A Strategic Framework for Domain-Specific Small Language Models (SLMs) in India

Soumya Banerjee

Abstract

While the world is obsessed with the "AI arms race" to build the biggest models possible, India is quietly proving that smaller is often better. There has been a strategic shift toward Small Language Models (SLMs)—think of them as "Pocket Field Guides" rather than "Massive Encyclopedias." 📘✨

Here’s why this is a game-changer for India:

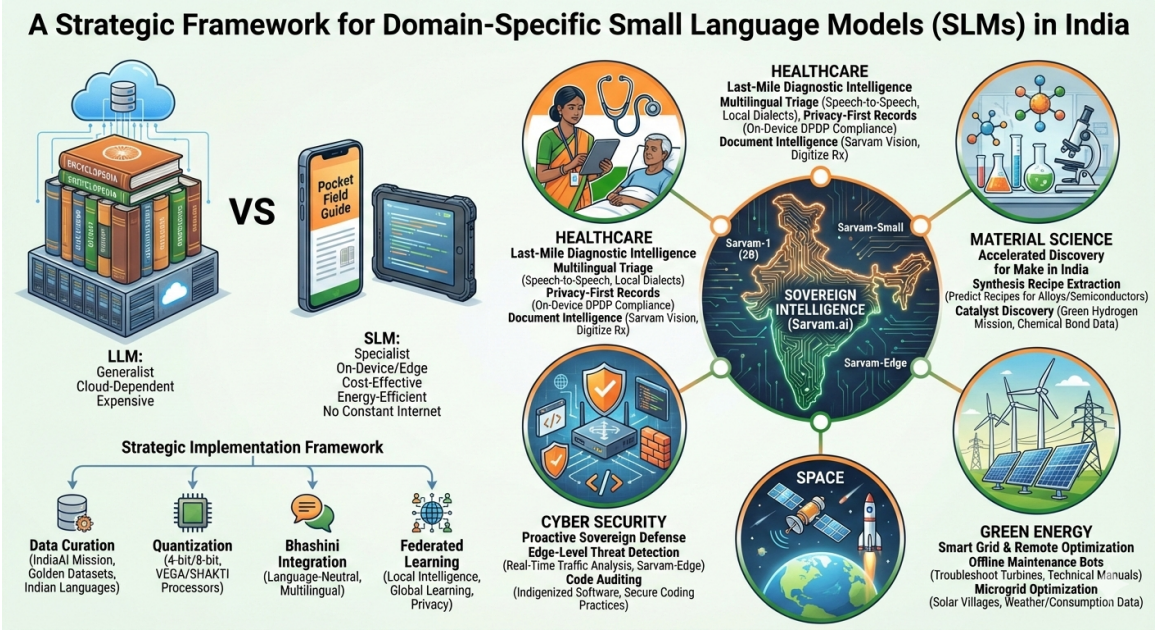
95% of the Profit: The real value isn't in building one giant brain; it’s in building millions of specialized tools that actually solve problems.

No Internet? No Problem: These models can run on a simple tablet in a rural clinic or a remote solar farm without needing a massive cloud connection.

Sovereign Intelligence: By using models like those from Sarvam.ai, we keep our data—and our solutions—local.

Where will we see the impact? 🏥 Healthcare: Helping village health workers triage patients in local dialects. ⚡ Green Energy: Managing smart grids in remote "Solar Villages." 🚀 Space: Processing satellite data right there in orbit. 🛡️ Cyber Security: Detecting threats instantly at the source.

The future of AI in India isn't just about being "smart"—it's about being practical, affordable, and inclusive. 🇮🇳



Overview

India’s strategic pivot toward Small Language Models (SLMs) like those from Sarvam.ai marks a significant shift from "all-purpose" AI to "purpose-built" sovereign intelligence. In a country with the scale of India, SLMs are often more practical than Large Language Models (LLMs) because they are cost-effective, energy-efficient, and can run on-device (at the "edge") without needing constant high-speed internet.

One can think of the difference between Large Language Models (LLMs) and Small Language Models (SLMs) as the difference between a massive, multi-volume encyclopedia and a pocket-sized field guide for a specific job. LLMs (like ChatGPT) are generalists—they are incredibly smart but require massive supercomputers, cost millions of dollars to run, and need a constant high-speed internet connection. In contrast, SLMs (like those from Sarvam.ai) are specialists—they are much smaller, faster, and designed to do one or two specific tasks exceptionally well, such as diagnosing a crop disease or processing a loan in a local language.

For a country like India, the "95% of the profit" doesn't lie in building the world's largest general-purpose model, which is an expensive "arms race" for tech giants. Instead, the real value lies in the application layer: building millions of affordable, energy-efficient SLMs that can run directly on a farmer's smartphone or a village clinic's tablet without needing a cloud. By focusing on these "nimble" models, India can solve massive social problems at a fraction of the cost, turning AI from a luxury high-tech toy into a practical, profitable tool for the everyday citizen.

As of early 2026, Sarvam AI has released models like Sarvam-1 (2B) and Sarvam-Small,

specifically optimized for Indian contexts. Here is an outlined approach for how India can deploy these domain-specific models across your requested sectors.

1. Healthcare: Last-Mile Diagnostic Intelligence

In India, the primary challenge is the doctor-to-patient ratio in rural areas. SLMs can bridge this gap by running on low-cost tablets or smartphones used by ASHA workers.

- Multilingual Triage: Using Sarvam's speech-to-speech capabilities to allow patients to describe symptoms in local dialects (e.g., Bhojpuri or Gondi). The SLM extracts clinical entities and flags high-risk cases for doctors.
- Privacy-First Records: Because SLMs can run locally, sensitive patient data never has to leave the clinic or the device, ensuring compliance with India's Digital Personal Data Protection (DPDP) Act.
- Document Intelligence: Deploying Sarvam Vision to digitize and summarize handwritten prescriptions or old medical reports, making them searchable for longitudinal care.

2. Material Science: Accelerated Discovery for Make in India

Material science requires processing massive amounts of unstructured scientific data. SLMs can act as specialized "research assistants" for labs like CSIR.

- Synthesis Recipe Extraction: Similar to the MIT DiffSyn approach, Indian researchers can use SLMs trained on metallurgical and chemical journals to predict the best temperature and pressure "recipes" for new alloys or semiconductors.
- Catalyst Discovery for Green Hydrogen: SLMs can be fine-tuned on chemical bond data to suggest potential new catalysts, significantly shortening the R&D cycle for India's National Green Hydrogen Mission.

3. Cyber Security: Proactive Sovereign Defense

In cyber security, latency is the enemy. SLMs allow for real-time analysis at the network edge rather than waiting for cloud-based threat detection.

- Edge-Level Threat Detection: Deploying Sarvam-Edge on government routers and defense hardware to analyze traffic patterns in real-time for "Zero-Day" anomalies.
- Code Auditing for Indigenized Software: SLMs trained on secure coding practices can be used by Indian startups to auto-audit code, ensuring no backdoors exist in software being integrated into the India Stack.

4. Green Energy: Smart Grid & Remote Optimization

India’s renewable energy infrastructure is often located in remote areas (Rajasthan's solar parks or Tamil Nadu’s wind farms) with limited connectivity.

- Offline Maintenance Bots: SLMs on ruggedized handhelds can help technicians troubleshoot turbine failures by querying technical manuals and sensor data without an internet connection.
- Microgrid Optimization: Using SLMs to process local weather patterns and consumption data to optimize the distribution of solar energy in "Solar Villages".

5. Space: Telemetry & Scientific Retrieval

For ISRO and the growing private space sector (SpaceTech), SLMs can handle the high-velocity data generated by satellites.

- Telemetry Summarization: Instead of sending massive raw logs to Earth, an SLM on a satellite could summarize health status and only transmit critical anomalies, saving bandwidth.
- Knowledge Retrieval: A specialized "ISRO-GPT" (based on Sarvam-M architecture) trained on decades of Indian space missions to help young engineers quickly retrieve specific propulsion or orbital mechanics data.

The Strategic Implementation Framework

To make this work, the approach should follow four "Pillars of Sovereignty":

- Data Curation: Use the IndiaAI Mission to create "Golden Datasets" for these 5 domains, with at least 20% content in Indian languages.
- Quantization: Apply 4-bit or 8-bit quantization so these models run on the VEGA/SHAKTI indigenous processors.
- Bhashini Integration: Ensure every domain-specific SLM is "Language-Neutral," allowing a scientist in Kerala to query a Material Science model in Malayalam.
- Federated Learning: Train models across different hospitals or energy plants without sharing the raw data, keeping the intelligence local but the learning global.

Closing

By focusing on domain-specific SLMs and combining the above pillars, India can build a resilient, sovereign AI ecosystem that prioritizes accessibility, privacy, and real-world impact over trophy-model size.