भारत सरकार

इलेक्ट्रॉनिकी और सूचना प्रौद्योगिकी मंत्रालय

लोक सभा

अतारांकित प्रश्न संख्या 486

जिसका उत्तर 03 दिसंबर, 2025 को दिया जाना है ।

12 अग्रहायण, 1947 (शक)

तमिल भाषा डेटासेट

486. थिरु दयानिधि मारन:

क्या इलेक्ट्रॉनिकी और सूचना प्रौद्योगिकी मंत्री यह बताने की कृपा करेंगे कि:

(क) क्या सरकार ने राष्ट्रीय भाषा प्रसंस्करण (एनएलपी) के लिए भारत एआई मिशन के अंतर्गत वाक् पहचान, मशीन अनुवाद और सामग्री मॉडरेशन के लिए भारतीय भाषा एआई बनाए या वित्तपोषित किए हैं और यदि हां, तो ऐसे डेटासेट की स्थिति और उपलब्धता क्या है तथा इसके लिए भाषा-वार और मद-वार कितनी धनराशि आवंटित की गई है;

(ख) गत पांच वर्षों के दौरान इन डेटासेटों को विकसित करने के लिए भाषा-वार कितनी धनराशि आवंटित की गई और वास्तव में कितनी धनराशि उपयोग की गई तथा कार्यान्वयन एजेंसियों का ब्यौरा क्या है;

(ग) क्या सरकार का तमिल डिजिटल पारिस्थितिकी तंत्रों में एआई सटीकता, सुरक्षा और संतुलन में सुधार लाने के लिए बोली जाने वाली भाषाओं सहित तमिल में समर्पित, उच्च गुणवत्ता वाले डेटासेट बनाने का विचार है और यदि हां, तो इसके लिए उठाए गए विशिष्ट कदमों, समय-सीमा और प्रदान की गई बजटीय सहायता का ब्यौरा क्या है;

(घ) क्या ऐसे डेटासेट के विकास के लिए मार्गदर्शन हेतु तमिल भाषा विशेषज्ञों, विश्वविद्यालयों अथवा अनुसंधान संस्थानों के साथ कोई परामर्श किया गया है और यदि हां, तो तत्संबंधी ब्यौरा क्या है; और

(ङ) तमिल भाषा डेटासेट पर कुल कितनी धनराशि का बजट बनाया गया और कितनी धनराशि आवंटित, स्वीकृत और खर्च की गई है?

उत्तर

इलेक्ट्रॉनिकी और सूचना प्रौद्योगिकी राज्य मंत्री (श्री जितिन प्रसाद)

(क) से **(ङ)** भारत की एआई कार्यनीति प्रौद्योगिकी के उपयोग का लोकतंत्रीकरण करने के माननीय प्रधानमंत्री के दृष्टिकोण पर आधारित है। इसका उद्देश्य भारत केंद्रित चुनौतियों का समाधान करना और सभी भारतीयों के लिए अवसर पैदा करना है।

<u>इंडियाएआई मिशन:</u>

यह सात स्तंभों- इंडियाएआई कंप्यूट, एआईकोष, इंडियाएआई फाउंडेशन मॉडल्स, इंडियाएआई फ्यूचर स्किल्स, स्टार्टअप फाइनेंसिंग, एप्लिकेशन डेवलपमेंट और सुरक्षित एवं विश्वसनीय एआई के माध्यम से भारत के विकास लक्ष्यों के साथ एक सुदृढ़ और समावेशी एआई इकोसिस्टम स्थापित करने की एक रणनीतिक पहल है।

एआईकोश पर लगभग 275 डेटासेट अपलोड किए गए हैं। इसमें प्रमुख संगठनों द्वारा योगदान किए गए 100+ तमिल भाषा डेटासेट शामिल हैं।

इंडियाएआई मिशन एनएलपी, स्पीच, अनुवाद और सामग्री मॉडरेशन सहित विभिन्न उपयोग के मामलों के लिए भारतीय भाषा के एआई डेटासेट के निर्माण और वित्त पोषण का समर्थन करता है।

<u>मिशन भाषिणी:</u>

डिजिटल इंडिया कार्यक्रम के अंतर्गत मिशन भाषिणी भारतीय भाषाओं के लिए एआई-संचालित स्पीच और टेक्स्ट प्रौद्योगिकियों को विकसित करके बहुभाषी डिजिटल पहुंच को सक्षम बनाता है। यह राष्ट्रीय भाषा अनुवाद मिशन (एनएलटीएम) के अंतर्गत आधिकारिक रिपॉजिटरी के रूप में कार्य करता है।

भाषिणी को 70 से अधिक अनुसंधान भागीदार संस्थानों के राष्ट्रीय सहयोग के माध्यम से विकसित किया गया है। यह **350 से अधिक एआई-आधारित भाषा मॉडल** का कोष होस्ट करता है और **22+ विशेष भाषा सेवाएं प्रदान करता है।**यह ऑटोमेटिक स्पीच रिकग्निशन (एएसआर), मशीन अनुवाद (एमटी), टेक्स्ट-टू-स्पीच (टीटीएस), ऑप्टिकल कैरेक्टर रिकॉग्निशन (ओसीआर) और लिप्यंतरण प्रदान करता है।

· डेटासेट कॉर्पस में शामिल हैं:

Ø 246 मिलियन समानांतर वाक्य युग्म
Ø 3.7 मिलियन मोनोलिंगुअल टेक्स्ट प्रविष्टियां
Ø 14,000 घंटे का एएसआर ऑडियो
Ø 2.5 मिलियन ओसीआर इमेज सैंपल
Ø 476 घंटे का टीटीएस ऑडियो
Ø भारतीय भाषाओं में 20.56 मिलियन लिप्यंतरण प्रविष्टियां

सभी डेटासेट और मॉडल भाषिणी प्लेटफॉर्म के माध्यम से या एआईकोश प्लेटफॉर्म पर डिजिटल इंडिया भाषिणी डिवीजन अकाउंट के माध्यम से सार्वजनिक रूप से उपलब्ध हैं।

तमिल और बोली जाने वाली भाषाओं सहित भारतीय भाषाओं के लिए समर्पित डेटासेट बनाए गए हैं। एएसआर के लिए, तमिल सहित 22 अनुसूचित भारतीय भाषाओं के लिए ट्रांसक्राइब्ड स्पीच डेटासेट और प्रशिक्षित मॉडल उपलब्ध हैं, जिनका आईआईटी मद्रास में ए14भारत द्वारा योगदान किया गया है।

तमिल डेटासेट में समानांतर और मोनोलिंगुअल कॉर्पोरा, एएसआर (लेबल/अनलेबल), टीटीएस, ओसीआर, लिप्यंतरण, शब्दावली, **एनईआर (नामित इकाई मान्यता)** और शब्दावली संसाधन शामिल हैं, जो भाषिणी और एआईकोश प्लेटफार्मों के माध्यम से सार्वजनिक रूप से सुलभ हैं।

भाषाई विशेषज्ञों, अनुसंधान और अकादमिक भागीदारों के साथ संरचित परामर्श के माध्यम से विकास कार्य किया गया है।

अनुवाद, स्पीच रिकग्निशन, स्पीच सिंथेसिस और ओसीआर गतिविधियों में 22 अनुसूचित भारतीय भाषाओं के लिए डेटासेट बनाने के लिए कुल 47 करोड़ रुपये आवंटित किए गए थे और पूरी राशि का पूरी तरह से उपयोग किया गया है।

*****