GOVERNMENT OF INDIA
MINISTRY OF EDUCATION
DEPARTMENT OF HIGHER EDUCATION
LOK SABHA
STARRED QUESTION NO. 1604
ANSWERED ON 09.02.2026

AI LANGUAGE ASSAMESE AND OTHER LANGUAGES

†1604. SHRI PARIMAL SUKLABAIDYA:

Will the Minister of EDUCATION be pleased to state:

(a) the details of initiatives undertaken by the Government to promote all the 22 Scheduled Indian languages through AI-based language platforms such as Bhashini and BharatGen and the progress achieved so far, in this regard;

(b) the extent of digitisation of linguistic data, development of multilingual AI tools and creation of open language datasets under these initiatives, with special reference to languages spoken in Assam, including Assamese, Bengali, Bishnupriya Manipuri and other languages of the Barak Valley and Cachar district;

(c) the role played by academic institutions, startups and public sector agencies, particularly from the North Eastern Region, in the development and deployment of these language technologies; and

(d) whether the Government proposes to expand the use of these platforms for education, governance and public service delivery in linguistically diverse districts such as Cachar district in Assam and if so, the details thereof?

ANSWER

MINISTER OF STATE IN THE MINISTRY OF EDUCATION

(DR. SUKANTA MAJUMDAR)

(a) to (d): The National Education Policy (NEP) 2020 highlights the importance of multilingualism and places strong emphasis on the promotion of all Indian languages. In alignment with the objectives of NEP 2020, the Government of India has undertaken several initiatives promoting education, preservation and research on Indian languages. These efforts have further been augmented through the adoption of Artificial Intelligence and Machine Learning (AI/ML) technologies, including the development of Large Language Models, translation tools and language-enabled digital services in Indian languages.

Government of India has started the BharatGen project, which is a multimodal Large Language Model Project focused on developing efficient and inclusive AI solutions that support all 22 Scheduled languages and enable the creation of a robust digital AI infrastructure for India's unique socio-cultural context and diverse sectors. BharatGen is anchored at IIT Bombay, where the core development of language technologies is undertaken. The initiative is implemented through a consortium of leading academic institutions including IIT Kanpur, IIT Madras, IIT Hyderabad, IIIT Hyderabad, IIM Indore, and IIT Mandi.

Further, to facilitate the translation of content into Indian languages, technological advancements include the development of AI-based translation tools such as ANUVADINI by the All India Council of Technical Education (AICTE) and BHASHINI, an initiative under the Digital India Programme by M/o Electronics and Information Technology.

BHASHINI has developed state-of-the-art AI models for Indian languages through a collaboration of over 70 research partner institutes. BHASHINI platform hosts a repository of over three hundred fifty AI-based language models and provides more than 22 specialized language services. These services include Automatic Speech Recognition (ASR), Machine Translation (MT), Text-to-Speech (TTS), Optical Character Recognition (OCR) and Transliteration. The dataset corpus includes 246 million parallel sentence pairs and 3.7 million monolingual text entries. All datasets and models are publicly accessible via BHASHINI platform or through the Digital India BHASHINI Division account on the AIKosh platform. Further, the National Language Translation Mission through the BHASHINI platform is digitising large volumes of text and speech data across all 22 scheduled languages.

BHASHINI follows a multi-stakeholder participation model involving academic and research institutions including from the North Eastern Region for fundamental research, dataset preparation and development of baseline AI models. BHASHINI has taken steps to support Assamese, Bengali and other Indian languages in collaboration with Government of Assam for use in digital governance platforms and citizen services, including those in linguistically diverse districts such as Cachar. These include integration of translation plugins, facilitating access to APIs etc.

Further, extensive linguistic resources for scheduled Indian Languages have been developed by the Central Institute of Indian Languages (CIIL) of the Ministry of Education under the Linguistic Data Consortium for Indian Languages (LDC-IL) scheme. Since 2019, LDC-IL has provided resources to government agencies, government-promoted initiatives, researchers, and commercial and industrial users engaged in developing language technologies. The resources released for the Assamese language include a Raw Text Corpus, Raw Speech Corpus, Sentence-Aligned Speech Corpus, TTS Corpus and Mother Tongue Parallel Text Corpus. Apart from the datasets, LDC-IL has also developed web-based applications and tools to support research and development in Indian languages. Several of these tools, hosted at CIIL data centres, are accessible through the Medha Bhashika website https://medha.ciil.org.

******