2,000 वर्ष पुरानी, फिर भी ‘लो-रिसोर्स’: तमिल की AI जंग

समाचार-संदर्भ: 7 नवंबर 2026 | विषय: विज्ञान और प्रौद्योगिकी
प्रकाशन: 10 नवंबर 2026
मुख्य परीक्षा: GS-III: विज्ञान-प्रौद्योगिकी, AI और भाषा-तकनीक



मौलिक प्रतीकात्मक चित्र: pcsmantra.com

खबरों में क्यों

तमिलनाडु सरकार द्वारा आयोजित "Tamil for AI" सम्मेलन में विशेषज्ञों ने कहा कि 2,000 वर्ष पुरानी, शास्त्रीय-दर्जा प्राप्त तमिल भाषा AI की दुनिया में "लो-रिसोर्स" भाषा बनी हुई है। सम्मेलन में भाषा-डेटा, शास्त्रीय ज्ञान और नीति-हस्तक्षेप पर चर्चा हुई।

कम-संसाधन दर्जे का तमिल-विरोधाभास

तमिल को 2004 में भारत सरकार ने शास्त्रीय भाषा का दर्जा दिया। इसकी 2,000 वर्ष पुरानी लिखित परंपरा, 8.5 करोड़ से अधिक बोलने वाले और संगम साहित्य से लेकर आधुनिक साहित्य तक समृद्ध ग्रंथ-परंपरा है। फिर भी AI भाषा-मॉडलों के प्रशिक्षण-डेटा में तमिल की उपस्थिति अंग्रेजी, चीनी या स्पेनिश की तुलना में नगण्य है।

विरोधाभास यह है कि भाषा ऐतिहासिक रूप से समृद्ध है, लेकिन डिजिटल रूप से विपन्न। शास्त्रीय ग्रंथों का डिजिटलीकरण, मशीन-पठनीय स्वरूप और एनोटेटेड डेटा की कमी AI प्रशिक्षण में बाधा है।

सम्मेलन की मुख्य बातें

डेटा-संकट: तमिल का उच्च-गुणवत्ता डिजिटल कोर्पस सीमित है। अधिकांश सामग्री अनौपचारिक, अमानकीकृत या कॉपीराइट-बंधित है।

शास्त्रीय ज्ञान: तोल्काप्पियम, तिरुक्कुरल और संगम ग्रंथों में व्याकरण, चिकित्सा, नैतिकता और राजनीति का ज्ञान है। इसे AI-प्रशिक्षण में जोड़ने से भाषा-मॉडल की सांस्कृतिक गहराई बढ़ेगी।

तकनीकी समाधान: विशेषज्ञों ने ओपन-सोर्स तमिल NLP टूल्स, ऑप्टिकल कैरेक्टर रिकग्निशन (OCR) से पुराने ग्रंथों का डिजिटलीकरण और सामुदायिक डेटा-योगदान के सुझाव दिए।

नीति-प्रतिबद्धता: तमिलनाडु सरकार ने तमिल भाषा-तकनीक के लिए ₹50 करोड़ के कोष की घोषणा की, जिसमें डेटा-बुनियादी ढांचा, शोध-छात्रवृत्ति और स्टार्टअप समर्थन शामिल है।

AI4Bharat का योगदान: IIT Madras के AI4Bharat समूह ने Nemotron-Pretraining-Indic-v1 जारी किया, जिसमें 22 भारतीय भाषाओं के लिए प्रशिक्षण-डेटा है। इसमें तमिल का हिस्सा बढ़ाया गया है।

चुनौती: भाषा-तकनीक में कुशल श्रमिकों की कमी, तमिल-उच्चारण की विविधता और तमिल-इंग्लिश मिश्रित बोलचाल (code-mixing) मॉडल-प्रशिक्षण को जटिल बनाती है।

25 वर्ष पहले की तुलना

2001 में तमिल यूनिकोड में आया। उस समय चुनौती टाइपिंग और फॉन्ट थी। आज चुनौती AI-प्रशिक्षण डेटा है। तकनीक बदली, लेकिन भाषा-न्याय का प्रश्न वही है: क्या डिजिटल दुनिया में सभी भाषाओं को समान अवसर मिलेगा?

शास्त्रीय-दर्जा संदर्भ

भारत में 6 भाषाओं को शास्त्रीय दर्जा मिला है: तमिल (2004), संस्कृत (2005), कन्नड़ और तेलुगु (2008), मलयालम (2013), ओडिया (2014)। शास्त्रीय दर्जे से शोध-कोष, केंद्रीय विश्वविद्यालयों में अध्ययन-केंद्र और पुरस्कार मिलते हैं, लेकिन AI-युग में इसका अर्थ बदल रहा है।

आगे की राह

तमिल का AI-भविष्य तीन कारकों पर निर्भर है: (1) डिजिटल डेटा की मात्रा और गुणवत्ता, (2) शास्त्रीय ज्ञान का AI में समावेश, (3) भाषा-तकनीक में निवेश। सम्मेलन ने दिशा दिखाई; अमल अब नीति और समुदाय पर है।

इससे जुड़ा लेख: Nemotron-Pretraining-Indic-v1: भारतीय भाषाओं के लिए AI4Bharat की नई पहल

प्रिलिम्स परीक्षा अभ्यास प्रश्न

तमिल भाषा के संदर्भ में विचार कीजिए:
1. तमिल को 2004 में शास्त्रीय भाषा का दर्जा मिला।
2. AI भाषा-मॉडलों में तमिल "हाई-रिसोर्स" भाषा है।
3. तिरुक्कुरल तमिल का शास्त्रीय ग्रंथ है।

(a) केवल 1

(b) केवल 2

(c) केवल 1 और 3

(d) सभी

उत्तर और व्याख्या

उत्तर: (c) केवल 1 और 3

AI भाषा-मॉडलों में तमिल "लो-रिसोर्स" (कम-संसाधन) भाषा है, हाई-रिसोर्स नहीं।

भारत में शास्त्रीय भाषा दर्जे के संदर्भ में सही कथन चुनिए:
1. संस्कृत को 2005 में दर्जा मिला।
2. केवल 4 भाषाओं को यह दर्जा मिला है।
3. मलयालम को 2013 में दर्जा मिला।

(a) केवल 1

(b) केवल 1 और 3

(c) केवल 2 और 3

(d) सभी

उत्तर और व्याख्या

उत्तर: (b) केवल 1 और 3

भारत में 6 भाषाओं को शास्त्रीय दर्जा मिला है: तमिल, संस्कृत, कन्नड़, तेलुगु, मलयालम, ओडिया।

मुख्य परीक्षा अभ्यास प्रश्न

शास्त्रीय भाषा होते हुए भी तमिल AI-युग में "लो-रिसोर्स" क्यों है? भाषा-न्याय और तकनीकी समाधान पर चर्चा कीजिए। (250 शब्द)

उत्तर की रूपरेखा

भूमिका में: तमिल की ऐतिहासिक समृद्धि और डिजिटल विपन्नता का विरोधाभास बताएं।

मुख्य भाग में: डेटा-संकट, शास्त्रीय ज्ञान का AI-समावेश, OCR/डिजिटलीकरण और नीति-हस्तक्षेप (तमिलनाडु का ₹50 करोड़ कोष) समझाएं; AI4Bharat जैसी पहलों का उल्लेख करें; भाषा-न्याय का बड़ा प्रश्न उठाएं।

निष्कर्ष में: तकनीक भाषा-विविधता को मजबूत करे, कमजोर नहीं - यह संदेश दें।

स्रोत

1. The Hindu (7 नवंबर 2026): Tamil for AI conference highlights low-resource challenge
2. AI4Bharat, IIT Madras: Nemotron-Pretraining-Indic-v1