AN INQUIRY

“We will quickly

lose even the social permission

to actually take something like energy, which is a scarce resource, and use it to generate these tokens,” said Satya Nadella. “If these tokens are not improving health outcomes, education outcomes, public sector efficiency, private sector competitiveness across all sectors, small and large.” [1]

“There will be very hard parts like

whole classes of jobs going away

Sam Altman wrote in a July blog post. “But on the other hand the world will be getting so much richer so quickly that we’ll be able to seriously entertain new policy ideas we never could before. [...] “That’s how capitalism works, and the ecosystem and economy would be fine; we plan to be a wildly successful company, but if we get it wrong, that’s on us.”[2]

Jamie Dimon, has said artificial intelligence

“may go too fast for society”

and cause “civil unrest” unless governments and business support displaced workers. While advances in AI will have huge benefits, from increasing productivity to curing diseases, the technology may need to be phased in to “save society”, he said.[3]

"We remain nervous that the economy is subject to

potentially large shocks

that may not have an immediate effect" but could build over time, IMF Chief Economist Pierre-Olivier Gourinchas said in a briefing with reporters. However, if the hoped-for productivity gains from the new technology don't pan out, it could trigger an "abrupt" slump in markets that could spread to other sectors and erode household wealth, the fund said. [4]

"I'm on the Meta oversight board.

We need AI protections now.

We know that AI can be dangerous; chatbots advise teens on suicide and may soon be capable of instructing on how to create biological weapons. Yet there is no equivalent to the Federal Drug Administration, testing new models for safety before public release." — an opinion published by Suzanne Nossel, on The Guardian.[5]

AN INQUIRY

Over the past few weeks, unless you're an undeterred optimist that's privy to the vision of a future that is *not* in fact entirely dystopian, it has felt like the artificial intelligence race has lost some of its direction and ended up at something akin to a juncture. Not necessarily a juncture from where the trajectory will change radically, but very much a juncture where one wouldn't be dismissed as a Luddite for asking questions. And so, through this series, we ask questions — questions about where we're headed and about what that might cost us, today and tomorrow. We start at home. By all means, India should have a lot going on. Sometimes, it seems like it does, but sometimes, it seems like it will. That's the premise of this first part — the 'does', the 'will', and what it would take to bridge that gap.

BCL
TRANSFORMING LIVES WITH WISDOM

Established in 2011, the BCL group has presence in India and United Arab Emirates. The team comprises of ~ 300 personnel including ~30 Chartered Accountants and 8 Company Secretaries. Under the leadership of its 16 partners, BCL offers the full spectrum of corporate financial services, corporate & start-up advisory, and transfer pricing and international taxation. Over the years, our services have been sought and availed by businesses around the world, making BCL the perfect partner for start-ups and established enterprises alike.

Report Author: Ranjani, Senior Associate, BCL India.

INTRODUCTION

That India has a research and development problem is a tale as old as time. We don't spend nearly enough. Corporates and academic institutions operate in silos, with little to no structural linkages. Recent empirical evidence is stark — as per the Artificial Intelligence Index Report, 2025[6] issued by Stanford:

1. Despite publishing 9.22% of all artificial intelligence publications in computer science between 2013 and 2023 (ranking only marginally behind United States' 9.20%), India has a meagre share of 0.37% (way, way behind United States' 14.16%, and China's 69.70%) in all granted AI patents between 2010 and 2013 — a disparity that is likely created and sustained by the incentivization of publications, but perhaps not as much what should result from such publications.
2. India's contribution to AI projects on GitHub has increased steadily between 2011 and 2024, and stands at 19.91%, marginally ahead of Europe and only 3 percentage points behind the United States. The country ranked second among all nations in "AI Skill Penetration" (by definition, a measure of the intensity of AI skills in a particular country or by industry or gender that signals the prevalence of AI skills across occupations), only behind the United States. And yet, India saw among the lowest private investments in AI in 2024, at $ 1.16 billion. Compare that with the United States (109x) or China (9.3x) or United Kingdom (4.5x), and the gulf in the number of newly funded companies explains itself. Neither contributions to GitHub nor AI Skill Penetration are particularly compelling or incontrovertible indicators of our talent pool — but neither the existence *of* a talent pool nor a lack of support therefor are disputable.

But this is not a gap that can, or should subsist. And so, more than two years out from ChatGPT's release, and nearly a year from the official announcement of the program, about a few weeks after DeepSeek's official release, the India AI mission started to gain momentum. India's positioning in a global race could no longer merely hinge on its standing as an ITeS behemoth, or as a talent pool that pervaded international borders. It needed its own ecosystem.

INTRODUCTION

The IndiaAI mission identifies seven pillars:

1. Creating subsidized access to compute infrastructure through an effective private-public partnership — the current target stands at ~38,000+ GPUs
2. Develop India-specific foundational models and large language models (LLMs) tailored to Indian languages and domains, advancing IP creation
3. Build anonymized, consent-based, interoperable public datasets to fuel model training — or 'AIKosh'
4. Development of sector-specific AI applications and solutions
5. Upskilling with a view to future-proofing
6. Start-up financing, and
7. Create frameworks and toolkits to ensure explainable, ethical, and privacy-preserving AI.

The ensuing sections of this Report will focus on the first three of these, those that will ultimately constitute the 'building blocks', of our ecosystem when we do have one. That said, it is important to acknowledge that development of a sovereign ecosystem (the importance of which cannot be overstated in an age where AI companies are going back and forth with the Department of War of a country no less than the United States of America) is only one facet of India's AI story. Much of the immediate economic outcomes will come from the existing traction with #4 ("Primary research indicates that for almost 67% of AI startups in India, the core area of work is in the AI application model layer.")[7] And that has not missed the government's attention: Union Minister for Electronics and Information Technology Ashwini Vaishnaw's remarks at Davos, 2026 are of note — "At the application layer, India will probably be the biggest supplier of services to the world," he said, adding that return on investment (ROI) in AI comes from enterprise-level deployment and productivity gains, not from creating very large models alone. The application layer as a lever for economic growth (development and deployment alike) is also among the key propositions of the Niti Aayog. Their September 2025 report[8] suggests that:

"AI adoption could contribute an

additional $500–600B

to India's GDP by 2035, beyond the projected growth trajectory, driven by productivity improvements, operational efficiencies, and the reallocation of human effort to higher-value tasks"

"Deployment of AI in eighteen selected sectors can accelerate commercialization, disrupt legacy value chains, and create a lasting competitive edge. Analysis suggests that such breakthrough innovations could potentially contribute at least an

incremental $280-475B

to India's GDP by 2035."

COMPUTE

At this juncture, it would be good to zoom out and recognize how integral physical infrastructure, and by extension, access to extraordinary amounts of capital is, to the development of this technology. Artificial intelligence rests on a monumental stack of chips (the intelligence), data centers (its residence), power (to keep the lights on), and water (to keep it from implosion). McKinsey estimates the amount required for data centers equipped to handle AI processing loads to require a capital expenditure of a modest $5.2 trillion by 2030[9]. Now while the estimate itself is constrained by several uncertainties about the end outcome of this race, it is a good indicator of what the stack takes. (If we needed any further indication than the severe geographic asymmetry in the distribution of foundational models, that is.) As it stands now, for India, none of these are advantages, or attainable at that. The 2025-26 Economic Survey says it well —

> *For India, as power, finance, and especially water remain binding constraints, scaling compute indiscriminately carries opportunity costs. Investment in AI infrastructure competes directly with other sources of demand, such as households and industries. This creates a trade-off between centralised scale and distributed efficiency, strengthening the case for smaller, task-specific models that can run on limited hardware and decentralised compute networks.*

And such limitations are perhaps why the IndiaAI mission has chosen to focus on just one layer of the stack — the chips. Of the total ~INR 10,372 crore allocated to the mission, “India AI Compute Capacity” is allocated ~INR 4,600 crore, roughly 44% of the total. What began with a target of 10,000+ GPUs has now almost quadrupled, to 38,000+ GPUs, (with an additional 20,000 GPUs committed in the AI summit, bringing it up to ~60,000 GPUs) being made available at ‘one-third’ the global cost, and ‘unlike many countries where big tech controls GPU access.’ While this is a sensible enough proposition, one must recognize that subsidising procurement to chips that are controlled, in their entirety, by a global supply chain (which in turn is subject to geopolitical volatility) is a transitory fix to a complicated problem.

COMPUTE

A problem that exists, and will intensify at the interaction of inordinate concentration of market power (Nvidia's share of the GPU market is pegged at 85% as of January, 2026)[10], mercurial government policies on export control (as of 14th January 2026, the US Commerce Department on Tuesday[11] opened the door for Nvidia to sell advanced artificial intelligence chips in China with restrictions, a development that came through about a 15 days after CNBC[12] reported a '$160 million AI-chip smuggling ring') and an unprecedented, uncertain upsurge in demand (there's as much surety about the existence of a compute power demand curve as there is uncertainty about its slope — while McKinsey expects a 3x increase in global demand for data centers by 2030, the expectation is also heavily disclaimed by uncertainty over what the actual business impact of AI might look like.)[13] India does not make chips of its own and while Budget 2026 has indicated further government intent with the India Semiconductor Mission 2.0 being given an a further outlay of Rs. 40,000 crore, any possibility of a sovereign supply chain is distant. To that extent, the declaration that India's compute infrastructure is not controlled by 'big tech' is... well, ironic. The Economic Survey of 2025-26 has taken an interesting approach to modelling bottlenecks in expansion of compute power, as a policy stress test. And even in the baseline scenario, about 10 quarters in:

> *"The primary constraint shifts toward GPU availability. (from financing / power). Even with "normal" global supply conditions and India's GPU share in global demand set at a realistic 4%, the probabilistic access mechanism and minimum lead times result in a persistent and growing pool of operators waiting for hardware. This bottleneck dominates for most of the simulation horizon, suggesting that hardware access, rather than pricing or power availability, becomes the central limiting factor in sustained capacity expansion."*

In the second scenario, where foreign demand is elevated, the result is more stark: "The result is a decoupling between financial readiness and execution capability, underscoring that easing domestic financial constraints alone is insufficient when hardware supply remains externally constrained." ***Subsidising procurement, then, would work only to the extent that procurement is ... well, possible.***

MODEL

Between Nandan Nilekani's declaration in August 2024 that "Our goal should not be to build one more LLM. Let the big boys in the (Silicon) Valley do it, spending billions of dollars", to the government's selection of Sarvam to build the country's 'sovereign large language model' in April 2025, to the Economic Survey's proposition in January 2026 that India is better off taking a 'bottom-up approach — one that prioritises application-specific, small models that are tailored to defined uses and sectoral needs, which will allow broader development and diffusion of AI solutions, free from concerns about resource constraints or high entry barriers', it is safe to say that the response to 'What should India build' has seen some flux over the last three years. In our examination of this question in this Report, it would serve us well to answer two intersecting sub-questions —the merits of small language models, and the hows and whys of artificial intelligence in regional languages.

[1] Small Language Models

Against the background of the preceding section on India's very real compute limitations, the best place to start this section would be a primer on what distinguishes small language models from large language models —

> *Small language models are more compact and efficient than their large model counterparts. As such, SLMs require less memory and computational power, making them ideal for resource-constrained environments such as edge devices and mobile apps, or even for scenarios where AI inferencing—when a model generates a response to a user's query—must be done offline without a data network.*[14]

Artificial general intelligence (AGI) —a hypothetical state where intelligence is no longer necessarily exclusively human — has long been made out as the very exciting outcome of this race. The acceptance of, and rapid advancements in agentic artificial intelligence — a state where cognitive effort is not necessarily exclusively human — seems, then, like an inevitable consequence. And it may well be the primary axis of AI, for the foreseeable future. And if research suggests that small language models are well-suited for application, that might be the way forward for India.

MODEL

It makes sense, then, to reproduce extracts of the paper published by Nvidia's researchers in late 2025[15] in our defence of that argument —

> ***On the democratization it would enable:*** *One particularly notable and desirable consequence of SLM flexibility when put in place of LLMs is the ensuing democratization of agents. When more individuals and organizations can participate in developing language models with the aim for deployment in agentic systems, the aggregate population of agents is more likely to represent a more diverse range of perspectives and societal needs.*
> ***On the economic considerations:*** *Due to their small size and the associated reduction in pre-training and fine-tuning costs, SLMs are inherently more flexible than their large counterparts when appearing in agentic systems. As such, it becomes much more affordable and practical to train, adapt, and deploy multiple specialized expert models for different agentic routines.*
> ***And finally, perhaps most interestingly, on why we the buzz is still muted:*** *Large capital bets have been made on the centralized LLM inference being the leading paradigm in providing AI services in the future. As such, the industry has been much quicker at building the tools and infrastructure to that end, omitting any considerations for the possibility that more decentralized SLM or on-device inference might be equally feasible in the near future.*

For the boundaries within which India's AI story has to take origin and evolve — scarce compute, fledgling research ecosystems and almost prohibitive diversity — carving out compact models for precise problem statements that don't need data center acreage seems infinitely more plausible, grounded in reality and solution-oriented than any attempt to build a frontier model with a wide sweep. And given their ability to run on private servers (or on edge), these models are inherently architected to be secure, allowing safe deployment critical sectors like healthcare and defence.
All put together, maybe India's AI story scales new heights by starting small.

MODEL

[2] Regional Language Models

When one is asked to name mankind's three greatest inventions, *language* is probably not at the top of that list — but maybe it should be. Now while there are by-the-book definitions, (Cambridge defines language as 'a system of communication by speaking, writing, or making signs in a way that can be understood, or any of the different systems of communication used in particular regions') in its simplest (and perhaps oversimplified) form, language is a documented consensus on expression. If one acknowledges that the act of expression is intrinsic to being human, the systematic erosion of language that has already been caused by the overwhelming proliferation of English should be cause for concern. (According to UNESCO's World Atlas of Languages, there are 7,000 languages, spoken or signed, in use in the world today – and only 351 languages are used as the medium of instruction. One language disappears every two weeks.) And this erosion will only be exacerbated when technology that is purported to change our lives as we know it is predominantly trained in English:

> *"The assumption is that English is the de facto standard for everything because it is particularly used in academic settings, and many of the builders are targeting U.S. and European usage, and the effect has been that the data skews toward particular types of English. The models are much less performant beyond that." — Sanmi Koyejo, assistant professor of computer science at Stanford University and an affiliate of the Stanford Institute for Human-Centered AI.*[16]

Now while this is clearly a priority, and one recognized (and pursued) by the Indian government at that, this is not a divide that can be merely addressed by training. Such training needs data — and that's data that we don't have. A research paper[17] published by Stanford University attributes the disparity in LLM training to the 'resourcedness gap' — unlike English, several of the world's languages are under-resourced and disadvantaged in terms of both quantity (insufficient quantity of both labelled and unlabeled data) and quality (absence of diverse resources.)

MODEL

And of the twenty two languages officially recognized by the government, data paucity is likely for a majority — but that is not to say that this is a lost cause. There are workarounds. The simplest (and potentially the avoidant) would be to integrate with machine translation technology ("As English language models are improving at an unprecedented pace, which in turn improves machine translation, it is from an empirical and environmental stand-point more effective to translate data from low-resource languages into English, than to build language models for such languages."[18]) The second — as an extension of the first — is to use machine translation to generate data to train. The third — and the ideal (the second approach, to the extent that it uses synthetic data suffers from both the reinforcement of quality deficiencies, and the omission of 'important local contextual knowledge and linguistic nuances.'[19]) — would be to assemble more labelled data. If generative artificial intelligence really is going to be all that it is made out to be, digitization of native languages may not be the choice that it has been all these years.

Expression emerges from deep cultural context — and even for technology that is designed to be as patronizing as AI, in the absence of such context, global (in its truest sense) penetration will fail. And there is evidence for the failure of expression (in its most artistic form, albeit)[20] —

> *Take Tamil, for example, a language spoken by over 78 million people and the official language of Sri Lanka and the Indian state of Tamil Nadu. When asked to write a poem in the traditional Tamil style of metered poetry called Venpa, ChatGPT's English version was a far better representation of the structure and phrasing typical of Venpa than its Tamil counterpart, despite being a style of poetry originating in Tamil.*

After all, like Turkish artist Refik Anadol said at the WEF Meeting in Davos, ***"How on Earth can we create an AI that doesn't know the whole of humanity?"***

DATA

Data has been the currency of the global economy for a while now — but its importance in an age where generation of statistically probable text has been made out to be singularly, critically important cannot be understated. While research and development in generative artificial intelligence is already hurtling towards the use of synthetic data (artificially generated information that mimics real-world data) as a workaround for data gaps (The Epoch AI research team projects, with an 80% confidence interval, that the current stock of training data will be fully utilized between 2026 and 2032.)[21] and privacy laws alike, use of synthetic data may come with significant downsides, including, not insignificantly but perhaps disturbingly, autophagy — 'a phenomenon (and a future?) where generative AI systems may increasingly consume their own outputs without discernment, raising concerns about model performance, reliability, and ethical implications.' [22] And that may mean opportunity for India — after all, for a country as large and diverse as ours, 'becoming the data capital' of the world may well be the lowest hanging fruit. Here's Niti Aayog's suggestion:

> *'By placing quality, trusted, and interoperable data at the core, India could become the data capital of the world and set new global benchmarks for breadth, depth, and quality of trusted data ecosystems. While this [AI Kosh] is a good foundation, scaling the breadth and depth of data could position it to move into high-value domains such as genomics, manufacturing telemetry, and cross sector financial data, implying that datasets are certified for quality, tagged for privacy, and interoperable.'*

As of the date of this report, AIKosh, the platform owned and operated by the Indian government under the AI mission to procure and host non-personal datasets, claims availability of more than 7,000 datasets. But while this is a good start, the real value lies in the regulated sectors (healthcare, banking, financial services and insurance) — and the regulatory boundaries may not permit the kind of collection, leverage and cross-border flow that it would take to really become the 'data capital.'

DATA

Which brings us back to this — is autophagy and model collapse a given when synthetic data is used? Stanford's AI Index Report[23] indicates otherwise.

> *Newer research suggests that when synthetic data is layered on top of real data, rather than replacing it, the model collapse phenomenon does not occur. [...] As the prevalence of synthetic data grows, particularly with an increasing share of web content being AI-generated, future models will inevitably be trained on non-human-generated material. One approach to expanding datasets is data augmentation, which modifies real data—such as tilting or image mixing—to create new variations while preserving essential characteristics.*

If there is indeed hope in that direction, then Ernst and Young's[24] suggestion of 'synthetic data' as a strategic choice may be sound:

> *Synthetic data offers a way to unlock the value in systems while protecting the people behind the data. For sectors where privacy is paramount and pressure to modernize is mounting, it is no longer a question of if but how fast can be the adoption. If India wants to build trustworthy, inclusive, and regulation- aligned AI, especially in high-stakes sectors, synthetic data could be the most underrated unlock.*

That said, like with many of the courses charted out for advancements in artificial intelligence, there are constraints with synthetic data as well, especially in fields like healthcare. An article published in The Lancet[25] summarizes it well:

> *This shift from a patient-focused approach to compositional data practices will raise new ethical and epistemological questions about how synthetic data represents the realities of health care.*[10]*Key considerations will include how synthetic data should be regarded and used compared with real-world data, how to generalise research outcomes based on synthetic data to patients, how to assess and evaluate synthetic data quality, and how to establish accountability.*

The question then, is this.

India has data. Can we make enough of it, quickly enough, within the sandboxes we've forged and must have?

REFERENCES

1. *Volenik, A. (2026, January 22). YahooFinance. Retrieved from https://finance.yahoo.com/news/microsoft-ceo-satya-nadella-warns-205620968.html*
2. *Robins-Early, N. (2026, January 25). The Guardian. Retrieved from https://www.theguardian.com/technology/ng-interactive/2026/jan/25/sam-altman-openai*
3. *John Collingridge and Graeme Wearden. (2026, January 21). The Guardian. Retrieved from https://www.theguardian.com/technology/2026/jan/21/rollout-ai-slowed-save-society-jp-morgan-jamie-dimon-jensen-huang*
4. *Rosario, J. D. (2026, January 19). Bloomberg. Retrieved from https://www.bloomberg.com/news/articles/2026-01-19/imf-warns-ai-trade-pose-risks-to-solid-global-growth-outlook*
5. *Nossel, S. (2026, March 2). I'm on the Meta oversight board. We need AI protections now. Retrieved from The Guardian: https://www.theguardian.com/commentisfree/2026/mar/02/meta-oversight-board-ai*
6. *AI Index Steering Committee, Institute for Human-Centered AI, Stanford University. (April 2025). The AI Index 2025 Annual Report. CA.*
7. *Competition Commission of India. (September, 2025). Market Study on Artificial Intelligence and Competition.*
8. *Niti Aayog, in partnership with McKinsey. (September, 2025). AI for Viksit Bharat| The Opportunity for Accelerated Economic Growth.*
9. *McKinsey and Company. (2025, April 28). The cost of compute: A $7 trillion race to scale data centers. McKinsey Quarterly.*
10. *Hires, J. (2026, January 26). Nvidia's 85% GPU Market Share Faces Growing Competition: Is This AI Stock Still a Buy for 2026?. Retrieved from YahooFinance*
11. *AFP. (2026, January 14). US allows Nvidia to send advanced AI chips to China with restrictions. The Hindu.*
12. *Magdalena Petrova, E. J. (2025, December 31). How $160 million worth of export-controlled Nvidia chips were allegedly smuggled into China. CNBC.*
13. *McKinsey and Company. (2025, April 28). The cost of compute: A $7 trillion race to scale data centers. McKinsey Quarterly.*
14. *Caballar, R. D. (n.d.). What are small language models? Retrieved from IBM Think: https://www.ibm.com/think/topics/small-language-models*
15. *Nvidia Research. (15 September, 2025). Small Language Models are the Future of Agentic AI.*
16. *Stanford University Human-Centered Artificial Intelligence. (2024, April 2022). Improving Equity and Access to Non-English Large Language Models. Retrieved from https://hai.stanford.edu/news/improving-equity-and-access-non-english-large-language-models*
17. *Stanford University | HAI. (April 22, 2025). Mind the (Language) Gap: Mapping the Challenges of LLM Development in Low-Resource Language Contexts.*
18. *Tim Isbister, F. C. (21 April, 2021). Should we Stop Training More Monolingual Models, and Simply Use Machine Translation Instead? ARXIV.*
19. *AI Index Steering Committee, Institute for Human-Centered AI, Stanford University. (April 2025). The AI Index 2025 Annual Report. CA.*
20. *Stanford University Human-Centered Artificial Intelligence. (2024, April 2022). Improving Equity and Access to Non-English Large Language Models. Retrieved from https://hai.stanford.edu/news/improving-equity-and-access-non-english-large-language-models*
21. *AI Index Steering Committee, Institute for Human-Centered AI, Stanford University. (April 2025). The AI Index 2025 Annual Report. CA.*
22. *Xiaodan Xing, F. S.-B. (15 May, 2024). When AI Eats Itself: On the Caveats of AI Autophagy. ARXIV.*
23. *AI Index Steering Committee, Institute for Human-Centered AI, Stanford University. (April 2025). The AI Index 2025 Annual Report. CA.*
24. *Ernst and Young. (n.d.). The AIdea of India: Outlook 2026.*
25. *Governing synthetic data in medical research: the time is now / Boraschi, Daniela et al. /The Lancet Digital Health, Volume 7, Issue 4, e233 - e234*