
MAY 2026

Executive Summary

- Modern computing has operated on a three-layer architecture since the 1950s. AI introduces a new intelligence layer on top of it.
- GPUs, originally designed by NVIDIA for graphics processing, became the foundation for modern AI systems.
- ImageNet 2012 marked the breakthrough moment when GPUs sharply improved image recognition accuracy.
- AI's current boom rests on four pillars: cheap compute, vast internet data, modern neural networks and scalable infrastructure.
- Most Indian firms are renting AI capability from hyperscalers, sending both data and value creation abroad.
- India lacks frontier AI models but may benefit from open-source systems that reduce dependence on closed platforms.
- The central challenge is whether India can build AI for its own languages, users and economic needs.
- India's opportunity lies in augmenting engineers with AI rather than replacing labour through full automation.
- India's AI capacity has risen from near zero to roughly 2 GW in three years, with ambitions of 8 GW by 2030.
- The investments made now will shape whether India becomes a builder of AI systems or only a consumer of them.

The AI Reality Check is an IMA India research series examining how artificial intelligence is changing the operational reality of Indian enterprises. It is based on in-depth interviews with CXOs and senior operational leaders from IMA member companies, supplemented by secondary research. The series is concerned not with where AI is headed but with where it stands today: what has moved from pilot to deployment, what constraints are shaping adoption and what choices will determine whether Indian enterprises emerge from this period as builders of AI capability or consumers of it.

This is the first of five papers.

One Long Thread: Computing, Intelligence and India

Most conversations about AI's use in business start in the present tense: Which tools are being deployed? Which processes have been automated? Which vendors have won which contracts? Vishal Dhupar, Managing Director of NVIDIA India and a 40-year veteran of the technology industry, starts somewhere different. He starts in 1964.

IMA India spoke with Mr Dhupar as part of a research series examining how AI is changing the operational reality of Indian enterprises. Whereas most such conversations would look at the AI landscape from the inside-out, outlining company-level deployment, constraints and challenges, this interview offered something rare: a coherent account of the entire arc of computing history. Mr Dhupar explained why AI has arrived in the form it has, why India's position is both promising and precarious and what choices will determine whether the country builds something durable, or simply rents capability from others.

The Same Idea, Six Times

The story of modern computing begins in April 1964 with a mainframe, The IBM System/360 (S/360), the first computer designed to separate the application layer from the hardware layer using an operating system. Before the S/360, software was written for specific hardware. When you replaced the machine, you replaced everything. After it, an application could travel. The hardware could be upgraded without disrupting the programs running on it. Everything that came afterwards, from the departmental minicomputers of the 1970s, the personal computers of the 1980s, the internet and the smartphones of today, are built on the same 3-layer foundation of hardware at the bottom, operating system in the middle and applications on top.

What sustained this architecture through six decades of change was Moore's Law: the observation, made by Intel co-founder Gordon Moore in 1965, that the number of transistors on a chip would roughly double every two years at roughly the same cost. This meant that every two years the same application ran twice as fast and cost half as much to run. The IT industry's long growth cycle was, in large part, a direct consequence of this. As computing became cheaper, more applications became economically viable, which generated demand for more compute.

Less commonly understood is how decisive the move away from this architecture has been. Over the last decade, Moore's Law on its own would have delivered roughly a hundred times more compute.

Accelerated computing actually delivered approximately a million times; ten thousand times beyond what physics alone would have permitted in the same window.

Mr Dhupar describes this gap as 'the substrate of the AI moment.' Without it, he argues, none of the present generation of AI systems: large language models, generative AI, ability to reason and agentic frameworks would have been economically viable to train or operate. The AI moment, in this telling, is not a software accident. It is the product of a thirty-year hardware bet.

> ***The substrate of the AI moment is not Moore's Law. It is what we built beyond it.***

By the early 1990s this dynamic had made the personal computer market enormous but structurally undifferentiated. Every manufacturer was buying from the same Intel/Microsoft supply chain; the only competitive variable was price. Mr Dhupar describes the three founders of NVIDIA as 'contrarians' who looked at this race to the bottom and asked a different question. Instead of *how to compete* within the architecture, they sought to find out *where it was weak*.

The answer was specialisation. The CPU was designed to handle any calculation, but optimised for none. Certain categories of computation, such as computer graphics, fluid dynamics and image processing, required thousands of simpler calculations run simultaneously, rather than one complex calculation run fast.

NVIDIA built a processor for that: the Graphics Processing Unit, or GPU, a term the company coined itself. Its first commercial focus was gaming, on the logic that making digital environments look physically real required solving exactly the kind of parallel mathematical problems the GPU excelled at.

What Happened in 2012

For its first fifteen years, the GPU was a specialist graphics chip with a growing sideline in scientific research. The moment that changed its trajectory came at the ImageNet Large Scale Visual Recognition Challenge in 2012. The competition asked entrants to classify 1 million images across a thousand categories. Top-5 error rates had hovered

around 26% for two years, improving by roughly one percentage point annually through incremental refinements to hand-coded algorithms. In 2012, a team from the University of Toronto led by Geoffrey Hinton submitted an entry built on a different premise. Rather than writing an algorithm to recognise images, they built a system that could learn to recognise them, using a deep neural network trained on NVIDIA GPUs via the CUDA programming framework. The result was a top-5 error rate of 15.3%, against the prior year's 26.2%: a reduction of nearly eleven percentage points in a single iteration.

What made this possible was the GPU's parallel processing architecture. Originally developed to render realistic game graphics, it was precisely suited to the matrix multiplications at the heart of neural network training. The same hardware that made a game's sunlight look real could learn to distinguish a cat from a dog. The research community's attention shifted rapidly. Within 5 years, deep learning had moved from a niche academic technique to the dominant approach in computer vision, speech recognition and natural language processing.

The next structural shift came in 2017, when Google researchers published 'Attention Is All You Need,' introducing the transformer architecture. Transformers enabled unsupervised learning at scale: a model could now be trained on the entire text of the internet without labelled data, encoding the statistical relationships between words, concepts and contexts into a vast parameter space. The result, combined with continued increases in GPU compute, was the large language model (LLM), the technology that underpins ChatGPT, Gemini, Claude and their competitors.

Other People's Intelligence

The availability of powerful AI models through cloud APIs has created, in Mr Dhupar's view, a deceptively comfortable situation for Indian enterprises. The barrier to accessing state-of-the-art AI capability has never been lower: a company can call an OpenAI or Google API and get generative AI functionality in hours, without owning any of the underlying infrastructure. But this convenience has a structural cost.

The data sovereignty problem

When an enterprise sends its proprietary data to a foreign model for processing, the data and learnings derived from it enrich a system it does not own. Take, for instance, Google Maps. In its early years in India, the navigation product spoke with a generic accent and recommended routes

that did not account for local landmarks or road naming conventions. Over time, as Indian users corrected it and fed it local context, it adapted. The model got smarter. But who owned that intelligence?

The same logic applies to every Indian enterprise currently feeding its proprietary knowledge into a foreign AI model. The specific knowledge that differentiates one cement company from another, one bank from its competitors, one pharmaceutical manufacturer's formulation expertise from its rivals' — all of this is data that, if fed into an external model without sovereignty controls, leaves the enterprise's boundaries and does not return.

When critical workflows depend on external models, Indian enterprises surrender control over where data is processed, how long it is retained, which jurisdiction governs access, how pricing changes, and whether proprietary knowledge becomes a durable internal asset. Even if raw data remains contractually protected, capability can still migrate outward: model behaviour, integration logic, fine-tuning methods and operational intelligence sit inside someone else’s stack. For India, AI sovereignty therefore means more than data localisation. It means retaining enough control over compute, models, data pipelines and application IP to decide what should be rented, what should be owned, and what must never become a dependency

The rent versus build calculus

Mr Dhupar frames the choice facing Indian enterprises in terms familiar from real estate. You can either rent AI capability (which is fast, cheap and requires no infrastructure commitment); or you can build it, investing in proprietary data architecture, fine-tuned models and sovereign compute. Renting is certainly the path of least resistance. However, building is what generates IP that compounds.

What has changed in the last twelve months, Mr Dhupar argues, has made this choice substantially less binary than it appeared two years ago. The performance gap between open-frontier models and proprietary closed models has narrowed considerably. Open-source models like *DeepSeek*, *Llama*, *Qwen*, *Mistral*, NVIDIA's own *Nemotron* are now competitive with closed frontier models on the benchmarks that matter for enterprise work. By October 2025, monthly downloads on Hugging Face, the principal open model repository, had surpassed 160 million.

This opens what Mr Dhupar describes as a third path between training a frontier model from scratch and renting capability from a foreign hyperscaler. The third path is to take open-frontier models and post-train

them on proprietary data in Indian languages, deploying them on Indian compute. Indian labs are already operating on this path: Sarvam in 22 Indic languages, BharatGen building India-specific foundational models ecosystem relies on. Three years ago, this option did not exist at the frontier of capability. Today it does.

> ***If the substrate is open, the gatekeeper goes away.***

Among the other companies we interviewed, the ones currently making the most durable progress are those that made this choice deliberately. (These will be explored in the papers that follow in this series.)

BSV Group (Bharat Serums and Vaccines) built its data architecture before deploying AI, and credits that discipline with the quality of the output it now generates. HT Media had already constructed a centralised data lake before generative AI became commercially viable; it describes this as a key structural advantage. Even beyond this research, a visible pattern seems to be that the companies still experimenting with AI 6 months into a pilot are, almost without exception, the ones that handed their data to a third-party service and hoped for the best.

The 5-layer AI Factory

A framework for understanding AI infrastructure describes five layers that must be present for a functioning AI capability:

1. Energy: the power, cooling and grid capacity that physically enables AI workloads
2. Chips and Systems: Codesign Chips, System, Network. Software, Cooling as one computer.
3. Infrastructure: data centres, networking, cooling fabric, fibre and the physical plant
4. Models: the trained AI systems
5. Applications: the software built on top of them where the economic value sits

Measuring India's current position against this framework offers a mixed diagnosis. On applications, India is strong. Its software engineering talent base, built over 50 years from the mainframe era through to the global IT services industry, means it has more people capable of building AI-powered applications than almost any other country. On models, it is beginning to invest.

The government's IndiaAI Mission, backed by a Rs 100 bn budget allocation announced in 2024, includes provisions for computing infrastructure and Indic language model development. On land, the 2025 Union Budget introduced provisions to incentivise data centre construction, offering tax benefits to operators serving both domestic and export markets. However, on processors, GPUs, the hardware that actually trains and runs AI models, India remains almost entirely dependent on foreign supply and foreign cloud infrastructure.

This gap matters more than it might appear. Processors are a strategic bottleneck. Access to GPU compute during the current period of rapid AI model development determines who gets to train frontier models, who can fine-tune existing ones on proprietary data and who is simply a consumer of whatever capability others choose to offer. India remains heavily dependent on hyper-scaler APIs, with limited domestic GPU capacity and sits solidly in the consumer column.

The Augmentation Argument

Mr Dhupar draws a distinction that most AI commentary ignores: between the India that is building its own AI capability and the India that is simply consuming capability built elsewhere. The two stories are unfolding in parallel, at different speeds and with different stakes.

The first is government-led. India’s Bhashini platform, the National Language Translation Mission, the emerging ecosystem of Indic language models, among them Sarvam AI, AI4Bharat from IIT Madras, Krutrim, are all designed to make AI functional for the 900 million Indians who do not interact primarily in English. From a farmer checking the timing of the monsoon to a first-generation smartphone user navigating government services, the value here depends entirely on AI working in Hindi, Tamil, Telugu and dozens of other vernacular languages.

The second, and currently the less impressive of the two, is the enterprise track. large Indian companies are experimenting with chatbots and summarisation tools, pulling down API calls to foreign models and describing the results as transformative without having made any of the structural investments that would make that true. The Budget’s data centre incentives are a signal that the government wants to change this. Whether enterprises follow depends on whether boards treat AI infrastructure as a capital allocation decision rather than an IT procurement line item.

Underneath both stories is a question about India's natural competitive asset. One of the persistent anxieties in India’s AI discourse is about

displacement: that AI will eliminate the software engineering jobs that have anchored India's export economy for five decades. Mr Dhupar's view is almost the opposite, and it rests on the same augmentation logic that originally justified the GPU.

China's factories are staffed by labour that is already, by global standards, productive and cost-effective. Yet Chinese manufacturers are deploying robots faster than anyone else in the world. As of 2025, there were ~2 million industrial robots working in China’s factories. Why automate labour that is already cheap? Because augmenting it with intelligent systems widens the productivity gap further. The combination of a human worker and a robot is not simply a human worker who does not get tired; it is a qualitatively different production unit.

The same logic applies to Indian software engineers: a developer using a code generation tool does not produce the same output as a developer without one. A controlled study by GitHub and MIT found that developers completed applicable tasks up to 55% faster when using AI code generation tools. But crucially, the value of that tool depends on the context it operates in. A generic code generation model produces generic code. A model trained on the specific architecture of an Indian bank's core banking system, or on the proprietary protocols of a logistics company's freight management platform, produces something more valuable. ‘The ones who hold the data,’ according to Mr Dhupar, ‘are the ones who will extract the most from augmentation.’

> ***“The ones who hold the data are the ones who will extract the most from augmentation”***

Specific examples illustrate how internationally developed AI infrastructure, including NVIDIA’s software stack, could support India’s own ecosystem without requiring every foundational component to be built domestically. In logistics, optimisation libraries such as cuOpt can help Indian operators improve route planning, fleet utilisation and delivery efficiency at national scale. In foundation models, frameworks such as Megatron and NeMo can give Indian labs a faster path to training, adapting and deploying models for local languages, sectors and regulatory contexts. In life sciences, genomics tools such as Parabricks can accelerate large-scale sequencing and analysis efforts, including population-scale initiatives such as GenomeIndia. The strategic point is not that India must own every layer of the global AI stack from day one. It is that Indian institutions can deliberately wield proven infrastructure to create India-specific

capability on top of it: models, datasets, workflows and applications that serve domestic priorities and compound into local IP.

This is why India's IT services majors, including Wipro, TCS, Infosys and Tech Mahindra among others, are better positioned than the narrative of AI disruption suggests. They hold, in aggregate, decades of enterprise application code and operational data from the industries they serve. That data is the raw material for differentiated AI. The question is whether they use it to build proprietary intelligence or continue to act as intermediaries for someone else's.

Sixty Years and Counting

Mr Dhupar's most ambitious claim is that India could become, in AI, what Paris is to design: a recognised global centre of a particular kind of capability. He grounds it in three assets. An engineering talent base of genuine scale, a digital public infrastructure stack comprising Unified Payments Interface (UPI), Aadhaar and Open Network for Digital Commerce (ONDC) that has generated datasets few other countries can match, and a degree of policy intent that, while not yet matched by private investment, provides a real framework. He is also candid about what India lacks: infrastructure and the willingness to make the rent-versus-build choice decisively.

> ***"We are testing the waters. We've got talent here. We lack infrastructure and we need to bring that up"***

Several developments over the past 12 months have sharpened the choice. Indian AI factory capacity has moved from near-zero to approximately 2 gigawatts operational, with a path to 10 gigawatts by 2030, supported by capital commitments now totalling roughly $ 300 bn across operators, hyperscalers and the Indian operators. Globally, the question of whether AI is commercially self-funding has effectively been answered: in the last twelve months, gross margins at every major AI-native company, OpenAI, Anthropic, and Cursor among them, have turned strongly positive. The trillion-dollar capital expenditure commitments reported almost weekly are no longer speculative; they are the rational response to demand curves that are now widely understood within the industry. For Indian enterprises and policymakers, this changes the cost of delay. A decision deferred two years ago could be revisited at modest expense. A decision deferred today is taken inside a market already substantially built.

The arc from the IBM S/360 to the GPU and the transformer is a story about specialisation. Each step adds a layer the previous architecture could not handle, without discarding what came before. India's position inside that arc is open. It has, in its software talent and its digital infrastructure, assets that matter in the AI era. It faces, in its compute dependency and its current preference for renting over building, a structural risk that is not yet critical but becomes harder to reverse with each passing year. The rest of this series will examine how Indian firms are building AI capabilities, from data architecture and governance to the evolving division of labour between humans and machines. Mr Dhupar's sixty-year account is the lens through which all of these should be read.