सत्यमेव जयते

Office of the Principal Scientific Adviser
to the Government of India

ADVANCING INDIGENOUS FOUNDATION MODELS

March 2026

India’s AI Policy Priorities White Paper Series

Acknowledgement

This white paper has been prepared with inputs and feedback from domain experts, stakeholders, and colleagues during various stages of its development. We are grateful for the advice, reviews and inputs from Dr. Preeti Banzal, Adviser/Scientist 'G', Office of Principal Scientific Adviser to the Government of India; Mr Ankush Sabharwal, Founder of CoRover Pvt. Ltd.; Professor Sunita Sarawagi, IIT Bombay; and Professor Mayank Singh, IIT Gandhinagar. The views expressed do not necessarily reflect those of the reviewers, contributors, or their respective institutions.

About: White Paper Series
Emerging Policy Priorities for India’s AI Ecosystem

To foster informed deliberation and action among stakeholders engaged in shaping India's artificial intelligence (AI) policy and governance landscape, the Office of the Principal Scientific Adviser to the Government of India is producing this White Paper Series. These papers are conceived as explanatory briefs that examine specific policy issues and their associated nuances, with the aim of enabling broader understanding and meaningful societal engagement. The White Papers are developed by drawing on collective insights from the extended AI ecosystem, including inputs from multi-stakeholder consultations, bilateral and multilateral AI policy engagements, and subsequent expert reviews. They are intended solely as explanatory documents that highlight identified policy priorities and stimulate further discussion. The views presented in these white papers should not be construed as formal policy positions of the office.

Prepared by:

Mr. Animesh Jain, Senior Policy Fellow
Mr. Kunal Thakur, Policy Analyst

Office of the Principal Scientific Adviser to the Government of India

Table of Contents

1. Introduction

Foundation models (FM) are a core enabling layer in modern AI systems because they can be adapted for many applications, reducing the need to train separate models from scratch for_each task. It makes them transformative, but it also concentrates influence; choices made at the model's design and training stage can shape performance and risks across many downstream uses, affecting multiple sectors and services [1]. It raises the implications for sovereignty and inclusivity. Relying solely on foreign models risks under-representation of Indian languages and cultural contexts. Any biases in these models can cascade across all downstream applications that rely on them. This makes it critical to have a policy focus on these systems. All governance specifications need to be sufficiently broad, as they will reverberate across all sectors, and each sector may adopt and interpret those specifications as per its unique context and technical needs [2]. Dependence on foreign models limits India's ability to ensure transparency, inclusion, and alignment with national priorities. By investing in indigenous foundation models trained on more diverse data, designed for India's linguistic and social diversity, and governed through national frameworks, India can build AI systems that are less biased, trustworthy, and locally relevant, strengthening its technological autonomy amid a globally competitive AI ecosystem.

The India AI Governance Guidelines' report defines a foundation model as a large AI model trained on massive amounts of data (text, images, video, and audio) that can be used with or without fine-tuning for many downstream tasks [3]. A foundation model could work horizontally across domains (e.g., multilingual language models for banking, healthcare, or education) or vertically for specific sectors (e.g., diagnostic models for medicine). These models are pretrained on large-scale, largely unlabelled data to learn general representations and are then fine-tuned for over several tasks such as text translation, summarisation, question answering, or text classification. They may be unimodal (e.g., text) or multimodal (e.g., text with images, audio, or video) [4]. This versatility makes them a critical layer of today's AI ecosystem and a key area for innovation in India. Therefore, developing indigenous foundation models is a strategic priority. India's objective is to harness foundation models for inclusive growth and public good, while ensuring they are governed in a manner consistent with the country's values, legal framework, and security interests. This white paper tries to give an understanding of India's approach to advancing indigenous foundation models through public–private collaboration and to governing these systems that support trust, accountability, and responsible adoption.

Within this broader approach, India's requirements also need a specific emphasis on small language models (SLMs) with multimodal capabilities. SLMs are more focused, domain-orientated models that It can be fine-tuned for sector-specific tasks and is typically more economical to run and maintain. In India's context, dedicated SLMs (supervised learning models) for agriculture, health, education, and MSMEs (micro, small, and medium enterprises) can deliver high accuracy on local tasks with lower data and energy requirements. In practice, foundational LLMs can form an enabling layer over which domain-specific SLMs could be developed and fine-tuned for particular use cases, departments, and sectors. This combination of LLMs, SLMs, and multimodal models aligns with India's priorities of linguistic inclusion, affordability, energy efficiency, edge deployment, public-sector suitability, and sectoral innovation across areas such as agriculture, health, education, climate, and urban governance.

2. Building Indigenous Foundation Models

At present, many widely used foundation models (FM) in India are developed abroad and trained on datasets that insufficiently reflect India's linguistic and cultural diversity. Against this backdrop, India is building and developing its indigenous FM. This section tries to map this development through government and industry efforts. The government of India recognises indigenous FM as critical to its national AI infrastructure. These efforts include funding the development of large-scale multimodal models and building shared compute and data resources for research and startups. Since training and adapting the foundation model required sustained access to high-end compute and large, high-quality datasets, the government's approach is to build a shared compute and data ecosystem that can be accessed by startups, academic institutions, and public-sector entities. The shared infrastructure of Compute and Data is intended to reduce the cost and barriers of entry for Indian developers, expand participation beyond a few large players, and enable foundation models to be trained and fine-tuned on India-relevant datasets. In this effort, public investment in shared compute and data platforms serves as an enabling layer for indigenous foundation models, supporting both large and downstream task-specific models, while strengthening domestic capacity, resilience, and long-term technological autonomy.

The IndiaAI Mission, approved by the Union Cabinet in March 2024 with an outlay of ₹10,371.92 crore over five years and led by the Ministry of Electronics and Information Technology (MeitY), is the Government of India's primary initiative for strengthening domestic capabilities to build and deploy indigenous AI systems, including foundation models [5]. Developing foundation models requires substantial data and compute resources, which the mission addresses through two dedicated pillars, the India AI Compute Portal and the AI-Kosh platform.

2.1 Innovation Initiatives for Indigenous Foundation Models

India's indigenous FM ecosystem is being built through a set of initiatives that combine public support with private and academic innovation. Under the IndiaAI Mission, the government has used competitive selection processes to identify multiple teams to build foundation models trained on India-specific data.

The IndiaAI Mission released a Call for

Proposals for foundational AI models in January 2025 to invite startups, researchers, and entrepreneurs to develop large multimodal models, large language models, and small language models aligned to India-specific needs [7]. The Call received 506 proposals by April 2025, and selections were made through expert evaluation [8]. In the first phase of approvals, four initiatives were selected under the IndiaAI Foundation Models pillar : Sarvam AI, Soket AI, Gnani AI, and Gan AI [9]. They are developing a sovereign LLM ecosystem spanning multilingual text capabilities, voice AI, and advanced text-to-speech, respectively. Further, the second phase was announced in September 2025, and eight additional foundation-model initiatives were launched under the same pillar to build indigenous AI and large and small language models based on Indian datasets that span all 22 scheduled Indian languages. The detailed table with the names of the selected organisations and consortia, including startups, industry players, and academic institutions, is mentioned in Annexure 1. The selected projects cover multilingual foundational models, speech and voice models, multimodal AI, scientific models, healthcare reasoning systems, and agentic AI platforms. The above models are expected to be made available through the AIKosh platform, enabling startups and research institutions to access and build upon them. This venture is intended to strengthen the open ecosystem and accelerate innovation across India's AI community.

At the recent AI Impact Summit in New Delhi, several major indigenous models were launched. They showcased the breadth of India's emerging foundation-model capability. Like, Sarvam AI announced Sarvam-105B, trained from scratch and optimised for Indic language performance. Gnani.ai launched Inya VoiceOS, a voice-to-voice model is designed to process audio directly and reduce speech-to-text latency across more than 15 languages. Fractal launches Vaidya 2.0, a reasoning model optimised for complex medical diagnoses and STEM tasks. Tech Mahindra, in partnership with NVIDIA, introduced Project Indus (8B), a Hindi-first model focused on education and culturally relevant learning. These launches indicate that India's indigenous FM ecosystem is progressing in different directions, such as frontier-scale multilingual text models, voice-native speech systems, multimodal public-interest models, and domain-specialised reasoning models.

Further to these efforts, initiatives like BharatGen, led by IIT Bombay, have also released sovereign models across text and speech. This includes Param-1, a 2.9B-parameter core text model, along with

Shrutam for speech recognition, Sooktam for text-to-speech and Patram for document comprehension in Indian formats. It is also extending its model base into domain-specialised variants, including models such as AyurParam, to improve performance for public-interest and sector-specific workflows. BharatGen also released BharatGen v1, which is aligned to 22 languages and initially oriented to public-interest domains such as heathcare and agriculture. Soket AI is developing Project EKA as a large multilingual sovereign LLM initiative in the 120B range. Alongside these efforts, 'Tech Mahindra Makers Lab is developing Project Indus as an 8B Indic language model with an initial emphasis on Hindi and more than 37 dialects, while TechM Orion has been launched as an agentic AI platform for enterprise workflows. Together, these initiatives show that India's roadmap spans both capability-maximising large models and deployment-ready efficient models.

Apart from these public–private coordinated efforts, several private-sector players are also developing deployment-efficient small language models tailored for sector-specific use cases and trained on India-relevant datasets. For example, Zoho has released its in-house Zia LLM in 1.3B, 2.6B, and 7B parameter variants designed for enterprise workflows (such as extraction, summarisation, and RAG-style tasks) and integrated into its product ecosystem. CoRover.ai's BharatGPT has also contributed to these efforts. They are developing multilingual, deployment-orientated models for conversational AI. They have released the BharatGPT-mini as a 0.5B model and BharatGPT-3B-Indic as a 3B model. BharatGPT-3B-Indic is a multilingual model trained on Indian conversational data in 12 languages and released in BF16 format. The above initiatives showcase India's roadmap, which spans both capability-maximising large models and deployment-ready efficiency models.

2.2 Compute

Foundation models require sustained access to high-end compute. Under the mission's compute pillar, the government is building a scalable national compute ecosystem through public–private partnerships and an AI marketplace for access to compute and pre-trained models. The IndiaAI Compute Portal is operationalising shared access to high-end compute by providing compute-as-a-service to startups, academia, and public institutions. A significant share of Mission funding, ₹4,563.36 crore, is earmarked for compute capacity over five years. By December 2025, over 38,000 GPUs had been onboarded and were being offered at a subsidised rate of ₹65 per hour, expanding access for startups, academia, and public-sector use cases. This pillar is intended to ensure broad-based access for researchers, academia, startups, MSMEs, early-stage innovators, and government entities. The scale of this initiative marks a substantial shift in India's AI readiness. User onboarding data from IndiaAI illustrates the diversity of beneficiaries: 114 researchers and academic institutions, 47 startups and MSMEs, 8 IndiaAI Fellows, 32 students, 36 early-stage startups, 10 early-stage researchers, and 58 government entities had been onboarded by January 2026. These numbers demonstrate that the compute pillar is reaching both established institutions and emerging innovators. This compute push is reinforced by the India Semiconductor Mission (ISM), which aims to strengthen domestic capability across chip design and the wider semiconductor value chain, improving supply-chain resilience and long-term availability of AI hardware [6].

In addition to expanding public access to compute, the government also shapes the role of private cloud providers through empanelment, audit, procurement conditions, and compliance requirements. As set out in MeitY's Guidelines for Procurement of Cloud Services, these conditions help ensure that participating providers meet defined standards relating to data security, accessibility, and service delivery in support of AI development. The framework covers issues such as data location, legal compliance, security, service levels, and exit management [32]. Private cloud providers are also subject to broader cybersecurity and data-governance obligations, including CERT-In directions on log retention and incident-related compliance, and the data-processing obligations applicable under the Digital Personal Data Protection Act, 2023 [33].

2.3 Data

Data is the fuel for AI, and foundation models in particular require sustained access to large, high-quality datasets to train and fine-tune systems that perform reliably across languages and real-world contexts. Therefore, under the IndiaAI Mission, the government created the AIKosh dataset platform to strengthen India's data and model foundation by providing a shared national platform that reduces duplication, improves dataset quality, and enables systematic evaluation. AIKosh is a unified repository of datasets, AI models, and validated use cases, with sandbox capabilities to support testing, development, and benchmarking. Here, data should be understood not only as raw material for model training but also as a critical input for benchmarking, evaluation, and continuous performance monitoring. Accordingly, identifying compelling use cases and developing use-case-specific benchmarks to assess model quality and track progress should also be given importance. As of February 2026, it hosts over 10,021 datasets and 279 AI models across 20 sectors. By expanding access to large-scale and high-quality datasets, AIKosh is intended to support foundation model training and fine-tuning that better reflects India's linguistic and cultural diversity, while also enabling more systematic evaluation and risk mitigation. Many in academia and industry have been actively working to contribute to the AIKosh platform for boosting efforts on large-scale model training. Some specific efforts undertaken include Soket AI's EKA Pretraining Indic Corpus (v1) and IIT Gandhinagar's Triveni.

3. Governing India's foundational models- current landscape

The governance for FM in India is being shaped through a combination of legal, regulatory and technical instruments such as (i) cross-cutting AI governance guidance, (ii) data protection law, (iii) intermediary due diligence rules for online platforms and AI-enabled content, (iv) copyright and IP policy pathways relevant to training data, and (v) benchmarks that translate high-level principles into technical requirements. They create relevant obligations at different points in the foundation-model lifecycle, like data sourcing, training and fine-tuning, deployment through products and platforms, and post-deployment monitoring and remediation.

3.1 India AI Governance Guilines(2025)

The India AI Governance Guidelines (2025) emphasise accountability across the AI value chain. While the Guidelines are not an "AI law", they articulate governance expectations that can be operationalised through sector-specific regulatory directions and platform compliance processes. Internationally, similar value-chain framing is also reflected in multilateral and jurisdictional approaches. For instance, the OECD AI Principles emphasise responsible stewardship, transparency, robustness, and accountability across the AI system's lifecycle [10].

Accountability across the value chain: For FM, accountability cannot be limited to the downstream application developer because upstream design and training choices influence behaviour across many deployments. Therefore, the guidelines emphasise clarifying roles and responsibilities across actors and supporting compliance through documentation and governance artefacts. This is important for FM providers because they sit upstream of multiple deployers who may operate under different sectoral rules. Similar obligations are also being formalised in other jurisdictions. For example, under the EU AI Act, providers of general-purpose AI models also need to maintain technical documentation and make relevant information and documentation available to downstream providers that intend to integrate the model into their AI systems. However, it does not apply to models released under a "free and open-source licence", except where the model is classified as a general-purpose AI model with systemic risk [11].

Value-chain transparency: The Guidelines position transparency as a prerequisite for accountability. It emphasises that effective governance requires visibility across the AI value chain, including how an AI system is designed, which actors participate at different stages, the relationships among these actors, and the movement of key inputs and resources, such as data, across development and deployment. It further acknowledges transparency reports as a practical instrument to strengthen accountability. It also notes that such reporting can include disclosure of risk-related information, such as red-teaming outcomes, impact assessments, and mitigation measures to support scrutiny and build trust in deployed AI systems. The Guidelines also recommend a graded approach to accountability, under which responsibilities and liabilities are proportionate to an actor's role in the AI value chain, the characteristics of the AI system, and the level of risk involved. This approach is intended to align governance expectations with the realities of how AI systems are developed, deployed, and used across sectors.

3.2 DPDP(Digital Personal Data Protection Act)

The DPDP Act, 2023, is the primary legal framework that is applicable wherever a foundation-model pipeline processes digital personal data, including during dataset creation, training/fine-tuning, retrieval workflows, and inference-time logging [12]. The DPDP framework links processing to a specified purpose and requires limiting processing to what is necessary for that purpose. If personal data is present in training or fine-tuning data, the organisation must be able to justify the lawful basis and necessity and maintain controls for purpose and limitation. The Act requires data fiduciaries to implement reasonable security safeguards to prevent personal data breaches. Training datasets should be treated as protected assets with proportionate access controls, security measures, and breach response processes.

General data-protection laws also address the governance of personal data in AI systems internationally. In the European Union, GDPR (General Data Protection Regulation) obligations apply whenever personal data is processed, including in data collection, training, fine-tuning, and deployment pipelines [13]. China follows a similar approach through the Personal Information Protection Law (PIPL), which establishes duties for the processing of personal information. It acts as China's primary data legislation and establishes strict compliance obligations that directly impact the lifecycle of foundation model development. It functions as the country's primary personal-data legislation and establishes compliance obligations that affect the foundation-model lifecycle, from data collection and training to deployment. The law centres lawful processing on informed consent, which makes it legally risky to train models on unconsented personal data collected from open-web sources [14]. The law also reinforces data sovereignty through data localisation and cross-border transfer constraints for entities processing large volumes of personal information, which can materially limit the offshore movement of datasets used in large-scale or global training workflows [15].

3.3 IT Rule (Intermediary Guidelines and Digital Media Ethnics Code) Rules, 2021

The IT rules become relevant when foundation models are deployed through intermediary services (e.g., generative media tools integrated into social platforms or services enabling users to create and share content). The proposed amendment to the IT Rules, 2021, includes due diligence obligations, including provisions addressing harmful manipulated content (for example, obligations connected to the removal of content depicting individuals in certain altered or "morphed" forms) [16]. MeitY has issued a proposed amendment focused specifically on "synthetically generated information".

The proposal introduces a formal definition of synthetically generated information and establishes a set of due-diligence requirements for its identification and disclosure. It mandates the labelling and embedding of metadata, including permanent unique identifiers, for synthetic or modified content, and requires that such labels be displayed or made audible in a clear and prominent manner. The proposal further places obligations on Significant Social Media Intermediaries to obtain user declarations regarding whether content is synthetically generated and to adopt reasonable and appropriate technical measures to verify these declarations. It also prohibits the removal, suppression, or alteration of labels or identifiers associated with synthetically generated information. These measures aim to enhance transparency, traceability, and accountability in the production and distribution of AI-generated content. Even when a foundation model is not itself an "intermediary", the compliance burden falls on deployers and platforms that integrate it.

'Chinas Measures for the Labelling of Artificial Intelligence-Generated Content codifies a strict dual-labelling; it mandates that service providers must add explicit labels, presented in the form of text, sound, or graphics that are clearly perceptible to users, and implicit labels, embedded in the file's metadata and not easily perceived [17]. The measures impose a verification duty on online content platforms: if a platform detects these implicit labels or other traces of synthetic generation, it is legally compelled to add prominent labels to inform the public that the content is AI-generated, regardless of the user's claim. While the 'European Unions AI Act

(Article 50) adopts a transparency approach, requiring machine-readable formats primarily to protect consumer rights rather than for state tracking. India distinguishes its framework by adding a unique procedural layer: unlike the purely technical mandates of the EU or the automated enforcement of China, India's requirement for Significant Social Media Intermediaries (SSMIs) to actively validate "user declarations" distributes the accountability for content authenticity between the human creator and the hosting platform [18].

3.4 Balancing AI Innovation & Intellectual Property:

The India AI Governance Guidelines 2025 also recognise that copyright and content ownership are important aspects for generative AI and foundation models because AI models are often trained on large collections of publicly available data, and various lawsuits have been filed claiming that such practices constitute infringement. It describes copyright as a contested issue in AI governance, noting strong and divergent views on how legal frameworks can protect creativity without stifling innovation. It also notes that, under Section 52 of the Copyright Act, limited fair dealing exceptions apply for private or personal use, including research; however, these exceptions are restricted to non-commercial use and do not extend to organisational or institutional research. As a result, these exceptions may not cover many types of modern AI training, particularly those that involve large-scale data processing and machine learning models that require access to diverse datasets beyond personal use. The India AI Governance Guidelines 2025 suggested to the committee set up by DPIIT to consider a “balanced approach” that enables “Text and Data Mining (TDM)” with the objective of fostering innovation while also enabling provisions to protect the rights of copyright holders.

In continuation of that, DPIIT has released Part I of a working paper on the AI-copyright interface as a consultative document. The working paper assesses existing approaches, including blanket exemptions, text and data-mining exceptions with or without an opt-out right, voluntary licensing, and extended collective licensing. Owing to suitability concerns with these approaches, it proposes a policy framework aimed at striking a balance between the rights of content creators and AI innovators. The Committee does not endorse the 'zero price licence model', arguing that this would undermine incentives for human creativity and could lead to long-term underproduction of human-generated content. As an alternative, the Committee proposes a "hybrid model" under which:

- AI developers receive a "blanket licence" for the use of all lawfully accessed content for training purposes, without requiring individual negotiations.
- Royalties become payable only upon commercialisation of the AI tools, with rates set by a government-appointed committee and subject to judicial review; and
- A centralised mechanism handles royalty collection and distribution, aiming to reduce transaction costs, provide legal certainty, and support equitable access for both large and small AI developers.

This proposed framework offers a distinct alternative to existing global standards, unlike Singapore, which prioritises speed by granting a broad legal exception for computational data analysis, allowing developers to train on lawfully accessed data without permission. India seeks a middle path [19]. It also avoids the rigid restrictions seen in China, where the regulatory model mandates that all training data must come from legitimate sources [20]. This position was reinforced by the 'Guangzhou Internet Courts "Ultraman" ruling (2024), which held an AI service provider liable for copyright infringement because it generated images substantially similar to the copyrighted character [21]. It has also underscored platform liability and a "reasonable duty of care" in relation to infringing outputs, reinforcing the compliance significance of governance controls at the service layer. India's hybrid model attempts to balance these extremes: it grants a blanket license to ensure developers have immediate access to lawfully acquired data while mandating royalty payments upon commercialisation to ensure creators are compensated (unlike the pro-innovation exemptions in Singapore).

The India AI Governance Guidelines 2025 report and DPIIT Working Paper establish that copyright issues for generative AI systems trained on large-scale data are being actively addressed as a governance priority. It indicates that India is developing a framework to balance AI innovation with copyright protection for rights holders. in Singapore).

3.5 Developing India's Centric Benchmarks:

Benchmarking is crucial for the governance of foundation models, as it provides an objective, evidence-based way to evaluate and regulate their performance. Without such benchmarks, regulators cannot meaningfully compare models or set thresholds for acceptable performance. Benchmarks designed around specific governance priorities, such as fairness across languages and demographics, enable regulators to systematically test these aspects and set expectations. So benchmarking turns high-level principles like fairness, accountability, and transparency into measurable metrics that can be monitored.

The India AI Governance Guidelines 2025 encourage the use of benchmarks and evaluation frameworks to systematically assess AI systems against ethical, safety, and performance standards, enabling objective measurement of compliance with principles, such as fairness, robustness, and explainability. They highlight that India's Bureau of Indian Standards (BIS) is actively collaborating with industry stakeholders and academic institutions to develop AI-specific standards and testing protocols that are consistent with global frameworks, such as those established by ISO/IEC. Currently, India's benchmark ecosystem is developing through a mix of:

(i) Public digital platforms that institutionalise evaluation,
(ii) Academic benchmarks for Indic-language capability and risk, and
(iii) Domain and multimodal benchmarks aligned to India's public-service and market needs.

Digital India Bhasini and ULCA (Universal Language Contribution APIs) are developing a language technology evaluation stack. 'ULCAs tools include a benchmarking suite that enables benchmarking datasets to be submitted and used to evaluate submitted models, alongside model publication and inference workflows [22]. Bhashini-linked ecosystems are also building leaderboards for evaluation (for example, machine translation, ASR, and TTS) and creating benchmarks for linguistic tools, reinforcing an evaluation-first approach for India's language AI deployment [23]. India's research community has also further produced a growing set of benchmarks that test foundation-model capability in Indian languages and contexts, including:

- Indic-Bias: It is a comprehensive benchmark to evaluate the fairness of LLMs across 85 Indian identity groups, focusing on bias and stereotypes. They created three tasks: plausibility, judgement, and generation, and evaluated 14 popular LLMs to identify allocative and representational harms [24].

- IndicXTREME : It is a human-supervised benchmark of 9 diverse NLU tasks across 20 languages, featuring 105 evaluation sets in total [25].

- MILU (Multi-task Indic Language Understanding Benchmark) : It covers 42 subjects and eight domains in 11 Indic languages, showing both general and culturally specific knowledge. With an India-centric design, it incorporates material from regional and state-level examinations, covering local history, arts, festivals, and laws alongside standard subjects like science and mathematics [26].

- Indic-Glue: It is a natural language understanding benchmark for Indian languages [27].

- IndicGen-Bench: It is the largest benchmark for evaluating LLMs on user-facing generation tasks across a diverse set of 29 Indic languages covering 13 scripts and 4 language families. It is composed of diverse generation tasks like cross-lingual summarisation, machine translation, and cross-lingual question answering. It also extends existing benchmarks to many Indic languages through human curation, providing multi-way parallel evaluation data for many under-represented Indic languages [28].

- EKA-Eval: It is a comprehensive evaluation framework for benchmarking large language models in Indian languages. It integrates over 35 benchmarks, including 10 Indic-specific benchmarks, across areas such as reasoning, mathematics, tool use, The framework focuses on long-context understanding and reading comprehension. The framework is designed to go beyond English-centric evaluation and support a more inclusive assessment of LLM performance in linguistically diverse countries, such as India.

- IIT Gandhinagar's CoMi-Lingua, PI-Indic-Align, and SangrahaTox: CoMi-Lingua is a large, expert-annotated dataset for Hindi–English code-mixed NLP. It contains over 1,00,900 instances in both Devanagari and Roman scripts and supports tasks such as language identification, matrix language identification, part-of-speech tagging, named entity recognition, and translation. PI-Indic-Align is a benchmark for evaluating how well embedding models align personas and instructions in low-resource

Indian languages. It is available in 12 Indian languages and covers four evaluation tasks, including monolingual and cross-lingual retrieval as well as compatibility classification. SangrahaTox is a multimodal benchmark for evaluating vision-language systems on alignment and safety. It covers bias, stereotypes, and safety across a few countries, including India, and is designed to assess culturally sensitive image–prompt pairs.

India is also developing an advanced benchmark for speech and Indian-accented evaluation. AI4Bharat documents benchmark efforts such as Vistaar and Svarah, including Svarah's focus on gaps in ASR (Automatic Speech Recognition) performance on Indian accents [29]. These benchmarks are important for foundation-model ecosystems where speech interfaces and multilingual voice assistants are central to India's vision of inclusion and service delivery.

- Soket AI's CoSHE-Eval is an evaluation dataset curated for testing Automatic Speech Recognition (ASR) systems on Hindi-English code-mixed speech. It focuses on bilingual conversational contexts commonly found in India, where Hindi (in Devanagari) and English (in Latin script) co-occur naturally within the same utterance.

The Indian industry is also developing specialised benchmarks for computer vision applications:

- BharatGen's Patram-7B-Instruct is presented as a vision-language model for visual document understanding, supporting document-intelligence evaluation for India-relevant workflows [30].

- BharatBench: It is an evaluation framework that benchmarks, or measures, the performance of multilingual LLMs (large language models) across Indic languages and cultural contexts [31].

These India-specific benchmarks are essential to India's AI ecosystem to ensure that foundation models are evaluated against the realities of Indian deployment, particularly in language and speech recognition tasks:

- **Language coverage and linguistic behaviour:** Indian deployments require performance across multiple scripts, dialectal variation, and frequent transliteration patterns that are under-tested in global benchmarks.

- **Contextual and cultural validity:** India-facing systems must be evaluated on cultural norms and references (including public-service contexts) to avoid systematic failure modes that global tests do not capture.
- **Fairness and representational harms:** Indian society has distinct axes of social identity and risk; fairness benchmarks designed for Indian contexts enable more meaningful evaluation of bias and stereotyping in local deployments.
- **Governance and procurement readiness:** Standardised, repeatable benchmarks create a practical foundation for model comparison, procurement qualification, and sectoral adoption, supporting more consistent evaluation expectations across government and regulated domains.ent and regulated domains.

Conclusion

The global foundation-model landscape is no longer defined only by model performance or market adoption. It is increasingly shaped by who can secure and scale the enabling infrastructure required to build models, high-end compute and data centre capacity, access to specialised chips, and large, high-quality datasets. Across countries, the strategic direction is to strengthen domestic capacity to train and deploy models at scale, while also shaping the supply chains and platforms that determine access. Some jurisdictions are viewing advanced computing and chip supply chains as strategic assets by implementing stricter controls, while others are pairing rules with higher expectations for companies to be responsible and ready to follow regulations in the early stages of model development. Some countries are also scaling national compute capacity and directing AI deployment toward industrial integration. These patterns reinforce a common insight: foundation models are increasingly being treated as a sovereign capability, anchored in infrastructure and ecosystem depth.

India's approach is centred on building indigenous capability across the foundation-model stack. Rather than relying on a single model, India is developing an ecosystem that combines (i) shared compute access, (ii) India-centric data and model repositories, and (iii) multiple model-building efforts across text, speech, multimodal, and sectoral systems. This infrastructure layer is already being translated into model development, as we have seen in the above sections. Complementing this effort, private-sector innovation is also expanding the ecosystem through small language models and sector-specific systems trained on Indian datasets, complementing the larger foundation models with practical solutions that can be deployed quickly in real workflows.

Therefore, India's ambition is to establish a sustainable and competitive foundation model capacity that can support long-term national requirements that will reduce the structural dependence on external providers; enable inclusive and affordable adoption across Indian languages, regions, and sectors through a layered ecosystem of large models, multimodal systems, and small language models; and position India as a credible contributor to global innovation.

References

1. Bommasani, R., et al. (2021). *On the opportunities and risks of foundation models* (CRFM Report). Center for Research on Foundation Models, Stanford University.
2. Ada Lovelace Institute. (2024, July). What is a foundation model?
3. Ministry of Electronics and Information Technology. (2025, November). *India AI governance guidelines:* Enabling safe and trusted AI innovation. Press Information Bureau.
4. Lutkevich, B. (2025, January 6). Foundation models explained: Everything you need to know. TechTarget.
5. Press Information Bureau. (2024, March 6). Cabinet approves ambitious IndiaAI Mission to strengthen the artificial intelligence ecosystem in India [Press release].
6. India Semiconductor Mission. (February 9, 2026). India Semiconductor Mission.
7. IndiaAI. (2025, January 31). *IndiaAI Mission:* Call for proposals to build foundational AI models.
8. IndiaAI. (2025, May 15). Building India's foundational AI models*: IndiaAI innovation initiative*.
9. Press Information Bureau. (2025). *IndiaAI Mission: [*Content related to foundational AI models announcement*]* [Press release].
10. Organisation for Economic Co-operation and Development. (Accessed on 2026, February 9). AI Principles Dashboard: P9. OECD.AI.
11. artificialintelligenceact.eu. (Accessed on 2026, February 9.). Article 53.
12. Ministry of Electronics and Information Technology. (2023, August 11). *The* Digital Personal Data Protection Act, 2023.
13. European Parliamentary Research Service. (2020, June). The impact of the General Data Protection Regulation (GDPR) on artificial intelligence.

Personal Information Protection Law

14. of the People's Republic of China. (Accessed on 2026, February 9.). *Article 13*.
15. Personal Information Protection Law of the People's Republic of China. (Accessed on 2026, February 9.) Article 40.
16. Ministry of Electronics and Information Technology. (2025, October 22). *Explanatory note: Proposed amendments to the Information Technology (I*ntermediary Guidelines and Digital Media Ethics Code) Rules, 2021 in relation to synthetically generated information.
17. China Law Translate. (Accessed on 2026, February 9.). Measures for labeling of AI-generated synthetic content.
18. artificialintelligenceact.eu. (Accessed on 2026, February 9.). Article 50.
19. Attorney-General's Chambers. (2021). Copyright Act 2021 (ProvIds=pr244-).
20. Cyberspace Administration of China. (2023, July 13). Interim measures for the management of generative artificial intelligence services.
21. April, J. (2024, June 4). *A detailed analysis of A*rticle 50 of the EU's Artificial Intelligence Act. SSRN.
22. Bhashini. (Accessed on 2026, February 9.). *ulca* [Software]. GitHub.
23. Bhashini. (Accessed on 2026, February 9.). Anusandhan Mitra*: Project 13*.

24. AI4Bharat. (Accessed on 2026, February 9.). *Indic-Bias* [Data set]. Hugging Face.
25. AI4Bharat. (Accessed on 2026, February 9.). IndicXTREME [Data set]. Hugging Face.
26. IndiaAI. (2024, November 7). AI4Bharat and IBM Research India released an evaluation benchmark
27. AI4Bharat. (Accessed on 2026, February 9.). *I*ndicGLUE benchmark.
28. Singh, H., Gupta, N., Bharadwaj, S., Tewari, D., & Talukdar, P. (2024). *IndicGenBench:* A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. arXiv.
29. AI4Bharat. (Accessed on 2026, February 9.). Automatic Speech Recognition.
30. BharatGen. (Accessed on 2026, February 9). *patram-7b-instruct* [Software]. Hugging Face.
31. Olakrutrim AI Labs. (2024, February 4). *BharatBench:* Comprehensive multilingual multimodal benchmark suite for Indian languages and culture.
32. Ministry of Electronics & Information Technology. (Accessed on 2026, March 6.). Guidelines for procurement of cloud services (Version 2.2).
33. Government of India. Indian Computer Emergency Response Team. (2022, April 28). Directions under sub-section (6) of section 70B of the Information Technology Act, 2000. Ministry of Electronics & Information Technology, Government of India.

Annexure 1

Organisation	Model Type	Details
Sarvam AI	Sovereign LLM ecosystem (open-source)	Sarvam AI is developing a sovereign 105B-parameter LLM, referenced alongside 30B models designed for Indian languages, with a focus on governance, public service, and high-stakes deployment.
Soket AI	Multilingual text foundation model (open-source)	Soket AI Labs, under the IndiaAI Mission, is developing a 120-billion-parameter open-source multilingual foundation model tailored for India's linguistic diversity.
Gnani AI	V oice AI foundation model	14B-parameter multilingual, real-time speech processing with advanced reasoning capabilities.
Gan AI	Multilingual TTS foundation model	70B-parameter model targeting high-performance ("superhuman") text-to-speech.
A vataar/A vatar AI	AI A vatars	Creating specialised "AI Avatars" up to 70B parameters, optimised for Indian languages and domains such as agriculture, healthcare, and governance.
IIT Bombay Consortium – Bharat Gen	General-purpose multilingual, multimodal	Developing multilingual and multimodal models ranging from 2B to 1T parameters, with an open-source approach to support applications in agriculture, finance, legal, health, and education.
Fractal Analytics Ltd.	Large reasoning model	Building India's first large reasoning model of up to 70B parameters, designed for structured reasoning, STEM disciplines, and medical problem-solving.
Tech Mahindra Maker's Lab	Indic language model	Designing an efficient 8B parameter model for Indic languages (with a focus on Hindi dialects), alongside an agentic AI platform, Orion, for government applications.
Zenteiq	Science-driven multimodal foundation model	Developing BrahmAI, a science-driven multimodal foundation model (8B–80B parameters) to advance engineering intelligence, scientific computing, and industrial innovation.
GenLoop	Small language model suite	Creating small language models (2B parameters) – Yukti (Base), Varta (Instruction), and Kavach (Guard) – to support all 22 scheduled Indian

		languages with native reasoning and content moderation.
Intellihealth	Health-signal foundation model	Proposing a 20B parameter model for EEG signal analysis to enable early screening of neurological disorders and advance brain–computer interface research.
Shodh AI	Scientific discovery model	Developing a 7B parameter model for material discovery, integrating AI into experimental workflows to accelerate innovation in material sciences.

List of Abbreviations

AI — Artificial Intelligence

AIKosh — IndiaAI Datasets Platform (datasets/models/use -cases repository)

AI4Bharat — AI4Bharat initiative (Indic language AI ecosystem; IIT Madras)

API / APIs — Application Programming Interface(s)

ASR — Automatic Speech Recognition

BIS — Bureau of Indian Standards

DPDP — Digital Personal Data Protection (Act)

DST — Department of Science & Technology

EU — European Union

FM — Foundation Model(s)

GPU / GPUs — Graphics Processing Unit(s)

IP — Intellectual Property

ISM — India Semiconductor Mission

ISO/IEC — International Organization for Standardization / International Electrotechnical Commission

IT (Rules / Act) — Information Technology

LLM / LLMs — Large Language Model(s)

MeitY — Ministry of Electronics and Information Technology

MILU — Multi -task Indic Language Understanding Benchmark

MSME / MSMEs — Micro, Small and Medium Enterprise(s)

SLM / SLMs — Small Language Model(s)

TDM — Text and Data Mining

TTS — Text -to-Speech

ULCA — Universal Language Contribution APIs

सत्यमेव जयते

Office of the Principal Scientific Adviser
to the Government of India