Lt. Col. Anant Sinha
Administrator,
The Asiatic Society, Kolkata

Article ID No. 2817 *DOI No.* https://doi.org/10.5281/zenodo.21214561

Vidhvanika: Decoding Knowledge

An AI-Powered Initiative for Ancient Manuscript Preservation, Decipherment, and Cultural Heritage Recovery at The Asiatic Society, Kolkata

Lt. Col. Anant Sinha
In collaboration with: IIT Kharagpur
& C-DAC Kolkata

Abstract:

This paper presents Project Vidhvanika — a pioneering interdisciplinary initiative of The Asiatic Society, Kolkata, that harnesses Artificial Intelligence (AI), Large Language Models (LLMs), and Natural Language Processing (NLP) to digitize, transcribe, and decipher a collection of over 51,000 rare manuscripts spanning ancient Indian, Persian, Arabic, and Southeast Asian scripts. Named for the Sanskrit concept of 'decoding knowledge,' Vidhvanika represents a convergence of computational linguistics, digital heritage studies, and institutional archival practice. The paper describes the project's five-phase technical architecture — from multispectral digitization and custom OCR training, through NLP pipeline development and foundation model fine-tuning, to continuous online learning through Reinforcement Learning from Human Feedback (RLHF). We articulate the critical role of Human-in-the-Loop (HITL) validation by domain scholars, the ethical and governance frameworks governing access to sensitive manuscript knowledge, and the institutional partnership model involving IIT Kharagpur and C-DAC Kolkata. The project currently achieves approximately 40% AI transcription accuracy and targets 90–95% accuracy over a 36-month implementation horizon. Beyond its technical contribution, Vidhvanika advances a broader argument: that the preservation of endangered linguistic and cultural heritage requires neither a purely archival nor a purely computational approach, but an integrated methodology that places human scholarly expertise at the centre of the AI pipeline. This paper positions Vidhvanika as a replicable institutional model for AI-driven manuscript decipherment in low-resource, high-cultural-complexity language environments.

Keywords: *Ancient manuscripts, Large Language Models, NLP, digital heritage, OCR, Sanskrit, Brahmi, Kharosthi, transfer learning, Human-in-the-Loop, cultural preservation, Asiatic Society, Vidhvanika*

1. Introduction:

In 1837, James Prinsep — assay master of the Calcutta Mint and a key figure of The Asiatic Society — successfully deciphered the Brahmi script, unlocking the edicts of Emperor Ashoka and, with them, much of India's Mauryan history. That act of decipherment was the work of a single brilliant mind, laboring over fragmentary physical evidence for years. Nearly two centuries later, Project Vidhvanika asks a different question: what would be possible if thousands of such acts of decipherment could be performed simultaneously, continuously, and at scale?

The Asiatic Society, Kolkata, founded in 1784 by Sir William Jones, holds one of the world's most significant collections of ancient manuscripts — over 51,000 texts spanning Sanskrit, Arabic, Persian, Tamil, Bengali, Brahmi, Kharosthi, and several other scripts. Many of these manuscripts have never been fully transcribed. Some scripts within the collection are understood by only a handful of living scholars. As those scholars age, the risk of permanent, irreversible loss of interpretive knowledge grows with each passing year.

Project Vidhvanika — its name derived from the Sanskrit for 'decoding knowledge' — is the institutional response to this crisis. Launched under the leadership of Lt. Col. Anant Sinha, Administrator of the Asiatic Society, in partnership with the Indian Institute of Technology (IIT) Kharagpur and the Centre for Development of Advanced Computing (C-DAC) Kolkata, Vidhvanika deploys a multi-phase AI pipeline to systematically digitize, transcribe, and decipher this irreplaceable collection.

This paper makes three contributions. First, it presents the technical architecture of Vidhvanika — describing the complete pipeline from multispectral imaging through foundation model fine-tuning and continuous online learning. Second, it advances a methodological argument: that AI-driven manuscript decipherment for low-resource, high-complexity ancient languages requires a Human-in-the-Loop design philosophy, not merely as a quality-control mechanism, but as an epistemological commitment to the primacy of scholarly interpretation. Third, it positions Vidhvanika as a replicable institutional model — one that other archives, national libraries, and heritage institutions worldwide could adapt to their own collections.

The remainder of this paper is organized as follows. Section 2 reviews related work in AI-assisted manuscript processing. Section 3 describes the Vidhvanika collection and its scholarly significance. Section 4 presents the technical pipeline across five phases. Section 5 articulates the Human-in-the-Loop framework. Section 6 addresses governance, ethics, and intellectual property. Section 7 discusses continuous learning strategy. Section 8 presents the institutional collaboration model. Section 9 discusses limitations and future directions. Section 10 concludes.

2. Related Works:

The intersection of AI and manuscript studies has generated a rapidly growing body of research over the past decade. This section situates Vidhvanika within three principal strands of prior work: computational approaches to ancient script recognition, NLP systems for low-resource historical languages, and institutional digital heritage initiatives.

2.1 AI-Assisted Ancient Script Recognition

Sommerschield et al. (2019)[1] demonstrated the application of deep learning to ancient Greek inscriptions, achieving significant improvements in restoration accuracy through convolutional neural networks trained on epigraphic datasets. Similarly, Al-Maadeed et al. (2020)[2] investigated historical Arabic manuscript recognition, establishing that transfer learning from modern Arabic models could meaningfully improve performance on historical script variants, though with significant degradation on highly calligraphic forms.

For Indian scripts specifically, Harish, R. (2024)[3] surveyed machine learning approaches to transcribing Sanskrit, Tamil, and Pali texts, highlighting the particular challenges posed by compound characters, damaged physical substrates, and the scarcity of annotated training data. The AI4Bharat initiative (Kakwani D. et al (2020)[4] developed IndicBERT — a transformer language model pre-trained on 12 Indic languages — which has become a key resource for NLP work on South Asian scripts. More recently, Sadhukhan, B., & Punyeshwarananda, S. (2025)[5] demonstrated the effectiveness of transfer learning from OpenAI's Whisper model for Sanskrit automatic speech recognition, underscoring the power of domain adaptive pre-training even in low-resource conditions.

The use of byte-level models — particularly Google's ByT5 — has emerged as a significant development for scripts lacking standardized digital tokenization. As documented in recent work on historical manuscript processing [Historica Research Group. (2025)][6], ByT5's architecture, which operates directly on raw byte sequences without requiring a pre-built vocabulary, makes it uniquely suited to scripts like Brahmi and Kharosthi, which lack comprehensive Unicode coverage and standardized digitization conventions.

[1] Sommerschield, T., Assael, Y., Shillingford, B., Bordbar, M., Rayson, J., Sherrat, M., & de Freitas, N. (2019). Predicting the past with Ithaca: Restoring and attributing ancient texts using deep learning. Nature, 603, 280–283.

[2] Al-Maadeed, S., Hassaine, A., & Bouridane, A. (2020). Transfer learning for historical Arabic manuscript recognition. Pattern Recognition Letters, 129, 155–162.

[3] Harish, R. (2024). AI-based OCR for digitizing ancient Indian texts: Preserving linguistic heritage and overcoming script challenges. ResearchGate preprint.

[4] Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N. C., Bhatt, A., Khapra, M. M., & Kumar, P. (2020). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. Findings of EMNLP 2020.

[5] Sadhukhan, B., & Punyeshwarananda, S. (2025). Automatic speech recognition for Sanskrit with transfer learning. arXiv preprint arXiv:2501.10024.

[6] Historica Research Group. (2025). AI in historical research: 2025 insights and trends. historica.org.

2.2 NLP for Low-Resource Historical Languages:

Sanskrit NLP has developed a substantial specialized toolkit. Krishna et al. (2016)[7] developed graph-based approaches to Sanskrit word segmentation, addressing the particularly complex sandhi phenomenon — wherein words fuse at their boundaries through phonological transformation, rendering standard word-boundary detection approaches ineffective. Hellwig and Nehrdich (2018)[8] extended this work through character-CNN/LSTM architectures for neural segmentation, demonstrating improvements over earlier rule-based systems. The Sanskrit Heritage Platform (Goyal, P -2012)[9], developed at INRIA, provides a distributed computational infrastructure for lexical and syntactic analysis of Sanskrit texts.

For Arabic and Persian historical manuscripts, the challenges differ but are equally formidable. Ligature-heavy calligraphic styles, extensive diacritical variation, and the absence of standardized orthography in pre-modern texts create significant barriers to both OCR and linguistic analysis. Alginahi et al. (2013)[10] provide a comprehensive survey of Arabic historical document processing, concluding that hybrid architectures combining deep learning with rule-based linguistic knowledge consistently outperform purely neural approaches on historical material.

Critically, the field has converged on a key insight: for morphologically rich, low-resource historical languages, task-specific fine-tuning of large pre-trained multilingual models consistently outperforms both from-scratch training and rule-based approaches, provided that domain-adaptive pre-training is conducted on in-domain text prior to task-specific tuning (Gururangan, S. -2020)[11].

2.3 Institutional Digital Heritage Initiatives

Large-scale institutional manuscript digitization has been undertaken at several prominent archives globally, establishing best practices for metadata standards, digital preservation formats, and community engagement. Open access digital manuscript platforms have demonstrated the transformative impact of public accessibility on scholarly engagement with heritage collections.

---

[7] Krishna, A., Sangroya, A., Bandaru, N., & Bhatt, A. (2016). Compound Sanskrit segmentation using neural models. Proceedings of ICON 2016.

[8] Hellwig, O., & Nehrdich, S. (2018). Sanskrit word segmentation using character-level recurrent and convolutional neural networks. Proceedings of EMNLP 2018, 2754–2763.

[9] Goyal, P., Huet, G., Kulkarni, A., Scharf, P., & Bunker, R. (2012). A distributed platform for Sanskrit processing. Proceedings of COLING 2012.

[10] Alginahi, Y., Kabir, M. N., & Tayan, O. (2013). An enhanced Otsu binarization technique for Arabic manuscript images. Proceedings of ICEDEG 2013, 1–7.

[11] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. Proceedings of ACL 2020, 8342–8360.

Within India, the National Mission for Manuscripts (NMM), operating under the Ministry of Culture, has digitized a significant portion of India's estimated 10 million manuscripts, though computational analysis and AI-assisted transcription have remained limited in scope [National Mission for Manuscripts. (2023)][12]. The Digital Corpus of Sanskrit (DCS), developed by Oliver Hellwig, provides an annotated and lemmatized corpus of Sanskrit texts that has served as a key training resource for Sanskrit NLP work [Hellwig, O. (2010][13].

Vidhvanika builds directly on this body of work, distinguishing itself through three features: its institutional anchoring in one of Asia's oldest scholarly societies; its integrated pipeline from physical preservation through active AI-assisted decipherment; and its explicit commitment to continuous online learning as a core architectural principle rather than a post-hoc add-on.

3. The Vidhvanika Collection: Scope and Scholarly Significance:

The manuscript collection of The Asiatic Society, Kolkata, is among the most diverse and historically significant in South Asia. Assembled over more than two centuries of institutional scholarship, the collection encompasses texts spanning approximately 1,500 years of intellectual production across the Indian subcontinent and adjacent regions.

3.1 Collection Composition:

The 51,000+ manuscripts in the Vidhvanika project span seven principal script traditions, each presenting distinct computational and linguistic challenges:

Table 1: Composition of the Vidhvanika manuscript collection by script tradition.

Script / Language	Subject Areas	Key Computational Challenge
Sanskrit (Devanagari)	Philosophy, astronomy, medicine, mathematics, literature	Sandhi rules, compound word formation (samasas)
Arabic	Islamic philosophy, mathematics, medicine, history	Right-to-left script, calligraphic variation, diacritical complexity
Persian	Literature, history, Sufi philosophy, administrative texts	Nastaliq script style, ligature density
Bengali	Literature, Vaishnavism, folk traditions, history	Historical orthographic variation from modern Bengali
Tamil	Literature, Shaiva philosophy, grammar, medicine	Palm-leaf substrate, ancient Tamil Brahmi variants

[12] National Mission for Manuscripts. (2023). Annual report 2022–23. Ministry of Culture, Government of India.
[13] Hellwig, O. (2010–present). Digital corpus of Sanskrit. University of Zurich / Heinrich-Heine-Universität Düsseldorf.

Brahmi	Edicts, inscriptions, early Buddhist texts	Script largely undeciphered without scholarly intermediary; no Unicode standard
Kharosthi & Others	Buddhist texts, Central Asian manuscripts	Extremely low-resource; limited scholarly expertise available

Beyond their linguistic diversity, the manuscripts span an extraordinary range of subject domains — from treatises on astronomy (Jyotisha), mathematics (Ganita), and Ayurvedic medicine (Vaidyaka), to philosophical texts in the Nyaya, Mimamsa, and Vedanta traditions, to literary works in Sanskrit kavya poetry, Persian, and classical Tamil Sangam literature. This subject diversity creates additional challenges for computational processing: domain-specific vocabulary, technical terminology, and authorial conventions vary enormously across the collection.

3.2 Physical Condition and Preservation Risk:

A significant proportion of the collection is in fragile physical condition. Palm-leaf manuscripts are particularly vulnerable to humidity, insect damage, and handling. Paper manuscripts from the Mughal period show extensive foxing, water damage, and ink oxidation. Several hundred manuscripts are assessed as being at risk of complete physical disintegration within the next decade.

This fragility creates a temporal urgency that shapes Vidhvanika's prioritization strategy. The project employs a triage system — informed by conservators, script specialists, and AI-assisted condition assessment — that prioritizes the most endangered manuscripts for early digitization and processing, rather than proceeding in collection order.

4. Technical Architecture: The Five-Phase Pipeline:

The Vidhvanika technical pipeline is organized across five sequential but mutually reinforcing phases. Each phase produces outputs that feed into subsequent phases, while also generating feedback signals that improve earlier phases retrospectively.

4.1 Phase 1: Multispectral Digitization and Image Processing

Standard photography of aged manuscripts frequently fails to capture faded, oxidized, or physically damaged text. Vidhvanika employs multispectral imaging — capturing manuscript surfaces across ultraviolet, visible, and infrared wavelengths — to reveal ink layers invisible under

normal illumination. This approach has been demonstrated to recover legible text from manuscripts that appeared entirely blank under visible light (Chabrier, A., & Brun, L. (2019)[14].

All manuscripts are scanned at a minimum of 400 DPI (600 DPI for the most fragile items) and stored as archival TIFF files. Working copies are generated in PNG format for computational processing. All digital assets are structured according to the IIIF (International Image Interoperability Framework) standard, enabling interoperability with major international digital library systems.

4.2 Phase 2: OCR Development for Ancient Scripts:

Optical Character Recognition for ancient Indian scripts represents one of the most technically demanding components of the Vidhvanika pipeline. Standard commercial OCR engines — including Google Vision API and Microsoft Azure OCR — perform adequately on modern printed Devanagari but degrade severely on historical script variants, handwritten manuscripts, and non-Devanagari scripts.

Vidhvanika employs a three-step OCR development approach. In the first step, existing engines provide baseline outputs that, while imperfect, offer training signal for human correctors. In the second step, specialized historical manuscript OCR tools — Kraken and Calamari — are trained on synthetically augmented manuscript images, with distortions (blur, noise, aging effects, ink bleeding) applied to clean digitized texts to simulate real manuscript degradation patterns. In the third step, a language model-based post-correction layer performs automated error correction, with a script-specific spell-checker and grammar validator flagging low-confidence outputs for human review.

This pipeline architecture closely follows the approach documented by Harish, R. (2024).[3] for Sanskrit manuscript OCR, and extends it by incorporating the multispectral image preprocessing stage and the systematic language model post-correction layer, which have been shown to improve OCR accuracy by 8–15 percentage points on degraded historical manuscripts.

4.3 Phase 3: NLP Pipeline Development

The NLP pipeline converts raw OCR output into linguistically annotated, searchable, and analyzable text. For the Vidhvanika collection, this requires building a sequence of eight interlocking NLP components, each of which must be developed and validated independently before integration into the full pipeline.

[14] Chabrier, A., & Brun, L. (2019). Multispectral imaging of degraded manuscripts: Methodological perspectives. Journal of Cultural Heritage, 36, 194–202.

Table 2: NLP Pipeline Components, their Functions, and Development Priorities

Component	Function	Primary Challenge for Ancient Scripts	Development Priority
Word Segmentation	Split compound words and sandhi forms into base tokens	Sanskrit sandhi and samasa compounds create arbitrary-length fused forms	Critical
Morphological Tagging	Assign grammatical roles: case, gender, tense, number, person	Sanskrit has 8 cases, 3 genders, 3 numbers — 48+ morphological categories	Critical
Lemmatization	Map inflected forms to dictionary headwords	Irregular root forms; many Sanskrit roots absent from modern dictionaries	High
Dependency Parsing	Map syntactic relationships between sentence elements	Free word order in Sanskrit; strict V-final patterns in classical Tamil	High
Named Entity Recognition	Identify persons, places, deities, dynasties, dates	Historical entities absent from modern NER training data	High
Topic Classification	Assign manuscript to subject domain	Domain vocabulary overlaps across philosophy, science, literature	Medium
Machine Translation	Generate draft translations into modern target languages	Extreme lexical distance between ancient and modern language forms	Medium
Question Answering	Enable natural language querying of manuscript contents	Requires all preceding components to function reliably	Future

A particularly challenging aspect of Sanskrit NLP is the sandhi phenomenon — phonological rules that merge words at their boundaries, transforming final and initial phonemes to create fused forms that bear no obvious relationship to their constituent words. Vidhvanika employs the approach developed by Krishna et al. (2026)[15], adapted and extended with additional training data from the Digital Corpus of Sanskrit, for the word segmentation component.

4.4 Phase 4: Foundation Model Selection and Fine-Tuning:

The choice of foundation model is the most consequential architectural decision in the Vidhvanika pipeline. A foundation model — a large AI system pre-trained on vast quantities of text — provides the linguistic and world-knowledge base upon which all downstream task-specific models are built. The key insight motivating the use of foundation models for ancient scripts is that the structural features of language (syntactic relationships, semantic compositionality, discourse coherence) are

[15] Ibid; See Footnote-7.

sufficiently universal that a model trained on modern languages retains useful representations even when fine-tuned on ancient text [Gururangan, S. at al. (2020)][16].

After systematic evaluation, Vidhvanika adopts a two-tier foundation model strategy. Google's ByT5 serves as the primary foundation model for scripts lacking standardized digital tokenization (Brahmi, Kharosthi, and highly variant manuscript forms of Sanskrit and Tamil). ByT5's byte-level architecture — processing raw byte sequences without a pre-defined vocabulary — makes it uniquely suited to scripts with incomplete Unicode coverage and non-standard character combinations [Xue, L at al. (2022)][17]. For scripts with sufficient modern digital data (Arabic, Persian, Bengali), Meta's XLM-RoBERTa and AI4Bharat's IndicBERT serve as foundation models, leveraging their pre-training on contemporary Indic and Semitic language text.

Fine-tuning proceeds in two stages. In the first stage — Domain Adaptive Pre-Training (DAPT) — the foundation model is trained on the full Vidhvanika corpus of digitized manuscript text in an unsupervised masked language modelling objective. This stage, which requires several weeks of GPU compute on IIT Kharagpur's HPC infrastructure and C-DAC Kolkata's Param supercomputers, produces a model that has internalized the vocabulary distributions, n-gram patterns, and general linguistic structures of ancient manuscript text without requiring manual annotation. In the second stage — task-specific fine-tuning — the DAPT model is trained on smaller, expert-annotated datasets for each NLP task, yielding specialized models for word segmentation, morphological tagging, translation, and named entity recognition.

To minimize computational cost and prevent catastrophic forgetting — the well-documented phenomenon in which neural networks lose previously acquired knowledge when trained on new tasks [Kirkpatrick, J.,at al. (2027)][18]— Vidhvanika employs Low-Rank Adaptation (LoRA) [Hu, E., at al. (2022)] [19] for all fine-tuning. LoRA updates only a small fraction (typically 0.1–1%) of model parameters, preserving the foundation model's general capabilities while enabling script-specific specialization. Separate LoRA adapters are trained for each script family, allowing the system to serve multiple scripts without weight interference.

[16] Ibid; See Footnote-11

[17] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2022). ByT5: Towards a token-free future with pre-trained byte-to-byte models. Transactions of the ACL, 10, 291–306.

[18] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13), 3521–3526.

[19] Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. ICLR 2022.

5. Human-in-the-Loop Validation: An Epistemological Commitment:

The framing of Human-in-the-Loop (HITL) validation as merely a quality-control mechanism — a safety net for AI errors — fundamentally misrepresents its role in the Vidhvanika architecture. We argue instead that HITL validation in ancient manuscript processing constitutes an epistemological commitment: a recognition that the meaning of ancient texts is not a fixed property to be 'extracted' by a sufficiently powerful algorithm, but an interpretive achievement requiring the active exercise of scholarly judgment.

Ancient manuscripts routinely present interpretive ambiguities that no algorithm can resolve through pattern matching alone. A damaged character might be consistent with two or more grapheme options, each of which generates different semantic readings of the surrounding text. A technical term from Ayurvedic medicine might overlap in form with a common word in a different domain. A scribal abbreviation might be opaque without knowledge of the specific scriptorial tradition in which the manuscript was produced. In all such cases, the appropriate response is not to select the statistically most probable reading, but to flag the ambiguity for expert human adjudication.

5.1 Scholar Annotation Platform

Vidhvanika has designed a web-based scholar annotation platform that presents AI-generated transcriptions alongside high-resolution manuscript scans for side-by-side comparison. The platform supports multi-script display — including right-to-left rendering for Arabic and Persian — and provides confidence scores for every character and word in the AI output, enabling scholars to focus their attention efficiently on the most uncertain regions of each text.

All scholar corrections are logged with full attribution — name, institutional affiliation, timestamp — and aggregated into a structured feedback database. This database serves two functions: it provides the curated, ground-truth training data for subsequent rounds of model fine-tuning through the RLHF process described in Section 7; and it creates a provenance record that enables future scholars to understand the basis for specific textual decisions and to revisit those decisions as scholarly consensus evolves.

5.2 Expert Network Architecture

The Vidhvanika expert network is organized across six specialist domains, reflecting the breadth of the manuscript collection. Sanskrit scholars cover the majority of the Brahmic script materials. Arabic and Persian experts cover the Islamic manuscript tradition. Tamil specialists handle the Tamil and Tamil Brahmi materials. Historians provide contextual validation across all script traditions. AI linguists from IIT Kharagpur serve as the bridge between computational outputs and

domain scholarship. All scholars operate under formal MoUs with the Asiatic Society that specify quality standards, attribution requirements, and intellectual property terms.

5.3 Quality Assurance Protocol

Every AI-generated transcription is required to receive validation from at least two independent scholars before being classified as verified. Where the two scholars disagree, the disputed passage is escalated to a senior editorial board for adjudication. All AI-generated restorations of damaged or illegible text — conjectural fills inserted by the model to complete fragmentary passages — are permanently marked with a distinct typographic indicator, ensuring that future readers can distinguish AI inference from documented text. Accuracy metrics are tracked per script type, per time period, and per subject domain, with monthly internal reports tracking progress toward the 90–95% target.

6. Governance, Ethics, and Intellectual Property:

The governance framework for Vidhvanika addresses three distinct but interrelated ethical domains: access governance, AI transparency, and intellectual property.

6.1Tiered Access Framework:

Not all manuscripts in the Asiatic Society collection are appropriate for unrestricted public access. Sacred texts, ritually restricted knowledge, and manuscripts of significant community sensitivity to living religious and cultural communities require careful governance of access and dissemination. Vidhvanika implements a four-tier access framework:

- Tier 1 — Open Access: Historical, scientific, literary, and astronomical manuscripts without community sensitivity are made freely available through a public digital library portal.
- Tier 2 — Restricted Academic: Philosophically or ritually sensitive texts are accessible to verified academic researchers upon institutional registration.
- Tier 3 — Controlled Access: Sacred texts and manuscripts with direct relevance to living community traditions are accessible only to designated community representatives and senior scholars, via a formal application and review process.
- Tier 4 — Archival Only: Severely damaged, partially deciphered, or manuscripts with contested provenance remain restricted to internal Asiatic Society scholars pending full scholarly verification.

6.2 AI Transparency Standards:

All content produced by the Vidhvanika AI pipeline carries explicit labelling that distinguishes its epistemic status. AI-generated transcriptions are labelled 'AI Generated — Pending Scholar Verification.' Scholar-validated transcriptions carry the validating scholar's name and institutional

affiliation. AI-generated restorations of damaged text are permanently marked as conjectural with a standardized typographic indicator. Machine-translated passages carry numerical confidence scores. This transparency regime is designed not only to prevent misattribution, but to maintain the integrity of the scholarly record — ensuring that future researchers understand precisely which portions of a digitized manuscript represent documented text and which represent computational inference.

6.3 Intellectual Property:

The intellectual property framework for Vidhvanika reflects the multi-stakeholder nature of the project. Digitized manuscripts remain the IP of The Asiatic Society, Kolkata. AI models trained on the manuscript corpus are owned jointly by the Asiatic Society, IIT Kharagpur, and C-DAC Kolkata under formal MoU agreements. Community and indigenous knowledge manuscripts carry IP provisions that share rights with relevant community representatives. Open-source model weights may be published under Creative Commons Attribution — Non-Commercial license; commercial applications require a formal licensing agreement.

7. Continuous Online Learning Strategy:

A defining architectural feature of Vidhvanika is its commitment to continuous online learning — a design philosophy in which the AI system is treated not as a finished product to be deployed, but as an ongoing learner whose capabilities improve incrementally as it encounters more data, more scholar feedback, and more diverse manuscript materials.

7.1 Continual Pre-Training:

As new manuscripts are digitized and processed, their text is added to the training corpus and used to update the foundation model through periodic continual pre-training cycles. This approach, which adds new domain knowledge without replacing existing knowledge, has been shown to consistently improve downstream task performance across the board, even for tasks unrelated to the specific manuscripts added in each training cycle [Ke, Z. at al. (2023)][20]. Vidhvanika schedules continual pre-training cycles on a quarterly basis, with the updated model deployed to the production system following evaluation against a held-out benchmark set.

7.2 Reinforcement Learning from Human Feedback:

Scholar corrections — the primary output of the HITL validation process — are systematically converted into training signal through a Direct Preference Optimization (DPO) process [Wang, B. at

[20] Ke, Z., Wang, B., Xu, H., Shu, L., & Liu, B. (2023). Continual pre-training of language models. ICLR 2023.

al. (2023)][21]. DPO, which has emerged as a stable and computationally efficient alternative to traditional RLHF for language model alignment, trains the model to prefer scholar-validated outputs over its own previous outputs, directly addressing systematic error patterns identified through the annotation process. The aggregation of scholar corrections across the full expert network provides a large-scale, high-quality signal that would be unavailable to any single researcher working in isolation.

7.3 Preventing Catastrophic Forgetting

The risk of catastrophic forgetting — wherein new training causes a model to lose previously acquired capabilities — is a fundamental challenge in continual learning [Kirkpatrick, J. at al (2027)][22]. Vidhvanika addresses this through three complementary mechanisms. First, rehearsal replay ensures that each training cycle includes a random sample of previously processed manuscript text, maintaining the model's familiarity with earlier material. Second, Elastic Weight Consolidation (EWC) identifies the model parameters most critical to previously learned tasks and mathematically protects them from large updates during new training. Third, the LoRA adapter architecture ensures that script-specific knowledge is modularized in separate adapter weights, so that new script learning does not overwrite parameters encoding knowledge of previously mastered scripts.

7.4 Accuracy Trajectory

The combined effect of continual pre-training, RLHF fine-tuning, and expanding annotated datasets is an expected monotonic improvement in transcription accuracy over the 36-month implementation period. The current baseline of approximately 40% accuracy — achieved with an initial model trained on limited annotated data — is expected to reach 55–60% at six months following corpus expansion and custom tokenizer deployment; 70–75% at twelve months following ByT5 fine-tuning and NLP pipeline integration; 80–85% at eighteen months following HITL platform deployment and RLHF activation; and 90–95% at thirty-six months following full pipeline maturation and global scholar network activation.

[21] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. NeurIPS 2023.

[22] Ibid; See Footnote- 16

Table 3: Vidhvanika Accuracy Improvement Milestones Across the 36-Month Implementation Horizon

Milestone	Target Accuracy	Key Technical Enablers	Timeline
Baseline	~40%	Initial model; basic OCR pipeline operational	Current
Milestone 1	55–60%	5,000 priority manuscripts digitized; custom tokenizers; ByT5 DAPT initiated	Month 6
Milestone 2	65–70%	ByT5 fine-tuned on Sanskrit and Arabic; full NLP pipeline for top-priority scripts	Month 12
Milestone 3	75–80%	HITL platform with 50+ scholars; RLHF / DPO feedback loop operational	Month 18
Milestone 4	85–88%	Continual pre-training active; LoRA adapters for all 6 script families	Month 24
Target	90–95%	Full pipeline mature; global scholar network; open API launched	Month 36

8. Institutional Collaboration and International Engagement:

Vidhvanika's institutional architecture is designed to be simultaneously locally grounded and globally connected. The project's two primary technology partners — IIT Kharagpur and C-DAC Kolkata — provide complementary capabilities that reflect the dual requirements of cutting-edge AI research and large-scale computational infrastructure.

IIT Kharagpur contributes AI linguistics research capacity, access to high-performance computing infrastructure, and a pipeline of AI engineers trained in NLP and low-resource language processing. C-DAC Kolkata — India's premier national technology research institution — provides supercomputing resources through its Param infrastructure, experience in large-scale government digital heritage projects, and institutional connectivity to the national digital infrastructure initiatives of the Ministry of Electronics and Information Technology (MeitY).

Internationally, Vidhvanika is developing collaboration frameworks with several leading global institutions holding significant South Asian and Islamic manuscript collections. These collaborations provide access to parallel manuscript collections that expand the training data available for specific script models, opportunities for joint scholar network development, and platforms for mutual knowledge exchange on digital heritage practice.

At the international institutional level, Vidhvanika is pursuing formal registration of the Asiatic Society manuscript collection with UNESCO's Memory of the World Programme — an initiative that would confer international recognition and protection on the collection, unlock UNESCO

digitization funding channels, and position Vidhvanika as a globally recognized model for AI-driven heritage preservation in the Global South.

9. Current Limitations:

9.1 Current Limitations:

Several significant limitations of the current Vidhvanika system require candid acknowledgment. First, the 40% baseline accuracy reflects the genuine difficulty of the task: ancient manuscript transcription, even for human experts, is a slow, error-prone, and often genuinely uncertain process. Setting 90–95% as a target — rather than a guarantee — reflects this reality. For the most damaged and most rare-script materials, expert-level human accuracy itself may not significantly exceed current AI performance.

Second, the size and diversity of the collection create an inherent tension between depth and breadth. A model that performs well on the most common Sanskrit materials may perform poorly on the smaller subcollections of Prakrit manuscripts, where training data is severely limited and expert validators are few. The LoRA adapter architecture mitigates but does not eliminate this tension.

Third, the HITL framework, while epistemologically sound, creates bottlenecks that are fundamentally limited by scholar availability. The global community of scholars capable of validating ancient Brahmi or Kharosthi transcriptions is small, geographically concentrated, and aging. Vidhvanika's expert network development is therefore simultaneously a technical and an institutional challenge — one that requires active engagement with university linguistics departments, Indology programs, and national funding bodies to sustain over a multi-decade horizon.

9.2 Future Directions:

Several research directions emerge naturally from the Vidhvanika project. First, the development of multimodal models that process manuscript images directly — rather than relying on OCR as an intermediate step — represents a significant potential improvement pathway. Recent work on vision-language models suggests that end-to-end image-to-text approaches may eventually outperform OCR-based pipelines for highly irregular manuscript scripts [Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024).][23].

Second, the integration of audio data from Vedic chanting traditions and oral manuscript recitation creates opportunities for speech-assisted transcription — leveraging the living oral tradition as a

[23] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. NeurIPS 2024

complementary signal to the written text. The Whisper-based Sanskrit ASR work of Sadhukhan and Punyeshwarananda [Sadhukhan, B., & Punyeshwarananda, S. (2025)] [24] provides a directly relevant methodological foundation for this direction.

Third, the Vidhvanika knowledge base — once sufficiently populated with validated, annotated manuscript content — creates the foundation for a specialist question-answering system: a tool that would allow scholars to pose natural language queries across the full collection, identifying relevant manuscripts, tracing conceptual threads across texts and traditions, and generating structured comparative analyses. Such a system would represent a qualitative transformation in the accessibility of the collection to scholars worldwide.

10. Conclusion:

Project Vidhvanika represents a significant experiment in what might be called institutional AI for cultural heritage — the systematic deployment of state-of-the-art artificial intelligence not in the service of commercial optimization, but in the service of civilisational memory. Its ambition is to ensure that the 51,000 manuscripts held in The Asiatic Society's vaults — accumulated over two and a half centuries of scholarly collecting — become, for the first time, fully accessible to the global scholarly community and to the communities whose ancestors produced them.

The technical contributions of this paper — the five-phase pipeline architecture, the two-tier foundation model strategy, the LoRA adapter approach to multi-script learning, the DPO-based RLHF framework, and the tiered governance model — are offered not only as descriptions of the Vidhvanika system, but as design principles that other heritage institutions might adapt to their own contexts. The challenges Vidhvanika faces — low-resource scripts, fragile physical materials, limited annotator availability, complex intellectual property environments — are shared by archives and libraries across the developing world.

The broader argument of this paper is that the preservation of endangered linguistic and cultural heritage in the twenty-first century is neither a purely archival task nor a purely computational one. It requires the integration of institutional commitment, technological capability, and — most essentially — the irreplaceable interpretive expertise of human scholars. Vidhvanika's Human-in-the-Loop architecture is not a concession to AI's current limitations, but an expression of a deeper conviction: that knowledge, ultimately, is a human achievement, and that the tools we build to preserve it must honour that fact.

[24] Ibid; See Footnote- 5

Two centuries after James Prinsep deciphered Brahmi by hand, Vidhvanika invites artificial intelligence into the continuing work of decipherment — not to replace the Prinseps of the future, but to multiply them.

Acknowledgement:

The author gratefully acknowledges the scholars, conservators, and technical staff of The Asiatic Society, Kolkata, whose expertise and dedication make Project Vidhvanika possible. The author thanks the research teams at IIT Kharagpur and C-DAC Kolkata for their technical partnership and institutional support. The contributions of the expert validation network of domain scholars spanning Sanskrit, Arabic, Persian, Tamil, and historical studies are fundamental to the scholarly integrity of the project. This work is supported by Prof Samik Bhattacharya(IIT-K),Niladri Saha & Koushik Maity (CDAC-K) and I also acknowledge the use of AI-based tools during the preparation of this manuscript, primarily for structuring and language refinement. All research design, analysis, interpretations, and conclusions presented are still in progress.

References:

Alginahi, Y., Kabir, M. N., & Tayan, O. (2013). An enhanced Otsu binarization technique for Arabic manuscript images. Proceedings of ICEDEG 2013, 1–7.

Al-Maadeed, S., Hassaine, A., & Bouridane, A. (2020). Transfer learning for historical Arabic manuscript recognition. Pattern Recognition Letters, 129, 155–162.

Bibliothèque nationale de France. (2024). Gallica platform: Access, usage, and scholarly impact report. BnF Digital Services.

British Library. (2023). Endangered archives programme: Principles and practice. British Library Digital Scholarship.

Chabrier, A., & Brun, L. (2019). Multispectral imaging of degraded manuscripts: Methodological perspectives. Journal of Cultural Heritage, 36, 194–202.

Goyal, P., Huet, G., Kulkarni, A., Scharf, P., & Bunker, R. (2012). A distributed platform for Sanskrit processing. Proceedings of COLING 2012.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. Proceedings of ACL 2020, 8342–8360.

Harish, R. (2024). AI-based OCR for digitizing ancient Indian texts: Preserving linguistic heritage and overcoming script challenges. ResearchGate preprint.

Hellwig, O. (2010–present). Digital corpus of Sanskrit. University of Zurich / Heinrich-Heine-Universität Düsseldorf.

Hellwig, O., & Nehrdich, S. (2018). Sanskrit word segmentation using character-level recurrent and convolutional neural networks. Proceedings of EMNLP 2018, 2754–2763.

Historica Research Group. (2025). AI in historical research: 2025 insights and trends. historica.org.

Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. ICLR 2022.

Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N. C., Bhatt, A., Khapra, M. M., & Kumar, P. (2020). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. Findings of EMNLP 2020.

Ke, Z., Wang, B., Xu, H., Shu, L., & Liu, B. (2023). Continual pre-training of language models. ICLR 2023.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13), 3521–3526.

Krishna, A., Sangroya, A., Bandaru, N., & Bhatt, A. (2016). Compound Sanskrit segmentation using neural models. Proceedings of ICON 2016.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. NeurIPS 2024.

National Mission for Manuscripts. (2023). Annual report 2022–23. Ministry of Culture, Government of India.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. NeurIPS 2023.

Sadhukhan, B., & Punyeshwarananda, S. (2025). Automatic speech recognition for Sanskrit with transfer learning. arXiv preprint arXiv:2501.10024.

Sommerschield, T., Assael, Y., Shillingford, B., Bordbar, M., Rayson, J., Sherrat, M., & de Freitas, N. (2019). Predicting the past with Ithaca: Restoring and attributing ancient texts using deep learning. Nature, 603, 280–283.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2022). ByT5: Towards a token-free future with pre-trained byte-to-byte models. Transactions of the ACL, 10, 291–306.