ARTIFICIAL INTELLIGENCE IN INDIAN LANGUAGES AND KNOWLEDGE SYSTEMS: BRIDGING TRADITION AND TECHNOLOGY

Mr. Zunjare Gajanan Uttamrao
Assistant Professor and HOD Dept Of English, Shri Renukadevi Arts Commerce and Science College Mahur Tq Mahur Dist Nanded
gajananzunjare5@gmail.com

Abstract
Artificial Intelligence (AI) is rapidly reshaping how multilingual societies create, preserve, and distribute knowledge. India's linguistic richness—22 scheduled languages, multiple scripts, and thousands of speech varieties—intersects with long-standing intellectual traditions spanning Ayurveda, Yoga, classical arts, mathematics, astronomy, logic (Nyāya), and hermeneutics (Mīmāṃsā). This paper argues that AI, particularly Natural Language Processing (NLP), speech technologies, and multimodal learning, can both democratize access to contemporary knowledge and revitalize the Indian Knowledge System (IKS). We synthesize literature on Indic NLP and cultural heritage digitization; map initiatives that target low-resource settings; and propose a socio-technical framework grounded in openness, community participation, and cultural validity. Methodologically, we combine scoping review, landscape mapping of tools and datasets, and design principles distilled from case vignettes (Sanskrit OCR, Bhasha MT pipelines, Ayurvedic ontology building, and oral-history speech corpora). We discuss evaluation beyond accuracy—covering explainability, inclusivity, ethical safeguards, and community co-governance. The paper concludes with a roadmap for research, policy, and practice: creating open corpora across scripts, funding culturally aware evaluation suites, supporting bilingual education technologies, and embedding IKS epistemologies into AI architectures. Properly stewarded, AI can bridge tradition and technology, fostering linguistic inclusion, knowledge sovereignty, and sustainable innovation.

Keywords: *Artificial Intelligence; Indian Languages; Indian Knowledge System (IKS); NLP; Machine Translation; Speech; Cultural Heritage; Digital Humanities; Ethics.*

Introduction

Context and motivation

India houses one of the most complex linguistic ecologies in the world, with constitutionally recognized languages (e.g., Hindi, Bengali, Marathi, Tamil, Telugu, Urdu, Gujarati, Kannada, Malayalam, Odia, Punjabi, Assamese, Sanskrit, etc.), diverse scripts (Devanagari, Bengali–Assamese, Gurmukhi, Gujarati, Oriya, Perso-Arabic, Kannada, Malayalam, Tamil, Telugu, Roman), and rich code-switching practices. Simultaneously, India's intellectual heritage—the Indian Knowledge System (IKS)—comprises millennia of texts and practices in medicine (Ayurveda, Siddha, Unani), yoga and philosophy, grammar (Pāṇinian linguistics), mathematics and astronomy (Jyotiṣa), statecraft (Arthaśāstra), aesthetics (Rasa theory), and jurisprudence (Dharmaśāstra).

The digital turn has exposed a twofold gap: (a) most AI resources concentrate on high-resource languages; (b) heritage knowledge remains under-digitized, poorly annotated, or disconnected from contemporary educational and research infrastructures. AI, if designed with cultural sensitivity and technical rigor, can bridge both gaps—improving access for learners and researchers while ensuring living traditions are not flattened into mere data.

Problem statement

Mainstream models struggle with script diversity, morphological richness, agglutination, compounding, dialectal variation, and code-mixing common in Indian languages. Low-resource settings lead to brittle systems, domain drift, and bias. For IKS, the challenges include scarcity of machine-readable editions, complex commentarial traditions across centuries, polysemy in technical Sanskrit/Prakrit terminology, and the need for culturally grounded ontologies that avoid anachronistic mappings.

Research questions

1. RQ1: How can AI (NLP, MT, speech, multimodal models) be tailored for Indian languages at scale and in low-resource settings?
2. RQ2: What technical and methodological choices enable faithful, explainable integration of IKS sources into modern digital ecosystems?
3. RQ3: Which evaluation criteria—beyond raw accuracy—capture cultural validity, inclusivity, and ethical stewardship?
4. RQ4: What policy and governance mechanisms ensure equitable participation, open science, and long-term sustainability?

Contribution

This paper offers (i) an integrative review of Indic AI and IKS digitization; (ii) a design framework for culturally aware, low-resource AI; (iii) case-based principles for building datasets, tools, and

pedagogy; and (iv) a roadmap spanning research, education, and public policy.

Literature Review

Indic NLP foundations

Research on tokenization for Brahmic scripts addresses consonant clusters, ligatures, and abugida properties (inherent vowels, matras). Subword segmentation (BPE, unigram LM) must account for orthographic rules like sandhi and compound formation. Morphology-aware language models and character-level representations have shown promise for languages with rich inflection (e.g., Tamil, Marathi). For machine translation (MT), multilingual transformers and transfer learning enable cross-lingual generalization, but domain mismatch, limited parallel corpora, and named-entity fidelity remain issues. Speech technologies (ASR/TTS) face challenges in tonal variation, accent diversity, and background noise typical of field recordings.

Code-mixing and diglossia

Studies on Hindi–English, Bengali–English, and Tamil–English code-mixing show non-trivial syntactic interleaving, with pragmatic switching in social media and classrooms. Normalization must handle romanized text ("Hinglish"), variable spellings, and phonetic transliteration. Approaches include dual-vocabulary LMs, mixed-script tokenization, and contextual language identification.

Digital humanities and heritage

Libraries, archives, and community projects have begun digitizing manuscripts and inscriptions (e.g., palm-leaf manuscripts, copper plates, inscriptional corpora). Optical Character Recognition (OCR) for Devanagari and other Indic scripts has improved but still lags for degraded scans, cursive styles, and historical orthographies. Sanskrit NLP—tokenization, morphological analysis, sandhi splitting, lemmatization—benefits from Pāṇinian grammar formalisms, but ambiguity resolution and cross-commentary alignment remain open problems.

Indian Knowledge System (IKS) informatics

IKS spans multiple epistemic frameworks: pramāṇa theory (means of knowledge), ontology (padārtha categories in Vaiśeṣika), and hermeneutics (Mīmāṃsā). Encoding these within modern ontologies requires careful mapping to avoid reductionism. In biomedicine, Ayurvedic knowledge graphs (dravyaguṇa, rasa-vīrya-vipāka, doṣa-dhātu-mala) are being explored for evidence synthesis and decision support, while respecting traditional diagnostic logic (prakṛti, vikṛti, guṇa, and seasonal routines).

Objectives

1. Map the AI landscape for Indian languages and IKS digitization, identifying technical levers for low-resource settings.
2. Propose a design and evaluation framework that foregrounds inclusivity, cultural validity, and explainability.
3. Illustrate practices through case vignettes in OCR, MT pipelines, ontologies, and speech corpora.
4. Recommend policies for open data, education technology, and cross-sector collaboration.

Methodology

Research design

A qualitative, mixed-methods design combining:

- Scoping review of peer-reviewed and gray literature (toolkits, datasets, project reports).
- Landscape mapping of openly available resources (tokenizers, OCR engines, MT systems, ASR/TTS, transliteration libraries).
- Case vignettes coalescing best practices and pitfalls.
- Framework synthesis drawing on HCI, STS (science and technology studies), and digital humanities.

Data sources

- Academic repositories and conference proceedings in NLP, AI, and digital humanities.
- Government and institutional initiatives on Indian languages and IKS (national translation missions, digital libraries, IKS centers).
- Open-source codebases and corpora for Indic languages.
- Practitioner inputs (when accessible) from linguists, archivists, and traditional knowledge holders.

Analysis approach

We perform thematic coding across three layers:

1. Technical feasibility: data availability, accuracy, robustness, compute constraints.
2. Cultural and educational impact: accessibility, teacher/learner adoption, curricular alignment.
3. Ethical and governance alignment: consent, attribution, community participation, and openness.

Findings: The Current Landscape

Data and corpora

Indic corpora vary widely: some languages have newspapers, Wikipedia dumps, and parallel corpora; others rely on crawl-based monolingual text with heavy noise. Speech corpora are expanding but still sparse for dialects and tribal languages. Romanized code-mixed datasets are often small and domain-specific (e.g., social media). For IKS, digitized texts exist for many

Sanskrit works, yet metadata, TEI-XML markup, and commentary linkages are inconsistent, limiting machine readability.
Implication: Priorities include curated, balanced corpora with licenses that permit research and community reuse; robust metadata; and documentation in multiple Indian languages.

Tools and models

- Tokenization/segmentation: Indic-aware libraries handle script detection, normalization, and transliteration.
- OCR: Engine performance is respectable for clean modern prints but variable for manuscripts and historical fonts; post-OCR correction models and human-in-the-loop interfaces are crucial.
- MT: Multilingual transformer models offer strong baselines; domain adaptation and terminology control (glossaries) improve fidelity in technical/IKS content.

Socio-technical challenges

- Script diversity & orthographic variation increase preprocessing complexity.
- Code-mixing confounds LID (language identification) and decoding.
- Cultural semantics (e.g., polysemous Sanskrit technical terms) demand domain ontologies.

Case Vignettes

Sanskrit OCR for Manuscript Collections

Problem: Manuscripts with uneven inking, ligatures, and archaic orthography defeat general OCR.
Approach: Train OCR with font/handwriting-specific data; incorporate language models for post-OCR correction using sandhi-aware tokenization; apply active learning where human correctors verify uncertain segments flagged by uncertainty sampling.
Outcome: Significant reduction in character error rate; creation of an aligned digital edition with lemma indices and commentary cross-links.
Lesson: Performance hinges on tailored training data, post-OCR pipelines, and user-friendly annotation tools to sustain community participation.

Multilingual MT for Education

Problem: Translating STEM and humanities content from English to Indian languages with domain accuracy.
Approach: Build a terminology-constrained MT pipeline: curated glossaries (e.g., for Ayurveda, grammar, math), domain adaptation via fine-tuning on parallel educational texts, and constrained decoding to enforce glossary terms. Integrate quality estimation to route low-confidence sentences for human review.
Outcome: Higher adequacy and terminology fidelity; faster content localization for open courses.
Lesson: Terminology control and human-in-the-loop workflows outperform generic MT for specialized domains.

Oral Histories and Folk Traditions (ASR/TTS)

Problem: Documenting songs, stories, and local histories in under-resourced dialects.
Approach: Community-led data collection with mobile toolkits, speaker consent protocols, and federated learning to protect privacy. ASR trained with data augmentation (noise, speed, reverb); TTS supports revitalization by enabling pronunciation learning.
Outcome: Usable ASR for community archiving; TTS demos for education and cultural revival.
Lesson: Ethical, participatory methods are as important as model accuracy; projects must return value (training, tools, access) to the communities.

A Design Framework for Culturally Aware, Low-Resource AI

We propose the SAFAR framework—Sources, Alignment, Feedback, Accountability, Reuse.

1. Sources (Data Foundations)

- Balance across scripts, regions, registers (formal, colloquial), and modalities (text, audio, image).
- Document provenance, consent, and licensing; include romanized and native script variants.
- For IKS, include primary texts, commentaries, practitioner notes, and modern scholarship.

2. Alignment (Modeling Choices)

- •Build morphology-aware tokenizers; support mixed-script inputs.
- Use multitask/multilingual transfer; adapt with lo-RA or parameter-efficient fine-tuning to reduce compute.
- For IKS, incorporate symbolic components (grammars, ontologies) alongside neural models (neuro-symbolic).
- Feedback (Human-in-the-Loop)
- Active learning to prioritize uncertain samples for expert or crowd review.
- UI for teachers, archivists, and practitioners to correct outputs; all edits become training signals.
- Provide quality estimation and explanations (attention visualization, rationale templates).
- Accountability (Ethics and Governance)
- Consent & data minimization; community advisory boards; benefit-sharing agreements.

- Bias audits across languages, dialects, and social groups; misuse risk assessments for sensitive content.
- Transparency reports detailing datasets, model limits, and known failure modes.

Integration into Education, Research, and Public Services

Education

- Bilingual classrooms: AI-assisted translation enables lesson plans, glossaries, and formative assessments in mother tongues.
- Accessibility: TTS and ASR support students with visual or writing impairments; captioning assists inclusive classrooms.
- Teacher tooling: Summarization and question generation tailored to curriculum; terminology control to stabilize pedagogy.

Research and scholarship

- Scholarly editions: Automated collation of textual witnesses; alignment of sutras with commentaries; semantic search over IKS corpora.
- Data-driven humanities: Topic modeling and network analysis of historical texts; cross-lingual concordances.
- Open notebooks: Reproducible pipelines (OCR → correction → TEI markup → ontology linkage → public API).

Public services and citizen interfaces

- E-governance: Multilingual chat and form assistance; voice bots for public schemes in local languages.
- Healthcare: Decision support grounded in bilingual patient explanations; careful, non-diagnostic use of Ayurvedic knowledge graphs with practitioner oversight.
- Cultural heritage access: Museum labels, audio guides, and AR overlays in regional languages; crowdsourced annotation drives.

Risks, Ethics, and Governance

Epistemic humility

IKS embodies distinct ontological and methodological premises. Mapping directly to Western taxonomies risks distortion. AI systems must declare scope and limits, avoiding authoritative claims on contested interpretations.

Bias and exclusion

Under-representation of certain dialects or communities can amplify digital inequities. Audits must check for accuracy differentials by language, dialect, gendered speech, and sociolect.

Privacy and consent

Oral histories, medical records, and religious content require granular consent and context-sensitive access controls. Federated learning and differential privacy can reduce data exposure.

Attribution and benefit sharing

Digitization should not lead to extraction. Ethical guidelines must ensure attribution to text owners, archives, and communities, and tangible benefits—training, infrastructure, co-authorship, or revenue sharing for derivative products.

Hallucination and misuse

LLMs may fabricate citations, mistranslate sacred or technical vocabulary, or produce unsafe medical suggestions. Systems must include verification layers, human oversight, and guardrails (e.g., refusal to provide diagnostic advice).

Roadmap and Policy Recommendations

1. National Open Indic Corpus (NOIC): A multi-script, multi-domain corpus with permissive licensing, balanced by language and register; includes code-mixed and romanized data.
2. IKS Digital Stack:

- Texts: High-quality scans, OCR, post-OCR pipelines, TEI-XML editions.
- Ontologies: Community-curated vocabularies for Ayurveda, Yoga, Nyāya, aesthetics, etc.
- APIs: Public endpoints for search, alignment, and pedagogy.

3. Evaluation Fund: Grants to create culturally aware benchmarks and challenge sets (terminology, honorifics, dialect variation, politeness strategies).

Limitations

This synthesis is constrained by uneven documentation across languages and projects, variability in dataset quality, and the pace of AI progress. Some claims rely on general patterns observed across multiple initiatives rather than uniform benchmarks. The case vignettes are illustrative, not exhaustive. Future empirical work should include multi-site field studies, controlled classroom trials, and systematic evaluations with shared test suites.

Conclusion

AI offers a historic opportunity to advance linguistic inclusion and re-center India's intellectual heritage within contemporary knowledge infrastructures. Yet technical capacity alone is insufficient. Sustainable success requires culturally aware design, open and ethical data stewardship, and institutional partnerships that return tangible value to communities and educators. By adopting the SAFAR framework—investing in sources, alignment, feedback, accountability, and reuse—stakeholders can build AI systems that translate across scripts and centuries without

erasing nuance. The resulting ecosystem would enable a student in a Marathi-medium school to learn physics in her mother tongue, a Sanskrit scholar to search cross-commentary debates instantaneously, an Ayurvedic practitioner to explore textual lineages alongside modern evidence, and a museum visitor to hear folk narratives in the voice of the community itself. Properly governed, AI becomes not just a tool for efficiency but a medium for cultural continuity and creative flourishing.

References

1. AI4Bharat. (n.d.). Indic NLP catalog. GitHub. Retrieved August 22, 2025, from https://github.com/AI4Bharat/indicnlp_catalog
2. Anusaaraka. (n.d.). In Wikipedia. Retrieved August 22, 2025, from https://en.wikipedia.org/wiki/Anusaaraka
3. Anuvaad (Document Translation Platform). (n.d.). In Wikipedia. Retrieved August 22, 2025, from https://en.wikipedia.org/wiki/Anuvaad_(Document_Translation_Platform)
4. Bhashini. (n.d.). In Wikipedia. Retrieved August 22, 2025, from https://en.wikipedia.org/wiki/Bhashini
5. GRETIL. (n.d.). In Wikipedia. Retrieved August 22, 2025, from https://en.wikipedia.org/wiki/GRETIL
6. Jain, A., Kunchukuttan, A., & Bhattacharyya, P. (2020). iNLTK: Natural language toolkit for Indic languages. arXiv. https://arxiv.org/abs/2009.12534
7. Joshi, A., Bhattacharyya, P., & others. (2024). A review of the Indic AI research landscape. arXiv. https://arxiv.org/abs/2406.09559
8. Madhavan, A., Singh, S., & others. (2023). SanskritShala: A neural Sanskrit NLP toolkit. arXiv. https://arxiv.org/abs/2302.09527