2.8 ट्रिलियन टोकन: भारतीय भाषाओं के लिए सबसे बड़ा खुला AI डेटा-कोष क्या बदलता है

विज्ञान-प्रौद्योगिकी / डिजिटल समावेशन (मुख्य परीक्षा: सामान्य अध्ययन-III) | समाचार संदर्भ: तकनीकी रिपोर्ट arXiv (19 अक्टूबर 2026) और आधिकारिक ब्लॉग (16 अक्टूबर 2026) | प्रकाशित: 24 अक्टूबर 2026

भारतीय भाषाओं के बड़े भाषा-मॉडल (LLM) के लिए अब तक का सबसे बड़ा खुला प्री-ट्रेनिंग डेटा-कोष सामने आया है: Nemotron-Pretraining-Indic-v1 (NPI-v1) – लगभग 2.83 ट्रिलियन टोकन, 22 अनुसूचित भारतीय भाषाओं में देवनागरी-सरीखी मूल लिपियों में, और 12 भाषाओं में रोमन लिपि में। NVIDIA और IIT मद्रास की AI4Bharat प्रयोगशाला के संयुक्त कार्य का यह कोष और उसका निर्माण-पाइपलाइन खुले स्रोत में जारी किया गया है; तकनीकी रिपोर्ट arXiv पर 19 अक्टूबर 2026 को और घोषणा-ब्लॉग 16 अक्टूबर को आया। ध्यान दें: यह प्रीप्रिंट (सहकर्मी-समीक्षा से पहले की रिपोर्ट) और डेटासेट-घोषणा है, कोई 'सुपर-मॉडल' लॉन्च नहीं – लेखक खुद कहते हैं कि बने चेकपॉइंट प्रयोग-उपकरण हैं, सर्वोत्तम-मॉडल रिलीज़ नहीं।

खबरों में क्यों

AI की दौड़ में भारतीय भाषाएं डेटा-गरीबी से जूझ रही हैं। रिपोर्ट की तुलना खुलने वाली है: अंग्रेजी के लिए खुले FineWeb जैसे कोषों में 14 ट्रिलियन से अधिक शब्द हैं, जबकि बहुभाषी FineWeb2 में हिंदी के लिए मात्र 11.17 अरब शब्द – यानी हजार-गुना से अधिक की खाई। बिना डेटा के मॉडल नहीं बनता; और बिना मॉडल के नागरिक-सेवा, शिक्षा, स्वास्थ्य और न्याय तक AI की पहुंच भाषा-बाधा में अटकी रहती है। 22 अनुसूचित भाषाएं करीब 97% भारतीयों की प्रथम भाषा हैं (लगभग 1.2 अरब लोग, 2011 की जनगणना – पंद्रह वर्ष पुराना आकलन, यह ध्यान में रखें)। सरकारी परिदृश्य में भी भाषाई AI जोरदार है: डिजिटल इंडिया भाषिणी (राष्ट्रीय भाषा अनुवाद मिशन, MeitY) के तहत ही इसी माह BHASHINI और IISc बेंगलुरु ने तमिलनाडु की अल्पसंख्यक सौराष्ट्र भाषा के लिए टेक्स्ट-टू-स्पीच मॉडल बनाए। ऐसे में एक खुला, स्रोत-अनुरेखणीय, 22-भाषा कोष 'डिजिटल सार्वजनिक संपत्ति' के उम्मीदवार के रूप में नीति-चर्चा में आता है – साथ ही निजी-सार्वजनिक साझेदारी, डेटा-संप्रभुता और कृत्रिम डेटा की गुणवत्ता के सवाल भी।

तरीका: अनुवाद को गुणक बनाने वाली पाइपलाइन

समस्या सरल है: कई भारतीय भाषाओं में प्राकृतिक रूप से उपलब्ध गुणवत्तापूर्ण पाठ अत्यंत कम है। समाधान-रणनीति: क्युरेट किए गए अंग्रेजी दस्तावेज़ों (गणित, कोड, प्रश्नोत्तर) को GPU-त्वरित पाइपलाइन से अनुवाद करना – पर रुपये-की-चीज़ नहीं, गुणवत्ता-नियंत्रण के साथ। क्रम: स्रोत-नमूना चयन, बड़े पैमाने पर अनुवाद (gemma-4-31b-it और indic-translate जैसे मॉडलों से), भाषा-विशिष्ट सीमा-मानों के साथ स्वचालित गुणवत्ता-मूल्यांकन, खराब आउटपुट की लक्षित मरम्मत (रीराइटिंग), रोमनकरण, और अंत में छंटनी व डुप्लिकेट-हटाना। हर रिकॉर्ड अपने स्रोत-दस्तावेज़ तक अनुरेखणीय है, और हर आउटपुट अपने इनपुट से जांचा गया – यानी कृत्रिम पाठ वास्तविक सूचना से बंधा रहता है। प्राकृतिक पाठ-भाग Sangraha (AI4Bharat का पूर्व कोष), FineWeb2 और FinePDFs से आया। साथ ही बनाया गया Indic-LLM-Eval-Suite: मौजूदा भारतीय बेंचमार्क (जैसे MILU – भारत-केंद्रित ज्ञान) और अंग्रेजी बेंचमार्क के सावधान अनुवाद (MMLU-Pro, ARC, GSM8K), मूल-वक्ता समीक्षा के साथ; मूल्यांकन NVIDIA के NeMo-Gym में एकीकृत।

इसे एक तुलना से समझें: अंग्रेजी मॉडलों के लिए डेटा की समस्या 'छंटनी' की है – क्या छोड़ें; भारतीय भाषाओं के लिए समस्या 'कमी' की है – क्या बनाएं। इसीलिए रिपोर्ट अनुवाद को 'बल-गुणक' कहती है: एक बार क्युरेट किया गया अंग्रेजी दस्तावेज़ 22 भाषाओं में ज्ञान ले जाता है, बशर्ते गुणवत्ता-जाल काम करे। यही दृष्टिकोण पिछले दशक के मशीन-अनुवाद अनुसंधान (AI4Bharat के IndicTrans जैसे) को उत्पादन-अवसंरचना में बदलता है – अनुवाद अब सिर्फ सेवा नहीं, डेटा-फैक्ट्री है।

मिला क्या: डेटा-पर्याप्तता का प्रयोगात्मक प्रमाण

मुख्य प्रयोग: 12 भाषाओं में निरंतर-प्रीट्रेनिंग (कंटिन्यूड प्री-ट्रेनिंग, CPT) – 100 अरब टोकन के प्रशिक्षण-बजट में 10% (10 अरब टोकन) कृत्रिम+प्राकृतिक भारतीय पाठ मिलाकर। परिणाम: सार्वजनिक-कोष आधार-रेखा की तुलना में भारतीय ज्ञान और तर्क में सुधार; ब्लॉग के अनुसार पांचों रिपोर्ट किए गए बेंचमार्क – MILU, Indic MMLU-Pro, Indic ARC-Challenge, Indic ARC-Easy और Indic GSM8K – में अंक बढ़े। साथी प्रयोगों में डेटा-मात्रा, लिपि-कवरेज और भारतीय-हिस्सेदारी का अंग्रेजी प्रदर्शन पर असर मापा गया – यानी 'भारतीय डेटा डालने से अंग्रेजी क्षमता घटती है या नहीं' का प्रश्न भी संबोधित। रिपोर्ट पाइपलाइन की विफलता-प्रणालियां भी दर्ज करती है: अनुवाद अर्थ बदल सकता है, दस्तावेज़-संरचना टूट सकती है, रोमनकरण में असंगत वर्तनी आ सकती है – इसलिए केवल प्रवाही पाठ काफी नहीं, सूचना-संरक्षण जरूरी।

Eval-Suite से एक और उपयोगी खुलासा: मौजूदा खुले भारतीय और बहुभाषी मॉडलों की तुलनात्मक बेंचमार्किंग में कार्यों और भाषाओं के बीच 'भारी असमानता' दिखी – कोई एक मॉडल सभी भाषाओं-कार्यों में सर्वश्रेष्ठ नहीं। यानी भाषा-कवरेज का दावा और वास्तविक क्षमता अलग-अलग बातें हैं; खरीददारी-निर्णय (सरकारी प्रयोग) के लिए भाषा-वार बेंचमार्क अनिवार्य।

सीमाएँ और सावधानियाँ

कई बातें संयत रखें। पहली, यह प्रीप्रिंट है – arXiv पर सहकर्मी-समीक्षा से पहले; 'सबसे बड़ा खुला कोष' दावा लेखकों का अपना है, स्वतंत्र तुलना-लेखा बाकी। दूसरी, कोष अधिकांशत: कृत्रिम-अनुवादित है – यह ताकत भी और सीमा भी: लेखक खुद मानते हैं कि अनुवाद अर्थ-विकृति, अस्वाभाविक भाषा-वितरण और सांस्कृतिक-संदर्भ हानि ला सकता है; मशीन-अनुवादित बेंचमार्क में 'कार्य-संरक्षण' का सवाल उन्होंने स्वयं उठाया है। तीसरी, डेटासेट-पेज की जांच के समय सार्वजनिक पहुंच सीमित/अनुपलब्ध थी (ब्लॉग कहता है 'जल्द जारी') – 'खुला' दावे की पूर्ण पुष्टि रिलीज़ के बाद ही संभव। चौथी, कॉरपोरेट सह-लेखन: पाइपलाइन, अनुवाद-मॉडल, मूल्यांकन-ढांचा (NeMo-Gym) सब NVIDIA-परिसर के – संसाधन-दान और स्व-हित के बीच सीमा पाठक तय करे; स्वतंत्र बाहरी मूल्यांकन अभी नहीं। पाँचवीं, प्रयोग-मॉडल 'प्रयोग-उपकरण' हैं, उत्पाद नहीं – बेंचमार्क-बढ़त किसी तैनात सेवा की गारंटी नहीं। छठी, '97%/1.2 अरब' आंकड़ा 2011 जनगणना का है – भाषाई प्रोफाइल पुरानी; और CPT में केवल 12 भाषाएं परखी गईं, 22 नहीं। सातवीं, रोमन लिपि कवरेज 12 भाषाओं तक सीमित; और गुणवत्ता-सीमा मानव-समीक्षा आंशिक (पाइपलाइन मुख्यत: स्वचालित)। आठवीं, तिथि-परतें: ब्लॉग 16 अक्टूबर, arXiv v1 19 अक्टूबर 2026 – दोनों अलग; BHASHINI-सौराष्ट्र घोषणा 16 अक्टूबर का पृष्ठभूमि-समाचार है, इस कोष से अलग परियोजना।

आगे क्या देखना होगा?

तीन चीजें: (1) वास्तविक रिलीज़ – डेटासेट और रेसिपी का सार्वजनिक डाउनलोड-पृष्ठ खुलने पर समुदाय का स्वतंत्र ऑडिट; (2) क्या भाषिणी/IndiaAI मिशन जैसे सार्वजनिक प्रयास इस कोष या पाइपलाइन को अपनाते हैं – सार्वजनिक-निजी सीमा-रेका का असली परीक्षण वहीं होगा; (3) स्वतंत्र बेंचमार्किंग – Indic-LLM-Eval-Suite पर तीसरे-पक्ष के मॉडल-मूल्यांकन, और कृत्रिम बनाम प्राकृतिक डेटा के दीर्घकालिक प्रभाव के अध्ययन (मॉडल-पतन/भाषा-विकृति जोखिम)। नीति-दृष्टि से, यह काम 'भाषाई डेटा को सार्वजनिक अवसंरचना मानो' के तर्क को ठोस उदाहरण देता है – अगला बहस-बिंदु होगा ऐसे कोषों का शासन: कौन रखरखाव करे, किसके खर्च पर, किस जवाबदेही के साथ।

परीक्षा-दृष्टि से एक और संकेत: यह रिलीज़ उस बहस को आगे बढ़ाती है कि 'सार्वजनिक अवसंरचना' की परिभाषा में अब डेटा-सेट और पाइपलाइन भी शामिल हैं – जैसे UPI ने भुगतान-रेल दी, वैसे खुले भाषा-कोष भाषाई AI की रेल दे सकते हैं। पर अंतर भी समझें: UPI सार्वजनिक स्वामित्व में था; NPI-v1 निजी-सार्वजनिक सह-निर्माण है – लाइसेंस, रखरखाव और संस्करण-नियंत्रण के नियम इसके 'सार्वजनिक-अच्छे' चरित्र को तय करेंगे।

पाठक के लिए संदर्भ-बिंदु

चार बातें याद रखें। पहली, टोकन: मॉडल की पठन-इकाई – मोटे तौर पर शब्द-खंड; ट्रिलियन टोकन का आकार ही आधुनिक LLM-शक्ति का पहला ईंधन-माप है। दूसरी, प्री-ट्रेनिंग बनाम CPT: बुनियादी प्रशिक्षण और उसके बाद क्षेत्र-विशेष डेटा से अतिरिक्त प्रशिक्षण। तीसरी, अनुसूचित भाषाएं: संविधान की आठवीं अनुसूची की 22 भाषाएं – चार भाषा-परिवारों से; BHASHINI मिशन इसी समूह के डिजिटल-अनुवाद पर केंद्रित। चौथी, AI4Bharat: IIT मद्रास की ओपन-सोर्स भारतीय-भाषा AI प्रयोगशाला – IndicCorp, IndicBERT, Sangraha, IndicTrans2 इसकी पूर्व-परियोजनाएं; संस्था-परियोजना युग्म प्रिलिम्स में पूछा जाता रहा है।

मुख्य परीक्षा अभ्यास प्रश्न

'भाषाई डेटा-असमानता भारत के AI-लोकतंत्रीकरण की सबसे बड़ी संरचनात्मक बाधा है।' Nemotron-Indic जैसे खुले कोष-निर्माण प्रयासों के आलोक में इस कथन का मूल्यांकन कीजिए और सार्वजनिक-निजी साझेदारी में डेटा-शासन के सिद्धांत सुझाइए। (सामान्य अध्ययन-III: विज्ञान-प्रौद्योगिकी; 250 शब्द, 15 अंक)

उत्तर की रूपरेखा:

भूमिका में: डेटा-खाई की तुलना रखें: अंग्रेजी के लिए खुले FineWeb जैसे कोषों में 14 ट्रिलियन से अधिक शब्द हैं, जबकि बहुभाषी FineWeb2 में हिंदी के लिए 11.17 अरब शब्द। फिर Nemotron-Pretraining-Indic-v1 (NPI-v1) रखें: लगभग 2.83 ट्रिलियन टोकन, 22 अनुसूचित भाषाएँ मूल लिपि में और 12 भाषाएँ रोमन लिपि में, NVIDIA और IIT मद्रास की AI4Bharat प्रयोगशाला का संयुक्त कार्य। साथ में बताएँ कि यह प्रीप्रिंट और डेटासेट-घोषणा है, तैयार सुपर-मॉडल नहीं।

मुख्य भाग में: कथन का मूल्यांकन करें। पक्ष में: अनुवाद-आधारित पाइपलाइन (क्युरेट किए अंग्रेजी दस्तावेज़ों का गुणवत्ता-नियंत्रण के साथ अनुवाद, हर रिकॉर्ड स्रोत-दस्तावेज़ तक अनुरेखणीय) और 12 भाषाओं के निरंतर-प्रीट्रेनिंग प्रयोग में 100 अरब टोकन के बजट में 10 अरब टोकन भारतीय पाठ मिलाने पर पाँचों रिपोर्ट किए गए बेंचमार्क में अंकों का बढ़ना। सीमाएँ रखें: यह प्रीप्रिंट है और सबसे बड़ा खुला कोष होने का दावा लेखकों का अपना है; कोष अधिकांशत: कृत्रिम अनुवादित है, जिसमें अर्थ-विकृति और सांस्कृतिक-संदर्भ की हानि का जोखिम है; डेटासेट की सार्वजनिक पहुँच की पूर्ण पुष्टि रिलीज़ के बाद ही संभव है; पाइपलाइन, अनुवाद-मॉडल और मूल्यांकन-ढाँचा NVIDIA-परिसर के हैं और स्वतंत्र बाहरी मूल्यांकन अभी नहीं है; प्रयोग केवल 12 भाषाओं में हुए; 97 प्रतिशत और 1.2 अरब का आँकड़ा 2011 की जनगणना का है। यह भी लिखें कि लेख डेटा-पक्ष के प्रमाण देता है, इसलिए अन्य बाधाओं पर अपने आँकड़े न गढ़ें। डेटा-शासन के सिद्धांत लेख के बिंदुओं तक रखें: स्रोत-अनुरेखणीयता, लाइसेंस और खुलापन, रखरखाव और संस्करण-नियंत्रण, यह तय करना कि रखरखाव कौन करे, किसके खर्च पर और किस जवाबदेही के साथ, स्वतंत्र तीसरे-पक्ष का मूल्यांकन और भाषा-वार बेंचमार्क। निजी-सार्वजनिक सह-निर्माण के कारण इसे सार्वजनिक स्वामित्व वाली अवसंरचना की तरह न पढ़ें।

निष्कर्ष में: निष्कर्ष में कहें कि NPI-v1 जैसे प्रयास भाषाई डेटा को सार्वजनिक अवसंरचना मानने के तर्क को ठोस उदाहरण देते हैं, पर कथन की पूरी पुष्टि के लिए वास्तविक रिलीज़, स्वतंत्र बेंचमार्किंग और कृत्रिम बनाम प्राकृतिक डेटा के दीर्घकालिक प्रभाव का इंतजार है। अंतिम बहस-बिंदु शासन का रखें, और अपनी ओर से कोई कानून, योजना या आँकड़ा न जोड़ें।

स्रोत और सीमाएँ

- प्राथमिक तकनीकी रिपोर्ट (arXiv प्रीप्रिंट): NVIDIA & AI4Bharat, "Scaling Synthetic Pre-training Data and Evaluation for Indic LLMs", arXiv:2610.26006v1 (19 अक्टूबर 2026) – पूर्ण पाठ पठित; कोष-विवरण, पाइपलाइन, CPT-प्रयोग, Eval-Suite, लेखकों की अपनी विफलता-सूची।

- आधिकारिक घोषणा-ब्लॉग (Hugging Face/NVIDIA, 16 अक्टूबर 2026): लेखक-सूची (AI4Bharat: दोड्डपनेनी, खापरा, कुंचुकुट्टन आदि; NVIDIA टीम), 97%/1.2 अरब (2011 जनगणना) संदर्भ, पांच-बेंचमार्क बढ़त, 'जल्द जारी' रिलीज़-स्थिति।
- AI4Bharat संस्थागत पृष्ठ (ai4bharat.org): प्रयोगशाला-परिचय, पूर्व-परियोजनाएं (IndicBERT, Sangraha, IndicTransv2 आदि)।
- NVIDIA मॉडल-कार्ड (Hugging Face): Nemotron 3.5 Super VL में Indic CPT चरण (10 अरब टोकन हिस्सा) – प्रयोग-परिप्रेक्ष्य।
- आकाशवाणी समाचार (16 अक्टूबर 2026): BHASHINI + IISc बेंगलुरु का सौराष्ट्र TTS मॉडल – सरकारी भाषाई-AI परिदृश्य (अलग परियोजना, संदर्भ मात्र)।
- Bhashini.gov.in: राष्ट्रीय भाषा अनुवाद मिशन (MeitY) आधिकारिक पृष्ठ – कार्यक्रम-संदर्भ।
- सीमाएँ: प्रीप्रिंट (समीक्षा-पूर्व); 'सबसे बड़ा' दावा लेखकों का; कोष मुख्यत: कृत्रिम-अनुवादित (अर्थ-विकृति जोखिम लेखक-स्वीकृत); डेटासेट-पहुंच जांच-समय सीमित ('जल्द जारी'); NVIDIA कॉरपोरेट सह-लेखन/NeMo-Gym एकीकरण; CPT केवल 12/22 भाषाओं में; 97% आंकड़ा 2011 जनगणना का; स्वतंत्र मूल्यांकन बाकी; मशीन-अनुवादित बेंचमार्क की कार्य-संरक्षण सीमा लेखक-स्वीकृत।

अक्सर पूछे जाने वाले प्रश्न (FAQ)

संक्षिप्त उत्तर: भारतीय भाषाओं के बड़े भाषा-मॉडल (LLM) के लिए अब तक का सबसे बड़ा खुला प्री-ट्रेनिंग डेटा-कोष सामने आया है: Nemotron-Pretraining-Indic-v1 (NPI-v1) – लगभग 2.83 ट्रिलियन टोकन, 22 अनुसूचित भारतीय भाषाओं में देवनागरी-सरीखी मूल लिपियों में, और 12 भाषाओं में रोमन लिपि में। NVIDIA और IIT मद्रास की AI4Bharat प्रयोगशाला के संयुक्त कार्य का यह कोष और उसका निर्माण-पाइपलाइन खुले स्रोत में जारी किया गया है; तकनीकी रिपोर्ट arXiv पर 19 अक्टूबर 2026 को और घोषणा-ब्लॉग 16 अक्टूबर को आया।

प्रश्न: Nemotron-Pretraining-Indic-v1 क्या है?

भारतीय भाषाओं के बड़े भाषा-मॉडल (LLM) के लिए अब तक का सबसे बड़ा खुला प्री-ट्रेनिंग डेटा-कोष सामने आया है: Nemotron-Pretraining-Indic-v1 (NPI-v1) – लगभग 2.83 ट्रिलियन टोकन, 22 अनुसूचित भारतीय भाषाओं में देवनागरी-सरीखी मूल लिपियों में, और 12 भाषाओं में रोमन लिपि में। NVIDIA और IIT मद्रास की AI4Bharat प्रयोगशाला के संयुक्त कार्य का यह कोष और उसका निर्माण-पाइपलाइन खुले स्रोत में जारी किया गया है; तकनीकी रिपोर्ट arXiv पर 19 अक्टूबर 2026 को और घोषणा-ब्लॉग 16 अक्टूबर को आया।

प्रश्न: भारतीय भाषा डेटा बनाने की अनुवाद-रणनीति क्या है?

समस्या सरल है: कई भारतीय भाषाओं में प्राकृतिक रूप से उपलब्ध गुणवत्तापूर्ण पाठ अत्यंत कम है। समाधान-रणनीति: क्युरेट किए गए अंग्रेजी दस्तावेज़ों (गणित, कोड, प्रश्नोत्तर) को GPU-त्वरित पाइपलाइन से अनुवाद करना – पर रुपये-की-चीज़ नहीं, गुणवत्ता-नियंत्रण के साथ।

प्रश्न: मॉडलों की भाषा-वार क्षमता के बारे में मूल्यांकन ने क्या बताया?

Eval-Suite से एक और उपयोगी खुलासा: मौजूदा खुले भारतीय और बहुभाषी मॉडलों की तुलनात्मक बेंचमार्किंग में कार्यों और भाषाओं के बीच 'भारी असमानता' दिखी – कोई एक मॉडल सभी भाषाओं-कार्यों में सर्वश्रेष्ठ नहीं। यानी भाषा-कवरेज का दावा और वास्तविक क्षमता अलग-अलग बातें हैं; खरीददारी-निर्णय (सरकारी प्रयोग) के लिए भाषा-वार बेंचमार्क अनिवार्य।