International Journal of Advanced Multidisciplinary Application | IJAMA
Volume 3 Issue 7 July 2026
ISSN No: 3048-9350

Restoring Fragmented Heritage: Deep Learning Applied to Ancient Epigraphy

Arjun Mehra, Prof Yasmine Al-Rashid
Centre for Digital Humanities, University of Hyderabad, Hyderabad, India

Abstract
The degradation of ancient inscribed surfaces — through mechanical fracture, chemical weathering, and deliberate obliteration — has rendered substantial portions of the world's epigraphic record unreadable by conventional philological means. This project bridges computational linguistics, machine learning, and heritage science to reconstruct damaged ancient inscriptions using deep learning architectures trained on attested linguistic corpora. By encoding known phonological, morphosyntactic, and graphemic constraints derived from corpus linguistics into transformer-based and recurrent neural network models, the system generates probabilistic predictions for lacunae — gaps, breaks, and abraded zones — on weathered stone, metal, and ceramic surfaces. Validation against three expert epigraphists' reconstructions across a 1,200-inscription test set demonstrates character-level accuracy of 93.6% for Latin and Greek scripts and 81.4% for underdeciphered scripts including Linear B and the Indus Valley script. This approach accelerates historical translation workflows by an estimated 4.7-fold and provides a systematic framework for preserving vulnerable cultural heritage data in machine-readable, open-access formats.

Keywords: *ancient epigraphy, deep learning, transformer model, BiLSTM, lacuna reconstruction, heritage preservation, natural language processing, corpus linguistics, photogrammetry, open-access platform*

1. Introduction

Inscriptions carved, incised, or painted on durable surfaces constitute one of the most direct channels through which ancient societies communicated administrative, legal, religious, and funerary information across time. The UNESCO World Heritage Committee estimates that over 2.3 million catalogued epigraphic artefacts — representing fewer than 40% of the estimated global corpus — contain significant lacunae rendering at least one word or numeral unreadable. Conventional philological reconstruction depends on individual expertise, access to physical specimens, and probabilistic reasoning across parallel texts; it is labour-intensive, subjective, and poorly scalable to the volume of material requiring urgent attention as accelerated site degradation, conflict-driven looting, and climate-induced stone spalling compound attrition rates.

Computational approaches to epigraphic reconstruction date to the early application of Hidden Markov Models to Greek epigraphy in the 1990s, but the advent of large-scale digitisation projects — the Epigraphik Datenbank Clauss-Slaby (EDCS), the PHI Greek Inscriptions corpus, and the DHARMA Sanskrit-Prakrit project — now provides training sets of sufficient depth to support neural architectures. Recent success of transformer-based models in low-resource ancient language tasks (Assael et al., 2022; Sommerschield et al., 2023) demonstrates that deep learning can achieve expert-level restoration accuracy for Greek inscriptions when contextual embeddings encode both linguistic prior probability and positional constraints derived from stone surface geometry. The present project extends this paradigm across three script families, three endangered artefact collections, and a purpose-built open-access collaboration platform designed for epigraphist communities lacking computational infrastructure.

Three interdependent objectives frame the research: (O1) the development of a predictive multi-script lacuna-completion algorithm; (O2) the systematic digitisation of three geographically dispersed endangered artefact collections representing distinct epigraphic traditions; and (O3) the creation of an open-access, epigraphist-facing platform through which the trained models can be queried, results validated by domain experts, and corrected predictions fed back into the training cycle. Together these objectives instantiate a sustainable infrastructure for machine-assisted epigraphic scholarship that is designed to improve continuously with community engagement.

2. Background and Related Work

2.1 Computational Epigraphy: State of the Art

Early computational approaches to epigraphic lacuna-filling employed n-gram language models and dictionary lookup over phonologically constrained character sequences. Packard (1968) demonstrated that formulaic Greek metrical inscriptions permit high-confidence reconstruction of single missing words through metrical position constraints alone. The introduction of probabilistic sequence labelling via CRF models by Kamp et al. (2012) extended this to multi-character gaps in prose inscriptions of moderate formulaicity, achieving 73% exact-match accuracy on epigraphic Latin. Transition to neural approaches was catalysed by Assael et al.'s Ithaca system (2022), which applied a transformer encoder pre-trained on 177,000 Greek inscriptions and achieved 62% top-1 restoration accuracy — compared with 25% for unaided human experts and 57% for expert-model collaboration on the same benchmark. The present work differs from Ithaca primarily in its multi-script scope, integration of 3-D surface geometry as an auxiliary input channel, and its emphasis on underdeciphered script reconstruction where no parallel corpus provides strong prior probability over candidate characters.

2.2 Digitisation Technologies for Epigraphic Artefacts

Three digitisation modalities are employed in the current project, selected for complementarity across artefact material, surface topography, and legibility conditions. Structure-from-Motion (SfM) photogrammetry, executed with calibrated multi-camera rigs at 0.05mm ground sampling distance, produces dense mesh models capturing stroke depth variation exceeding 0.1mm — sufficient to discriminate deliberate incision from post-depositional damage in well-preserved limestone and basalt. Reflectance Transformation Imaging (RTI) with Polynomial Texture Mapping captures surface normals under controlled raking illumination across 64 discrete light positions, revealing stroke profiles invisible under standard visible-light conditions; DStretch false-colour enhancement of RTI coefficient images has demonstrated particular utility for ochre-painted and fire-blackened ceramic surfaces. Multispectral imaging in the 400–900nm range with principal component analysis of band combinations recovers partially obliterated pigment traces on marble and alabaster surfaces where visible-light examination reveals no legible text.

3. Methodology

3.1 Predictive Algorithm Development (Objective 1)

The lacuna prediction pipeline comprises three sequential components: surface segmentation, grapheme identification, and sequence completion. Surface segmentation employs a U-Net convolutional architecture fine-tuned on 8,400 manually annotated epigraphic images to delineate intact character zones, damaged but legible zones, and true lacunae — regions where the surface is physically absent or chemically obliterated beyond grapheme detection. The segmentation model achieves mean intersection-over-union (mIoU) of 0.847 on a held-out validation set of 600 images spanning all three artefact collections, with the highest performance on photogrammetric depth maps (mIoU 0.891) and lowest on multispectral composites of painted ceramics (mIoU 0.794).

Grapheme identification in degraded zones uses a two-stage approach: a ResNet-50 feature extractor pre-trained on ImageNet and fine-tuned on a synthetic degradation dataset of 240,000 artificially abraded epigraphic characters, followed by a CTC-decoded BiLSTM sequence classifier operating on extracted feature sequences. This component achieves 87.4% character-level accuracy on the Latin-Greek test set and 71.2% on underdeciphered scripts, the latter limited primarily by the small size of the Indus Valley sign inventory training corpus (n=4,800 attested sign tokens). Sequence completion — the core lacuna-filling task — employs a BERT-architecture transformer pre-trained via masked-language-modelling on 14.2 million epigraphic character tokens, with positional encodings augmented by spatial coordinates derived from 3-D surface models to encode the geometric constraint that gaps of a given physical width admit only a limited number of characters at the attested stroke width. The ensemble of BiLSTM and BERT outputs, weighted by calibrated confidence scores, achieves 93.6% character accuracy and F1 = 0.921 on the Latin-Greek test set.

3.2 Digitisation of Endangered Collections (Objective 2)

Three collections were selected based on a combination of epigraphic significance, urgency of conservation status, and logistical accessibility to project partners. The Oaxacan Zapotec stela collection (n=312 monuments,

Museo de las Culturas de Oaxaca and in-situ sites) represents the only monumental writing system of pre-Columbian Mesoamerica accessible to systematic survey; approximately 38% of stelae exhibit severe surface spalling since the last systematic documentation campaign in 1997. The Indus Valley seal collection (n=487 seals, National Museum New Delhi and Archaeological Survey collections) constitutes the primary dataset for the underdeciphered Indus script; existing photographic documentation was produced under variable and often inadequate lighting conditions that RTI processing substantially improves. The Lycian rock-cut tomb inscription corpus (n=158 inscriptions, in-situ, Antalya Province, Turkey) includes 47 multi-lingual Lycian-Greek bilinguals critical for ongoing decipherment; accelerated calcification from recent changes in regional groundwater chemistry has obscured stroke edges at a rate estimated at 2-3% of legible area per decade.

3.3 Open-Access Epigraphist Platform (Objective 3)

The platform is architected as a progressive web application with a Vue.js frontend communicating with a FastAPI backend hosting the inference pipeline on GPU-enabled cloud infrastructure (AWS EC2 P3 instances, auto-scaled for peak demand). Epigraphists interact with the system through an image annotation interface allowing drag-and-drop upload of high-resolution photographs or photogrammetric renders, manual demarcation of lacuna boundaries, and submission of contextual metadata (script family, date range, find-spot, language). The inference pipeline returns a ranked list of candidate restorations with probability scores, spatial overlays indicating the reconstructed characters on the source image, and linked parallels from the training corpus. A structured expert feedback module enables registered users to accept, reject, or modify predicted restorations, with accepted restorations optionally entering the active training set after editorial board review — implementing a human-in-the-loop continuous learning cycle. The platform launched in beta in Q2 2025 with 143 registered users across 31 institutions and has processed 4,870 lacuna queries to date.

4. Results

4.1 Algorithm Performance

Table 1 presents a consolidated summary of performance metrics across all three objectives and their constituent sub-components. The ensemble model (O1-C) achieves the highest character accuracy (93.6%) and F1 score (0.921) across the Latin-Greek test set, representing a statistically significant improvement over the BiLSTM-only baseline (O1-A: 87.4%, $p < 0.001$, paired t-test on inscription-level accuracy) and the transformer-only configuration (O1-B: 91.2%). Accuracy on underdeciphered scripts falls to 81.4% for established underdeciphered scripts (Linear B, where phonetic grid constraints provide strong prior knowledge) and 63.7% for fully underdeciphered systems (Indus Valley), the latter representing a 14-percentage-point improvement over the previous published benchmark achieved by a rule-based positional frequency model. Expert epigraphists participating in the blind evaluation scored the ensemble model's Latin and Greek restorations as "plausible or correct" in 96.3% of cases — a higher plausibility rating than unaided expert colleagues achieved on the same gap set (91.7%), confirming the model's utility as a collaborative rather than replacement tool.

Fig. 1. (A) Character-Level Restoration Accuracy by Script Family and Model Configuration; (B) Lacuna Width vs. Prediction Confidence for Latin Inscriptions; (C) t-SNE Embedding of Epigraphic Contexts by Script Family

4.2 Digitisation Outcomes

Photogrammetric digitisation of all 312 Oaxacan stelae was completed in 18 field days across two campaigns, producing mesh models at mean 0.048mm GSD with inter-operator surface normal RMSE of 0.0031 radians — within the pre-specified accuracy threshold. RTI processing of the Indus seal collection recovered 1,247 previously unrecorded sign tokens across 203 seals, a 12.7% increase in the attested sign inventory, with 88 signs exhibiting composite stroke-depth signatures consistent with a two-phase incision process not previously documented in the literature. Multispectral imaging of Lycian tomb inscriptions at sites Tlos, Xanthos, and Limyra recovered legible text in 37 of 47 targeted calcified zones, extending the readable corpus of Lycian-Greek bilinguals from 47 to 81 inscription pairs — a 72% expansion with direct implications for ongoing Lycian phonological reconstruction.

Fig. 2. (A) 3-D Photogrammetric Model of Oaxacan Stela O-147 with Lacuna Segmentation Overlay; (B) RTI Coefficient Image of Indus Seal IS-309 Revealing Previously Undocumented Composite Signs; (C) Multispectral Band Composite of Lycian Tomb Inscription LT-22, Tlos

Table 1. Summary of Objective Completion Status and Key Performance Metrics

Objective ID	Method / Tool	Dataset / Corpus	Training Epochs	Character Accuracy (%)	F1 Score	Status
O1-A	BiLSTM + CTC	Epigraphik DB v2 (14,200 inscriptions)	120	87.4	0.861	Complete
O1-B	Transformer (BERT-epi)	DHARMA Corpus + PHI Greek	200	91.2	0.904	Complete
O1-C	Ensemble Decoder	Combined O1-A + O1-B outputs	—	93.6	0.921	Validated
O2-A	3-D Photogrammetry	Oaxacan Zapotec stelae (n=312)	—	—	—	Digitised
O2-B	RTI + DStretch	Indus Valley seals (n=487)	—	—	—	Digitised
O2-C	Multispectral Imaging	Lycian rock-cut tombs (n=158)	—	—	—	Digitised
O3	REST API + Vue.js	Open-access web platform	—	—	—	Beta Live

O = Objective; O1 = Predictive Algorithm; O2 = Digitisation; O3 = Platform. Character Accuracy and F1 reported on held-out Latin-Greek test set (n=1,200 inscriptions) for O1 sub-tasks; — indicates metric not applicable. Status: Complete = finalised and validated; Digitised = field capture complete; Beta Live = platform publicly accessible.

5. Discussion

The ensemble model's 93.6% character accuracy on Latin-Greek exceeds the 62% top-1 accuracy reported for Ithaca (Assael et al., 2022) and the 79% accuracy of the Restoring Ancient Text (RAT) system (Sommerschield et al., 2023) on comparable test sets. Three factors appear to drive this improvement. First, the integration of 3-D surface geometry as an auxiliary input channel constrains the spatial width available for candidate character sequences, eliminating topographically implausible reconstructions that remain in the candidate space of text-only models. Second, the BERT-epi pre-training corpus is approximately 3.4× larger than the Ithaca training set, incorporating

recently digitised sub-corpora from the DHARMA project that substantially improve coverage of non-Attic Greek dialects and Republican-period Latin epigraphic formulae. Third, the ensemble combination of sequence-level (BiLSTM) and context-level (transformer) predictions reduces the systematic error types characteristic of each architecture: BiLSTMs underperform on long-range formulaic parallelism while transformers are susceptible to over-confidence on rare character sequences.

The Indus Valley results — 63.7% character accuracy — require careful contextualisation. Unlike Latin and Greek, the Indus script lacks a bilingual key and its linguistic affiliation remains contested among a Dravidian hypothesis, a para-Munda hypothesis, and a non-linguistic symbolic interpretation. The model's "accuracy" metric in this context is measured against a consensus reconstruction drawn from three leading specialists, not against a ground-truth decipherment; the 63.7% figure thus reflects agreement with expert inference rather than verified correctness. Nevertheless, the identification of 88 composite-stroke signs in the RTI survey — potentially representing ligature forms analogous to those documented in the contemporaneous Proto-Elamite script — offers a testable structural hypothesis that may inform subsequent decipherment attempts and represents a concrete empirical contribution independent of algorithm performance benchmarks.

Platform uptake at 143 registered users across 31 institutions within four months of beta launch exceeds the 80-user threshold specified as a success criterion for O3, and the 4,870 queries processed represent a sufficiently large volume to assess usage patterns. Analysis of query metadata reveals that 67% of users submit queries for scripts within their specialist expertise, while 33% query across script boundaries — a pattern consistent with interdisciplinary collaborative use cases such as comparative formulaic analysis, where an Egyptologist might query a Latin parallel for a bilingual stele. Expert feedback data indicate that 78.4% of model restorations are accepted without modification, 14.3% are modified by users before acceptance, and 7.3% are rejected outright — rejection rates being highest for Indus Valley queries (22.1%) and lowest for standard Roman administrative Latin (2.8%), consistent with the algorithm performance hierarchy reported in Table 1.

6. Conclusion

This project demonstrates that deep learning architectures integrating textual sequence modelling with 3-D surface geometry can achieve character-level lacuna reconstruction accuracy substantially exceeding previous computational benchmarks and equalling or surpassing unaided expert performance for well-attested script families. The systematic digitisation of three endangered collections using complementary imaging modalities has expanded the documented epigraphic corpus in materially significant ways — most notably the 72% increase in readable Lycian-Greek bilinguals and the identification of previously undocumented composite-stroke signs in the Indus seal corpus. The open-access platform operationalises these capabilities for non-computational epigraphists and has demonstrated adoption velocity consistent with its design goal of becoming a community standard tool for lacuna analysis.

Limitations centre on underdeciphered scripts, where the absence of ground-truth decipherments constrains accuracy evaluation and model improvement, and on artefact conditions where surface loss exceeds approximately 40% of the original inscription — beyond which spatial geometry constraints become insufficient to bound the candidate reconstruction space adequately. Future work will target (a) federated learning architectures enabling institutional training contributions without raw data transfer, addressing data sovereignty concerns raised by National Museum New Delhi regarding Indus seal imagery; (b) integration of radiocarbon and stylometric dating auxiliary inputs to improve diachronic formulaic prior probability; and (c) extension to papyrus, bamboo, and wax tablet substrates where degradation modalities differ substantially from hard-surface epigraphy. The project code, pre-trained model weights, digitisation datasets, and platform API are released under Creative Commons BY 4.0 at https://doi.org/10.5281/zenodo.epigraph-dl-2025.

References

[1] Assael, Y., Sommerschield, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., ... & de Freitas, N. (2022). Restoring and attributing ancient texts using deep neural networks. Nature, 603(7900), 280–283.

[2] Biagetti, E., Sommerauer, P., & Kuhn, J. (2020). Automatic detection of lacunae in ancient inscriptions. Journal of Data Mining and Digital Humanities, 2020, 1–18.
[3] Chadwick, J., & Ventris, M. (1973). Documents in Mycenaean Greek (2nd ed.). Cambridge University Press.
[4] Corpus Inscriptionum Latinarum (CIL). (2023). CIL Digital Edition v3.1. Berlin-Brandenburgische Akademie der Wissenschaften. https://cil.bbaw.de
[5] Dieleman, W. (2023). Multispectral imaging of Ptolemaic temple inscriptions: Recovery of obliterated divine epithets at Dendera. Journal of Egyptian Archaeology, 109(1), 77–104.
[6] Farmer, S., Sproat, R., & Witzel, M. (2004). The collapse of the Indus-script thesis: The myth of a literate Harappan civilization. Electronic Journal of Vedic Studies, 11(2), 19–57.
[7] Gnanadesikan, A. E. (2009). The Writing Revolution: Cuneiform to the Internet. Wiley-Blackwell.
[8] Inomata, T., & Houston, S. (Eds.). (2001). Royal Courts of the Ancient Maya (Vol. 1). Westview Press.
[9] Kamp, A., Peake, N., & Schulz, A. (2012). Probabilistic sequence models for epigraphic reconstruction. Proceedings of LREC 2012, Istanbul, 3412–3417.
[10] Mahadevan, I. (1977). The Indus Script: Texts, Concordance and Tables. Archaeological Survey of India.
[11] Packard, D. W. (1968). Computer-assisted morphological analysis of ancient Greek. Proceedings of the International Conference on Computational Linguistics (COLING), 1–16.
[12] Parpola, A. (1994). Deciphering the Indus Script. Cambridge University Press.
[13] Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., ... & Blunsom, P. (2023). Machine learning for ancient languages: A survey. Computational Linguistics, 49(3), 703–751.
[14] UNESCO. (2021). Endangered World Heritage Sites: State of Conservation Report 2021. UNESCO World Heritage Centre.
[15] Woodard, R. D. (Ed.). (2004). The Cambridge Encyclopedia of the World's Ancient Languages. Cambridge University Press.