GOVERNMENT OF INDIA
MINISTRY OF ELECTRONICS AND INFORMATION TECHNOLOGY
**RAJYA SABHA**
**UNSTARRED QUESTION NO. 3254**
TO BE ANSWERED ON: 20.03.2026

# INCENTIVISING DATA LOCALISATION AND AI READY DATASET

**3254. SHRI KARTIKEYA SHARMA:**

Will the Minister of ELECTRONICS AND INFORMATION TECHNOLOGY be pleased to state:

(a) the details of the proposed data governance framework to incentivise data localisation through mirrored datasets for large-scale Artificial Intelligence (AI) training;
(b) the number of indigenous Startups supported for curating high-quality, anonymised and machine-readable domestic datasets in the healthcare and agriculture sectors;
(c) the status of the 'AI-OS' initiative to convert Artificial Intelligence into a public good through shared open-source code repositories; and
(d) the details of financial contributions required from firms deriving substantial commercial value from Indian data to support domestic Research and Development (R&D) in AI?

## ANSWER

MINISTER OF STATE FOR ELECTRONICS AND INFORMATION TECHNOLOGY
(SHRI JITIN PRASADA)

(a) to (d): Government of India launched the IndiaAI Mission in March 2024 with the objective of building a robust ecosystem of AI in the country. The Mission seeks to expand access to AI technologies, support innovation and promote the development of AI solutions addressing India-centric challenges.

The Government supports domestic AI research and innovation under the IndiaAI Mission through development of common compute, AIKosh, applications, skilling and startup ecosystem development.

**AIKosha:**
AIKosha is platform which provide access AI models, development tools and other resources. These belong to various fields such as health, agriculture, and education, with safeguards for data privacy.

AIKosh also provides a secure API-based access, an AI Sandbox environment for model training and experimentation.

It also provides AI Guardrails toolkits, which support developers in monitoring, evaluating and constraining model behaviour during development and deployment.

At present, the non-personal data governance is governed by the National Data Sharing and Accessibility Policy (NDSAP) which aims to increase the accessibility and ease of sharing of non-sensitive data amongst registered users and their availability for scientific, economic, and social developmental purposes.

The details and features of AIKosh can be accessed from https://aikosh.indiaai.gov.in/.

**Digital India Bhashini**

Bhashini is part of National Language Translation Mission (NLTM) which focuses on creating AI-driven language solutions.

- Citizens contribute voice, text, and translations in 22 Indian languages on BhashaDaan platform
- In collaboration with over 70+ research institutions & sectoral experts, large volumes of annotated datasets are curated for different technologies
- These include speech recognition, machine translation, & other language technologies

₹47 crore was allocated for building datasets for 22 scheduled Indian languages across Translation, Speech Recognition, Speech Synthesis and OCR activities.

Data is collected from people across different regions, communities, and backgrounds to reflect India's true language diversity and avoid bias. It includes real-life dialects and spoken variations, capturing the richness of India's linguistic landscape.

The idea of an "AI-OS" platform under the broader national AI ecosystem strategy has been proposed by Economic Survey 2025-26. The aim is to promote open-source AI development and shared infrastructure in India.

**Development of Indigenous Foundational Model**

Under the **IndiaAI Foundation Models** pillar, the Government of India is supporting Indian startups and institutions.

This support is for developing large language models, multimodal models, and domain-specific small language models. These models are intended to strengthen India's domestic AI capabilities

The developed models are expected to contribute to the open-source ecosystem. They will be made available through the **AIKosh platform**.

This will enable access for other startups, researchers, and academic institutions.

Twelve organisations and consortia, including startups, industry players and academic institutions have been selected for developing Large and Small Language Models based on Indian datasets.

The selected projects cover multilingual foundational models, speech and voice models, multimodal AI, scientific models, healthcare reasoning systems, and agentic AI platforms.

The models are developed based on Indian datasets spanning all 22 scheduled Indian languages.

Sarvam AI, BharatGen, Gnani, and Socket launched their models at IndiaAI Impact Summit in February, 2026.

The Government is promoting open and collaborative AI innovation through shared platforms, open datasets, model repositories and tools under the IndiaAI ecosystem.

******