India’s big AI test is here: Making sovereign language models work

livemint.com/technology/ai-sovereign-language-models-sarvam-soket-gnani-ganai-indiaai-mission-genai-deepseek-openai-google-gemini-llama-krutrim-11750153684247.html

Leslie D'Monte, Shouvik Das June 17, 2025

Leslie D'Monte , Shouvik Das 9 min read

With over 1.4 billion people and 22 official languages, alongside thousands of dialects, India can ill afford to be an afterthought in the AI revolution. (iStockphoto)

Summary

Bengaluru/New Delhi: For years, the world’s most powerful artificial intelligence (AI) models have spoken in English. Trained on sprawling datasets like Wikipedia, Reddit, and Common Crawl, models such as OpenAI’s GPT-4, Google’s Gemini 2.5, Meta’s Llama, Microsoft’s Bing AI, and Anthropic’s Claude have mastered the dominant global internet dialect. But, they all falter when faced with the linguistic diversity of countries like India.

English-dominated AI models can hallucinate (fabricate facts), mistranslate key phrases, or miss the cultural context when prompted in Indian languages.

The concern is also over inclusion. With over 1.4 billion people and 22 official languages, alongside thousands of dialects, India can ill afford to be an afterthought in the AI revolution. The country is expected to total over 500 million non-English internet users by 2030. If AI models can’t understand them, the digital divide will only widen.

To address this, the Indian government launched a $1.2 billion IndiaAI Mission in February 2024. One of its central goals: to fund and foster the development of sovereign local language models and small language models (SLMs)—AI systems that are built, trained, and deployed entirely within India, on Indian data.

While large language models (LLMs), such as GPT-4, handle broad tasks, having been trained on copious amounts of data, SLMs are smaller, typically built for specific uses.

In January, the government opened a nationwide call for proposals to develop foundational AI models rooted in Indian languages and datasets. By April, more than 550 pitches had poured in from startups, researchers, and labs eager to build either SLMs or general-purpose LLMs.

In April, the government selected Sarvam AI to lead the charge. The Bengaluru-based startup will develop the country's first foundational model trained on local language datasets. It would build a massive 120-billion parameter open-source model to power new digital governance tools.

Parameters are settings that control how the AI model learns from data before making predictions or decisions. For instance, in a language model like ChatGPT, parameters help decide which word comes next in a sentence based on the words before it.

On 30 May, the government announced three more model-development efforts—from Soket AI, Gnani AI and Gan AI.

Soket AI, based in Gurugram, will build a 120-billion multilingual model focused on sectors like defence, healthcare, and education; Gnani AI, based in Bengaluru, will develop a 14-billion voice AI model for multilingual speech recognition and reasoning; Gan AI, also based in India's Silicon Valley, is working on a 70-billion parameter model aimed at advanced text-to-speech capabilities.

During the launch of the three additional models, union minister for electronics and information technology, Ashwini Vaishnaw, stressed the importance of more people being able to access technology and get better opportunities. "That's the philosophy with which IndiaAI Mission was created," the minister said.

A senior official from the ministry of electronics and information technology (MeitY), speaking on condition of anonymity, told Mint that a foundational sovereign language model can be expected within the next 12 months. "We will see many more sovereign models after a year or so, hosted on the government's AI marketplace platform," the official added.

Why it matters

Beyond the language gap, the global AI landscape is being shaped by rising concerns around sovereignty, data control, and geopolitical risk. As AI becomes the cornerstone of digital infrastructure, nations are racing to build their own models. In India, the move also aligns with India's broader vision of '*Atmanirbhar Bharat*' (self-reliant India).

India now joins a fast-growing club of countries that have developed or are developing sovereign LLMs—China (Baidu), France (Mistral), Singapore (SEA-LION), UAE (Falcon), Saudi Arabia (Mulhem), and Thailand (ThaiLLM).

View Full Image
Ashwini Vaishnaw, union minister for electronics and IT. (Photo: HT)

Even before Sarvam, India had seen an uptick in language model building activity. BharatGPT (by CoRover), Project Indus (Tech Mahindra), Hanooman (by Seetha Mahalaxmi Healthcare and 3AI), Krutrim (Ola), and Sutra (by Two AI) are some examples.

In October 2024, BharatGen, a government-backed project, released Param-1, a 2.9-billion parameter bilingual model along with 19 Indian language speech models. Led by IIT Bombay, BharatGen's mission is to boost public service delivery and citizen engagement using AI in language, speech, and computer vision.

Imagine a farmer in eastern Uttar Pradesh calling a helpline and interacting with a chatbot that understands and replies fluently in Bhojpuri, while also generating a clear summary for a government officer to act on. Or an AI tutor generating regional-language lessons, quizzes, and spoken explanations for students in languages like Marathi, Tamil, Telegu, or Kannada.

These efforts fit into India's broader digital stack, alongside Aadhaar (digital identity), UPI (unified payments interface), ULI (unified lending interface) and ONDC (the Open Network for Digital Commerce).

In a world where AI models are fast becoming a symbol of digital leadership, “a sovereign LLM is also about owning the narrative, the data, and the future of its digital economy", said Akshay Khanna, managing partner at Avasant, a consulting firm.

“Sovereignty will be a key requirement in all nations including India," says Mitesh Agarwal, Asia-Pacific managing director at Google Cloud. He points out that Google’s Gemini 1.5 processes data entirely within its India data centers. “For sensitive projects, we also offer open-source AI models and sovereign cloud options," he added.

Showing the way

Founded in July 2023 by Vivek Raghavan and Pratyush Kumar, Sarvam has raised $41 million from private investors. While the IndiaAI Mission won’t inject cash, it will take a minority equity stake in the startup.

View Full Image
File photo of Sarvam AI founders Vivek Raghavan (left) and Pratyush Kumar (right) with Nandan Nilekani (middle). (X)

For now, Sarvam will receive computing power—over 4,000 Nvidia H100 graphics processing units (GPUs) for six months—to train its model. The aim is to build a multimodal foundation model (text, speech, images, video, code, etc.) capable of reasoning and conversation, optimized for voice interfaces, and fluent in Indian languages.

“When we do so, a universe of applications will unfold," Sarvam co-founder Raghavan said at the launch on 26 April. “For citizens, this means interacting with AI that feels familiar, not foreign. For enterprises, it means unlocking intelligence without sending data beyond borders."

Sarvam is developing three model variants—a large model for “advanced reasoning and generation"; a smaller one for “real-time interactive applications", and “Sarvam-Edge for compact on-device tasks".

It is partnering with AI4Bharat, a research lab at the Indian Institute of Technology (IIT)-Madras, supported by Infosys co-founder Nandan Nilekani and his philanthropist wife Rohini, to build these models.

Sarvam has already developed Sarvam 1, a two-billion parameter multilingual language model, trained on four trillion tokens using Nvidia H100 GPUs.

The company claims its custom tokenizer (that breaks text into small units, like words or parts of words, so a language model can understand and process it) is up to four times more efficient than leading English-centric models when processing Indian languages, hence reducing costs.

Sarvam 1 supports 11 languages: Hindi, Bengali, Tamil, Telugu, Kannada, Malayalam, Marathi, Gujarati, Oriya, Punjabi, and English. It powers various generative AI (GenAI) agents and is also hosted on Hugging Face, enabling developers to build Indic-language apps.

Sarvam has already developed Sarvam 1, a two-billion parameter multilingual language model, trained on four trillion tokens using Nvidia H100 GPUs.

Hugging Face is a platform for sharing and hosting open-source AI models and datasets.

Gnani.ai, meanwhile, is building voice-to-voice foundational LLMs that aim to produce near instant autonomous voice conversations, with very low latency. The models also aim to enable “emotion aware conversations", which preserve intonation, stress and rhythm in the conversations, said Ganesh Gopalan, co-founder and CEO of Gnani.ai. “The model will enable realistic conversations in governance, healthcare and education," he added.

Wait and watch

Sovereign LLMs and SLMs are likely to find strong acceptance in public service delivery and citizen engagement services across the country, just like it happened with UPI. However, enterprises will likely wait till the models show maturity, are secure enough, and hallucinate less.

Current sovereign models, Sanchit Vir Gogia, founder of Greyhound Research explained, “lack deployment maturity, robust safety mechanisms, and domain-specific accuracy."

The Greyhound CIO Pulse 2025 survey found that 67% of enterprises exploring Indic LLMs report frequent failures in multilingual task execution, especially with mixed scripts (e.g., Devanagari+ Latin), identifying regional slang, or recognizing emotional cues in customer queries.

Further, language in India is hyper-local. Hindi spoken in Varanasi differs significantly from Hindi in Patna—not just in accent, but in vocabulary and usage. A health insurance aggregator in Bengaluru faced real-world fallout when its LLM couldn't differentiate between '*dard*' (pain) and '*peeda*' (suffering), leading to claim errors. The company had to halt rollout and invest in regionally-tuned data, Gogia said.

Moreover, there are limited safeguards against hallucinations. "Without deeper fine-tuning, cultural grounding, and linguistic quality assurance, these models are too brittle for nuanced conversations and too coarse for enterprise-scale adoption," Gogia added. "The ambition is clear—but execution still needs time and investment."

The missing millions

Building sovereign models without government or venture capital funding could also pose a big challenge since developing a foundational model from scratch is an expensive affair. For instance, OpenAI's GPT was in the works for more than six years and cost upwards of $100 million and used an estimated 30,000 GPUs.

Chinese AI lab DeepSeek did build an open-source reasoning model for just $6 million, demonstrating that high-performing models could be developed at low costs. But critics point out that the reported $6 million cheque would have excluded expenses for prior research and experiments on architectures, algorithms, and data.

Building sovereign models without government or venture capital funding could also pose a big challenge since developing a foundational model from scratch is an expensive affair.

Effectively, this means that only a lab which has already invested hundreds of millions in foundational research and secured access to extensive computing clusters could train a model of DeepSeek's quality with a $6 million run.

Ankush Sabharwal, founder and CEO of CoRover, says that its BharatGPT chatbot is a "very small sovereign model with 500-million parameters". He has plans to build a 70-billion parameter sovereign model. "But, we will need about $6 million to build and deploy it," Sabharwal says.

Long way to go

A glance at the download numbers for the month of May from Hugging Face underlines the wide gap between some of India's local language models and similar-sized global offerings.

For instance, Sarvam-1's 2-billion model saw just 3,539 downloads during the month. Krutrim, a 12-billion model from Ola-backed Krutrim SI Designs, fared similarly with only 1,451 downloads. Fractal AI's Fathom-R1 14-billion model showed the most promise with 9,582 downloads.

In contrast, international models with comparable or slightly larger sizes saw exponential traction. Google’s Gemma-2 (2-billion) logged 376,800 downloads during the same period, while Meta’s Llama 3.2 (3-billion) surpassed 1.5 million. Chinese models, too, outpaced Indian counterparts. Alibaba’s Qwen3 (8- billion) recorded over 1.1 million downloads, while a fine-tuned version of the same model—DeepSeek-R1-0528-Qwen3-8B—clocked nearly 94,500 downloads.

The numbers underline the need for a stronger business case for Indian startups.

The senior government official quoted earlier in the story said that sovereign models must stand on their own feet. “The government has created a marketplace where developers can access and build apps on top of sovereign models. But the startups must be able to offer their services first to India, and then globally," he said.

“API revenue, government usage fees, and long-term planning are key," Aakrit Vaish, former CEO of Haptik and mission lead for IndiaAI until March, said.

API revenue is what a company earns by letting others use its software features via an application programming interface. For example, OpenAI charges businesses to access models like ChatGPT through its API for writing, coding, or image generation.

Nonetheless, API access alone won’t cover costs or deliver value, Gogia of Greyhound Research said. “Sovereign LLM builders must focus on service-led revenue: co-creating solutions with large enterprises, developing industry-specific applications, and securing government-backed rollouts," he suggested.

Indian buyers, he added, want control—over tuning, deployment, and results. “They’ll pay for impact, not model access. This isn’t LLM-as-a-Service; it’s LLM-as-a-Stack."

In short, capability alone won’t cut it. To scale and endure, sovereign language models must be backed by viable business propositions and stable funding—from public and private sources alike.