कृत्रिम बुद्धिमत्ता के माध्यम से भारतीय भाषाओं और लोकज्ञान का संरक्षण: चुनौतियाँ और समाधान

रीता देवी

सहायक प्राध्यपिका, हिन्दी विभाग, रविंद्रनाथ टैगोर राजकीय महाविद्यालय सरकाघाट

प्रस्तावना:

भारतीय भाषाई और सांस्कृतिक विरासत का महत्व

भारत, अपनी विशाल भौगोलिक और सांस्कृतिक परिधि के साथ, एक अद्वितीय भाषाई विविधता का खजाना है। 2024 के आंकड़ों के अनुसार, देश में 453 जीवित भाषाएँ हैं । भारत की भाषाई सर्वेक्षणों ने इस विविधता को समय के साथ लगातार दर्ज किया है। 1961 की जनगणना में 1,652 मातृभाषाएँ दर्ज की गईं, जो 2011 तक उल्लेखनीय रूप से बढ़कर 2,843 हो गईं, जिसमें से 1,369 को वर्गीकृत किया गया और 1,474 को अवर्गीकृत माना गया । यह भाषाई विस्तार न केवल देश की समृद्ध विरासत को दर्शाता है, बल्कि यह भी इंगित करता है कि दस्तावेज़ीकरण के प्रयास बढ़ रहे हैं, जिससे पहले अज्ञात भाषाओं को पहचान मिल रही है।

हालांकि, इस विविधता का एक बड़ा हिस्सा गंभीर खतरे में है। यूनेस्को की सूची के अनुसार, हाई-टेक युग में भी भारत में 42 से अधिक भाषाएँ या बोलियाँ संकट में हैं, जिनके बोलने वालों की संख्या कुछ हजार तक सीमित है । लुप्तप्राय भाषाओं को आम तौर पर उन भाषाओं के रूप में परिभाषित किया जाता है जिन्हें 10,000 से कम लोग बोलते हैं । इस संकट में केवल भाषाएँ ही नहीं, बल्कि उनसे जुड़ी मौखिक परंपराएँ, लोक साहित्य और पारंपरिक ज्ञान भी शामिल हैं, जो पीढ़ी-दर-पीढ़ी मौखिक रूप से हस्तांतरित होते रहे हैं । भाषाओं के विलुप्त होने से मानव इतिहास, संस्कृति और सामूहिक ज्ञान के लिए एक अपूरणीय क्षति होती है ।

कृत्रिम बुद्धिमत्ता (AI) और विशेष रूप से प्राकृतिक भाषा प्रसंस्करण (NLP) जैसी प्रौद्योगिकियाँ इस संकट को कम करने के लिए एक अभूतपूर्व अवसर प्रदान करती हैं। यह रिपोर्ट इन प्रौद्योगिकियों की क्षमता, उनके कार्यान्वयन में आने वाली चुनौतियों और एक स्थायी समाधान के लिए रणनीतिक मार्ग का गहन विश्लेषण प्रस्तुत करती है, जिसका उद्देश्य भाषाई और सांस्कृतिक संरक्षण को तकनीकी नवाचार के साथ एकीकृत करना है।

भाग 1: भारतीय भाषाओं और लोकज्ञान के संरक्षण का संकट

1.1 भारत में भाषाई विविधता की स्थिति

भारत एक बहुभाषी देश है जहाँ लगभग 780 भाषाएँ और 86 लिपियाँ बोली जाती हैं । यह बहुभाषिकता ऐतिहासिक, भौगोलिक और सांस्कृतिक कारकों का परिणाम है । हालांकि, यह भाषाई विविधता खतरे में है। कई जनजातीय, द्वीप-आधारित और क्षेत्रीय भाषाएँ, जैसे कि ग्रेट अंडमानीज, जरावा, ओंगे (अंदमानी भाषाएँ), टोडा, कोरागा, कुरुबा (द्रविड़ भाषाएँ), बिरहोर, गुतोब (मुंडा भाषाएँ), और ताई नोरा (क्रा-दाई भाषाएँ) विलुप्त होने के कगार पर हैं ।

भाषा संरक्षण की वर्तमान स्थिति में एक अंतर्निहित विरोधाभास मौजूद है। 1961 से 2011 के बीच दर्ज की गई मातृभाषाओं की संख्या में वृद्धि हुई है, जो दस्तावेज़ीकरण के प्रयासों में सुधार को दर्शाती है और उन भाषाओं की पहचान करती है जो पहले अज्ञात थीं । यह एक सकारात्मक विकास है। हालांकि, यह वृद्धि भाषाओं के विलुप्त होने की तेज दर को कम नहीं करती, खासकर उन भाषाओं के लिए जिन्हें 10,000 से कम लोग बोलते हैं। एक प्रमुख कारण यह है कि भारत सरकार ऐसी किसी भाषा को मान्यता नहीं देती जिसे 10,000 से कम लोग बोलते हैं । यह नीतिगत विसंगति इन भाषाओं को वित्तीय और शैक्षणिक सहायता से वंचित कर सकती है, जिससे उनका विलुप्त होना और भी तीव्र हो जाता है।

1.2 लोकज्ञान और मौखिक परंपराओं का ह्रास

पारंपरिक ज्ञान (Traditional Knowledge), जिसमें मौखिक कथाएँ, लोक साहित्य, पारंपरिक प्रथाएँ और ज्ञान शामिल है, पीढ़ी-दर-पीढ़ी मौखिक रूप से हस्तांतरित होता रहा है । यह ज्ञान अक्सर लिखित रूप में दस्तावेज़ित नहीं होता, और भाषाओं के विलुप्त होने के साथ यह भी पूरी तरह से लुप्त हो रहा है, जिससे मानव इतिहास और संस्कृति को एक अमूल्य क्षति पहुँच रही है । भाषाओं का विलुप्त होना केवल एक सांस्कृतिक हानि नहीं है; यह एक आर्थिक हानि भी है क्योंकि आज की अधिकांश तकनीकें भाषा पर आधारित हैं । जिन भाषाओं को लिखित रूप में दर्ज नहीं किया गया है, उनका ज्ञान और उनसे जुड़े बौद्धिक अधिकार भी खो जाते हैं, जिससे समुदाय ज्ञान के दुरुपयोग से अपनी विरासत की रक्षा करने में असमर्थ हो जाते हैं।

पारंपरिक ज्ञान डिजिटल लाइब्रेरी (TKDL) का सफल उदाहरण इस बात को प्रमाणित करता है कि लोकज्ञान के दस्तावेज़ीकरण से उसे दुरुपयोग और बौद्धिक संपदा के हनन से बचाया जा सकता है । TKDL ने भारतीय चिकित्सा प्रणालियों पर पेटेंट के दावों में 44% की कमी लाकर पारंपरिक ज्ञान के संरक्षण को एक कानूनी और आर्थिक आवश्यकता के रूप में स्थापित किया है । यह दर्शाता है कि एक विशाल, डिजिटल परियोजना के माध्यम से लोकज्ञान का संरक्षण केवल एक सांस्कृतिक लक्ष्य नहीं, बल्कि एक रणनीतिक राष्ट्रीय आवश्यकता भी है।

तालिका 1: भारत की कुछ संकटग्रस्त भाषाओं की स्थिति और उनके क्षेत्र

भाषा का नाम	भाषा परिवार	स्थिति	क्षेत्र
जरावा	अंदमानी	अत्यधिक गंभीर रूप से खतरे में	अंडमान और निकोबार द्वीप समूह
बिरहोर	मुंडा	अत्यधिक गंभीर रूप से खतरे में	महाराष्ट्र, मध्य प्रदेश
गुटोब / गदबा	मुंडा	भेद्य / निश्चित रूप से खतरे में	ओडिशा, आंध्र प्रदेश
टोडा	द्रविड़ियन (प्रोटो-दक्षिण)	निश्चित रूप से खतरे में	तमिलनाडु (नीलगिरी हिल्स)
निहाली	असम्बद्ध	अत्यधिक गंभीर रूप से खतरे में	महाराष्ट्र, मध्य प्रदेश
खम्बा	तिब्बती-बर्मी	भेद्य (खतरे से आशंकित)	अरुणाचल प्रदेश
बालटी	तिब्बती	भेद्य (खतरे से आशंकित)	जम्मू और कश्मीर (लद्दाख)
सेंटीनलीज़	अंदमानी	अत्यधिक गंभीर रूप से खतरे में	अंडमान और निकोबार द्वीप समूह

भाग 2: संरक्षण और पुनरुद्धार के लिए AI की क्षमताएं

2.1 तकनीकी क्षमताएं: AI, NLP और LLM

कृत्रिम बुद्धिमत्ता (AI) भाषा संरक्षण में एक शक्तिशाली उपकरण के रूप में उभरा है । प्राकृतिक भाषा प्रसंस्करण (NLP) AI की एक शाखा है जो मशीनों को मानव भाषा को समझने, व्याख्या करने और उत्पन्न करने में सक्षम बनाती है । जनरेटिव AI और बड़े भाषा मॉडल (LLM) ने इस क्षेत्र में क्रांति ला दी है, जिससे अभूतपूर्व नवाचार संभव हुए हैं.

AI-संचालित उपकरण दस्तावेज़ीकरण, डिजिटलीकरण और पुनरुद्धार के लिए कई अनुप्रयोग प्रदान करते हैं:

- **दस्तावेज़ीकरण:** AI-आधारित अनुवाद उपकरण और भाषण संश्लेषण (speech synthesis) तकनीकों का उपयोग कम-ज्ञात भाषाओं में मौखिक और लिखित रूपों को रिकॉर्ड करने और अनुवाद करने के लिए किया जा सकता है । यह भाषाई विशेषज्ञों और समुदायों को आवश्यक डिजिटल संसाधन बनाने में मदद करता है जो पहले अनुपलब्ध थे ।
- **डिजिटलीकरण:** डीप लर्निंग-आधारित कंप्यूटर विजन और ऑप्टिकल कैरेक्टर रिकॉग्निशन (OCR) तकनीकों का उपयोग प्राचीन पांडुलिपियों और ग्रंथों को डिजिटल रूप से संरक्षित करने के लिए किया जा रहा है । आईआईटी जोधपुर में चल रही एक परियोजना प्राचीन संस्कृत, पाली और तेलुगु पांडुलिपियों को साफ, बढ़ा और बहाल कर रही है, जिससे इस अमूल्य सांस्कृतिक विरासत को संरक्षित किया जा रहा है ।
- **पुनरुद्धार और शिक्षा:** जनरेटिव AI मॉडल शैक्षिक सामग्री, भाषा सीखने वाले ऐप्स, और यहां तक कि इंटरैक्टिव ट्यूटर भी उत्पन्न कर सकते हैं । यह नई पीढ़ियों को उनकी भाषाओं से जुड़ने का एक आकर्षक और सुलभ तरीका प्रदान करता है, जिससे भाषा का उपयोग और प्रसार पीढ़ी-दर-पीढ़ी होता रहता है ।

2.2 कम-संसाधन वाली भाषाओं के लिए विशेष दृष्टिकोण

AI उन भाषाओं के संरक्षण में भी सहायता कर सकता है जिनका कोई लिखित रूप नहीं है । AI सिस्टम मौखिक पैटर्न को पहचानकर उन्हें पाठ में बदल सकते हैं, जिससे मौखिक परंपराओं को दस्तावेज़ित किया जा सकता है । कम-संसाधन वाली भाषाओं के लिए डेटा की कमी एक महत्वपूर्ण बाधा है । इस समस्या को दूर करने के लिए, शोधकर्ता इन-कॉन्टेक्स्ट लर्निंग (ICL) और अनसुपरवाइज्ड प्रॉम्प्टिंग जैसी उन्नत तकनीकों का उपयोग कर रहे हैं। ये तरीके उच्च-संसाधन वाली भाषाओं (जैसे अंग्रेजी) से ज्ञान को कम-संसाधन वाली भाषाओं में स्थानांतरित करने की अनुमति देते हैं । यह दृष्टिकोण पारंपरिक तरीकों के लिए आवश्यक विशाल, हाथ से लेबल किए गए डेटासेट की आवश्यकता को कम करता है, जिससे सीमित संसाधनों के साथ भी उपयोगी AI मॉडल बनाना संभव हो जाता है । इस प्रकार, ये तकनीकें सीधे तौर पर प्री-ट्रेनिंग डेटा असंतुलन की मूल समस्या का समाधान करती हैं, जिससे भारत जैसे देशों में AI नवाचार का लोकतंत्रीकरण होता है।

भाग 3: भारत में प्रमुख AI-संचालित पहलें: एक विश्लेषण

3.1 स्वदेशी LLM परियोजनाएं: भारत जेन और आदि वाणी

भारत ने AI का लाभ उठाने के लिए कई स्वदेशी पहलों को अपनाया है। भारत जेन देश का पहला सरकार द्वारा वित्त पोषित, मल्टीमॉडल लार्ज लैंग्वेज मॉडल है, जिसे विशेष रूप से 22 भारतीय भाषाओं के लिए भारत-केंद्रित डेटासेट पर प्रशिक्षित किया गया है । इसका लक्ष्य सरकारी सेवाओं (शासन, स्वास्थ्य सेवा, शिक्षा) को तेज़ और अधिक कुशल बनाना है, साथ ही भारतीय भाषाओं में उच्च-गुणवत्ता वाला टेक्स्ट और मल्टीमॉडल कंटेंट उत्पन्न करना है ।

इसी तरह, आदि वाणी जनजातीय भाषाओं के लिए भारत का पहला AI-संचालित अनुवादक है । इसे जनजातीय अनुसंधान संस्थानों (TRIs) के सहयोग से विकसित किया गया है ताकि जनजातीय समुदायों को सशक्त बनाया जा सके और सरकारी योजनाओं व स्वास्थ्य सलाह जैसी महत्वपूर्ण जानकारी उनकी मूल भाषाओं में उपलब्ध कराई जा सके ।

ये पहलें सिर्फ तकनीकी नवाचार नहीं हैं; वे गूगल और ओपनएआई जैसी विदेशी कंपनियों पर भारत की निर्भरता को कम करने की एक रणनीतिक प्रतिक्रिया हैं । ये परियोजनाएँ भारत की भाषाई और डिजिटल संप्रभुता को सुनिश्चित करती हैं। उनका मुख्य उद्देश्य "AI for All" और

"डिजिटल समावेश" है, जो AI को ग्रामीण और जनजातीय क्षेत्रों तक ले जाकर अंतिम-मील कनेक्टिविटी सुनिश्चित करता है और डिजिटल विभाजन को कम करता है।

3.2 अन्य डिजिटल और भाषाई संरक्षण परियोजनाएं

भारत में अन्य महत्वपूर्ण परियोजनाएँ भी चल रही हैं जो AI और डिजिटल तकनीकों का उपयोग करती हैं। भारतवाणी एक प्रमुख भाषा प्रौद्योगिकी परियोजना है जिसका उद्देश्य भारतीय भाषाओं का संरक्षण और पुनरुद्धार करना है। पारंपरिक ज्ञान डिजिटल लाइब्रेरी (TKDL) ने पारंपरिक ज्ञान को दुरुपयोग से बचाने में अपनी सफलता साबित की है, जिसके परिणामस्वरूप भारतीय चिकित्सा प्रणालियों पर पेटेंट के दावों में 44% की कमी आई है। इसके अलावा, आईआईटी जोधपुर की एक परियोजना प्राचीन पांडुलिपियों को डिजिटाइज़ करने और बहाल करने के लिए कंप्यूटर विजन और OCR तकनीकों का उपयोग कर रही है, जिससे भारत की सांस्कृतिक विरासत को सुरक्षित रखा जा रहा है।

तालिका 2: भारत में प्रमुख AI भाषा और लोकज्ञान संरक्षण परियोजनाओं का तुलनात्मक विश्लेषण

परियोजना	नेतृत्व करने वाली संस्था	लक्ष्य	मुख्य प्रौद्योगिकियाँ	सामाजिक प्रभाव
भारत जेन	IIT बॉम्बे (NM-ICPS)	22 भारतीय भाषाओं में LLM विकसित करना, शासन और शिक्षा में सुधार।	मल्टीमॉडल LLM, जनरेटिव AI, भारत-केंद्रित डेटासेट।	डिजिटल संप्रभुता को बढ़ावा, AI का लोकतंत्रीकरण, डिजिटल समावेश।
आदि वाणी	जनजातीय मामलों का मंत्रालय, IIT दिल्ली	जनजातीय भाषाओं का संरक्षण, सरकार की सेवाओं को जनजातीय समुदायों तक पहुँचाना।	AI-संचालित अनुवाद (NLLB, IndicTrans2), भाषण-से-पाठ, OCR।	जनजातीय सशक्तिकरण, लास्ट-माइल सेवा वितरण, सांस्कृतिक संरक्षण।
TKDL	CSIR	पारंपरिक ज्ञान को दुरुपयोग से बचाना।	डिजिटल अभिलेखागार, पेटेंट डेटाबेस।	बौद्धिक संपदा की सुरक्षा, पेटेंट दुरुपयोग में 44% की कमी।
IIT जोधपुर	Vision, Language and Learning Group	प्राचीन पांडुलिपियों को डिजिटाइज़ और बहाल करना।	कंप्यूटर विजन, OCR, डीप लर्निंग।	सांस्कृतिक विरासत का संरक्षण, अनुसंधान और अनुवाद को सक्षम करना।

भाग 4: तकनीकी और भाषाई चुनौतियाँ

4.1 डेटा की कमी और गुणवत्ता

भारतीय भाषाओं के लिए AI मॉडल विकसित करने में सबसे बड़ी तकनीकी बाधा प्रशिक्षण के लिए पर्याप्त, उच्च-गुणवत्ता वाले डेटासेट का अभाव है। जबकि अंग्रेजी के पास अरबों डिजिटाइज्ड टेक्स्ट हैं, अधिकांश भारतीय भाषाओं का डिजिटल पदचिह्न कमजोर है। कई भारतीय भाषाओं में एक मजबूत मौखिक परंपरा है, लेकिन उनका लिखित रिकॉर्ड सीमित है, जिससे पारंपरिक वेब-क्रॉलिंग विधियाँ अप्रभावी हो जाती हैं। यह डेटा की कमी AI की भाषाई पैटर्न सीखने की क्षमता को सीमित करती है, जिसके परिणामस्वरूप अनुवाद और टेक्स्ट जनरेशन में अशुद्धियाँ आती हैं।

4.2 भाषाई जटिलताएँ

भारतीय भाषाओं की अनूठी संरचना AI के लिए विशेष चुनौतियाँ पेश करती है:

- **स्क्रिप्ट की जटिलता:** कई भारतीय भाषाएँ एक ही लिपि साझा करती हैं, जैसे कि हिंदी, मराठी, नेपाली और संस्कृत सभी देवनागरी लिपि में लिखी जाती हैं। यह AI के लिए केवल वर्तनी के आधार पर भाषाओं में अंतर करना मुश्किल बना देता है।
- **कोड-स्विचिंग और बोली भिन्नता:** भारत में कोड-स्विचिंग (दो या दो से अधिक भाषाओं को मिलाना, जैसे हिंग्लिश) एक आम बात है। AI मॉडल, जो अक्सर एकभाषी डेटासेट पर प्रशिक्षित होते हैं, इन जटिलताओं को समझने में संघर्ष करते हैं। उदाहरण के लिए, एक वाक्य में Amazing जैसे अंग्रेजी शब्द को मराठी व्याकरण के नियमों के साथ समायोजित करने में AI को कठिनाई होती है। इन चुनौतियों के कारण AI द्वारा उत्पन्न अनुवाद अक्सर रोबोटिक और अप्राकृतिक लगते हैं।

4.3 तकनीकी सीमाएँ और संसाधन

बड़े भाषा मॉडल को प्रशिक्षित करने के लिए महत्वपूर्ण कंप्यूटेशनल संसाधनों (GPU और बड़ी मात्रा में बिजली) की आवश्यकता होती है। भारत जैसे विकासशील देश के लिए, उच्च-अंत वाले GPUs और आवश्यक बुनियादी ढांचे तक पहुँच एक महत्वपूर्ण बाधा हो सकती है। यह तकनीकी अंतर भारत को AI की वैश्विक दौड़ में पीछे छोड़ सकता है, भले ही नीतिगत और रणनीतिक इरादे मजबूत हों।

भाग 5: नैतिक, सामाजिक और नीतिगत चुनौतियाँ

5.1 सांस्कृतिक पूर्वाग्रह और एल्गोरिथम होमोजेनाइजेशन

वैश्विक AI मॉडल अक्सर अंग्रेजी-केंद्रित डेटा पर प्रशिक्षित होते हैं, जिसके परिणामस्वरूप भारतीय संदर्भ और

सांस्कृतिक बारीकियों के प्रति गहरा पूर्वाग्रह होता है। एक शोध से पता चला है कि GPT-3.5 जैसे मॉडल भारतीय कहानियों के बारे में जानकारी देने में विफल रहे, जिसके परिणामस्वरूप भ्रामक (hallucinated) और गलत जानकारी उत्पन्न हुई। यह इंगित करता है कि AI केवल डेटा को दोहराता नहीं है; यह सांस्कृतिक संदर्भ की कमी के कारण वास्तविक ज्ञान को विकृत कर सकता है, जिससे गैर-प्रमुख संस्कृतियों का लोप या सीमांतीकरण हो सकता है। विदेशी AI मॉडलों पर निर्भरता डिजिटल सांस्कृतिक उपनिवेशवाद को बढ़ावा दे सकती है, जहाँ प्रमुख भाषाएँ और संस्कृतियाँ एल्गोरिदम के माध्यम से हावी होती हैं।

5.2 डेटा स्वामित्व और संप्रभुता

जनजातीय और स्वदेशी समुदायों से प्राप्त भाषाई डेटा और ज्ञान का स्वामित्व और उपयोग एक महत्वपूर्ण नैतिक मुद्दा है। बिना उचित सहमति के डेटा का उपयोग औपनिवेशिक पैटर्न को कायम रख सकता है और समुदायों को उनकी ही विरासत पर नियंत्रण से वंचित कर सकता है। भारत का डेटा एम्पावरमेंट एंड प्रोटेक्शन आर्किटेक्चर (DEPA) जैसे मौजूदा ढाँचे सहमति-आधारित डेटा साझाकरण पर ध्यान केंद्रित करते हैं, जो एक सकारात्मक कदम है। हालाँकि, AI के लिए एक विशिष्ट राष्ट्रीय डेटा संरक्षण कानून और नियामक ढाँचे की अनुपस्थिति एक महत्वपूर्ण नीतिगत कमी है, जो डेटा के दुरुपयोग और गोपनीयता संबंधी चिंताओं को संबोधित करने में कमजोर है।

5.3 नीतिगत और फंडिंग संबंधी चुनौतियाँ

भारत में एक व्यापक राष्ट्रीय AI रणनीति और पर्याप्त वित्तीय निवेश की कमी है। हालाँकि इंडिया AI मिशन के लिए फंडिंग बढ़ी है, यह अभी भी अमेरिका और चीन जैसे वैश्विक नेताओं से बहुत पीछे है। यह वित्तीय अंतर AI के क्षेत्र में भारत की आत्मनिर्भरता के लक्ष्यों को बाधित कर सकता है। इसके अलावा, राष्ट्रीय शिक्षा नीति, 2020 जैसे प्रावधानों के बावजूद, जिन्हें स्थानीय भाषाओं में शिक्षा को बढ़ावा देने के लिए डिज़ाइन किया गया है, उनका प्रभावी कार्यान्वयन एक चुनौती बना हुआ है।

भाग 6: रणनीतिक समाधान और भविष्य की दिशा

भारतीय भाषाओं और लोकज्ञान के संरक्षण के लिए एक सफल रणनीति को तकनीकी नवाचार, नीतिगत समर्थन और सामुदायिक भागीदारी का एक बहु-आयामी दृष्टिकोण अपनाना होगा।

- **डेटा संग्रह और मानकीकरण:** पारंपरिक डेटा संग्रह विधियों से परे, समुदाय-आधारित, विकेन्द्रीकृत डेटा संग्रह मॉडल को अपनाया जाना चाहिए, जहाँ समुदाय अपने डेटा का नियंत्रण बनाए रखते हैं। इसके अतिरिक्त, कोड-मिश्रण और बोली भिन्नताओं के लिए मानकीकृत एनोटेटेड डेटासेट बनाने के लिए अनुसंधान को प्रोत्साहित करना महत्वपूर्ण है।
- **ओपन-सोर्स और सहयोगात्मक विकास:** भारत जेन जैसे ओपन-सोर्स प्लेटफॉर्म और भाषिनी चैलेंज जैसी पहलों को बढ़ावा देना आवश्यक है, जो शोधकर्ताओं, स्टार्टअप्स और उद्योग को नवाचार में भाग लेने के लिए प्रोत्साहित करती हैं। यह मॉडल AI के विकास को लोकतांत्रिक बनाता है और भारत की विशिष्ट आवश्यकताओं के अनुरूप समाधानों को बढ़ावा देता है।
- **नीतिगत सिफारिशें:** AI के लिए एक मजबूत नियामक ढाँचा स्थापित करना महत्वपूर्ण है जो डेटा संप्रभुता, बौद्धिक संपदा अधिकारों, और एल्गोरिथम पूर्वाग्रह को संबोधित करे। राष्ट्रीय शिक्षा नीति, 2020 के प्रावधानों को प्रभावी ढंग से लागू करना भी आवश्यक है, ताकि जनजातीय और अल्पसंख्यक भाषाओं में शिक्षा को बढ़ावा दिया जा सके, जिससे भाषा के जीवित रहने और अगली पीढ़ी तक हस्तांतरित होने की संभावना बढ़ जाती है।
- **फंडिंग मॉडल:** इंडिया AI मिशन जैसे कार्यक्रमों के लिए सरकारी फंडिंग में उल्लेखनीय वृद्धि आवश्यक है। इसके साथ ही, निजी क्षेत्र, शैक्षणिक संस्थानों और अंतर्राष्ट्रीय संगठनों के बीच भागीदारी को बढ़ावा देना भी महत्वपूर्ण है ताकि भाषा संरक्षण के प्रयासों को एक सतत और स्केलेबल मॉडल प्रदान किया जा सके।

तालिका 3: चुनौतियों और उनके प्रस्तावित समाधानों का सारांश

चुनौती	संक्षिप्त विवरण	प्रस्तावित समाधान
डेटा की कमी	अधिकांश भारतीय भाषाओं का सीमित डिजिटल पदचिह्न है।	समुदाय-आधारित डेटा संग्रह, मौखिक परंपराओं का दस्तावेज़ीकरण।
भाषाई जटिलताएँ	एक ही लिपि साझा करना, कोड-स्विचिंग और बोली भिन्नता।	एनोटेटेड डेटासेट बनाना, क्रॉस-लिंगुअल लर्निंग तकनीकों का उपयोग।
सांस्कृतिक पूर्वाग्रह	वैश्विक AI मॉडल अंग्रेजी-केंद्रित डेटा पर प्रशिक्षित होते हैं, जिससे सांस्कृतिक अज्ञानता होती है।	भारत-केंद्रित डेटासेट पर स्वदेशी मॉडल का प्रशिक्षण, सामुदायिक भागीदारी।
डेटा स्वामित्व	जनजातीय और स्वदेशी ज्ञान का अनियंत्रित उपयोग।	डेटा संप्रभुता और सहमति-आधारित डेटा साझाकरण के लिए नियामक ढाँचा।
फंडिंग और नीतिगत अंतराल	अपर्याप्त सरकारी फंडिंग और व्यापक राष्ट्रीय AI रणनीति की कमी।	फंडिंग में वृद्धि, सार्वजनिक-निजी भागीदारी, मजबूत नियामक ढाँचा।

निष्कर्ष

AI, भारतीय भाषाओं और लोकज्ञान के संरक्षण में एक अपरिहार्य और परिवर्तनकारी उपकरण है। भारत जेन, आदि वाणी, और TKDL जैसी सरकारी और शैक्षणिक पहलें इस दिशा में एक महत्वपूर्ण कदम का प्रतिनिधित्व करती हैं, जो भारत को तकनीकी आत्मनिर्भरता और डिजिटल संप्रभुता की ओर ले जा रही हैं। यह रिपोर्ट तकनीकी नवाचार के साथ-साथ एक समग्र दृष्टिकोण की आवश्यकता पर बल देती है। AI के सामने मौजूद डेटा की कमी, भाषाई जटिलता, और सांस्कृतिक पूर्वाग्रह जैसी तकनीकी चुनौतियों को डेटा संप्रभुता, सामुदायिक भागीदारी, और पर्याप्त फंडिंग जैसी सामाजिक और नीतिगत चुनौतियों के साथ मिलकर ही संबोधित किया जा सकता है।

AI को केवल एक तकनीकी समाधान के रूप में नहीं, बल्कि एक सहायक उपकरण के रूप में उपयोग करके, भारत अपनी भाषाई और सांस्कृतिक पहचान को सशक्त बना सकता है। यह सुनिश्चित करेगा कि सदियों का ज्ञान न केवल डिजिटल युग में जीवित रहे, बल्कि भविष्य की पीढ़ियों के लिए भी सुलभ और जीवंत बना रहे। यह एक मिशन है जो AI को तकनीकी नवाचार से परे, राष्ट्रीय विरासत और समावेशिता के एक साधन के रूप में स्थापित करता है।

References

1. भारत में हैं 453 भाषाएं, कितनी भाषाओं को मिली है आधिकारिक मान्यता? यहां जानिए डिटेल - Jagran, https://www.jagran.com/news/national-languages-india-official-recognition-know-details-23889285.html
2. Documenting India's Endangered Languages - Press Information ..., https://www.pib.gov.in/FeaturesDeatils.aspx?id=155013&NoteId=155013&ModuleId=2
3. भारत की कौन सी भाषाएं लुप्त होती जा रही हैं? इनके लुप्त होने के पीछे क्या कारण है? - Quora, https://hi.quora.com/%E0%A4%AD%E0%A4%BE%E0%A4%B0%E0%A4%A4-%E0%A4%95%E0%A5%80-%E0%A4%95%E0%A5%8C%E0%A4%A8-%E0%A4%B8%E0%A5%80-%E0%A4%AD%E0%A4%BE%E0%A4%B7%E0%A4%BE%E0%A4%8F%E0%A4%82-%E0%A4%B2%E0%A5%81%E0%A4%AA%E0%A5%8D%E0%A4%A4
4. लुप्तप्राय भाषाओं की सुरक्षा और संरक्षण के लिये योजना - Drishti IAS, https://www.drishtiias.com/hindi/daily-news-analysis/protection-and-preservation-of-endangered-language
5. Digital Archives And Preservation Of Traditional ..., https://ijirl.com/wp-content/uploads/2025/01/Digital-Archives-And-Preservation-Of-Traditional-Knowledge.pdf
6. भारत की विलुप्त होती बोलियां और भाषाएं - प्रवक्ता.कॉम - Pravakta.Com, https://www.pravakta.com/the-dissapearing-languages-and-dialects-of-india/
7. भारत में संकटग्रस्त भाषाओं की सूची - विकिपीडिया, https://hi.wikipedia.org/wiki/%E0%A4%AD%E0%A4%BE%E0%A4%B0%E0%A4%A4_%E0%A4%AE%E0%A5%87%E0%A4%82_%E0%A4%B8%E0%A4%82%E0%A4%95%E0%A4%9F%E0%A4%97%E0%A5%8D%E0%A4%B0%E0%A4%B8%E0%A5%8D%E0%A4%A4_%E0%A4%AD%E0%A4%BE%E0%A4%B7%E0%A4%BE%E0%A4%93%E0%A4%82_%E0%A4%95%E0%A5%80_%E0%A4%B8%E0%A5%82%E0%A4%9A%E0%A5%80
8. Endangered Languages of India - ClearIAS, https://www.clearias.com/endangered-languages/
9. Generative AI and Large Language Models in Language Preservation: Opportunities and Challenges - arXiv, https://arxiv.org/html/2501.11496v1
10. Machine Learning for Endangered TRENDS Research & Advisory, https://trendsresearch.org/insight/machine-learning-for-endangered-language-preservation/