Article

Contextualizing ancient texts with generative neural networks

https://doi.org/10.1038/s41586-025-09292-5

Received: 25 December 2024

Accepted: 16 June 2025

Published online: 23 July 2025

Open access

Check for updates

Yannis Assael[1,8]✉, Thea Sommerschield[2,8]✉, Alison Cooley[3], Brendan Shillingford[1], John Pavlopoulos[4], Priyanka Suresh[1], Bailey Herms[5], Justin Grayston[5], Benjamin Maynard[5], Nicholas Dietrich[1], Robbe Wulgaert[6], Jonathan Prag[7], Alex Mullen[2] & Shakir Mohamed[1]

Human history is born in writing. Inscriptions are among the earliest written forms, and offer direct insights into the thought, language and history of ancient civilizations. Historians capture these insights by identifying parallels—inscriptions with shared phrasing, function or cultural setting—to enable the contextualization of texts within broader historical frameworks, and perform key tasks such as restoration and geographical or chronological attribution[1]. However, current digital methods are restricted to literal matches and narrow historical scopes. Here we introduce Aeneas, a generative neural network for contextualizing ancient texts. Aeneas retrieves textual and contextual parallels, leverages visual inputs, handles arbitrary-length text restoration, and advances the state of the art in key tasks. To evaluate its impact, we conduct a large study with historians using outputs from Aeneas as research starting points. The historians find the parallels retrieved by Aeneas to be useful research starting points in 90% of cases, improving their confidence in key tasks by 44%. Restoration and geographical attribution tasks yielded superior results when historians were paired with Aeneas, outperforming both humans and artificial intelligence alone. For dating, Aeneas achieved a 13-year distance from ground-truth ranges. We demonstrate Aeneas' contribution to historical workflows through analysis of key traits in the renowned Roman inscription *Res Gestae Divi Augusti*, showing how integrating science and humanities can create transformative tools to assist historians and advance our understanding of the past.

The Roman world was a written world. Inscriptions were ubiquitous in public and private spaces, their communicative power shaped not only by the written text but also by their physical form and placement[2,3]. It is estimated that about 1,500 new Latin inscriptions are discovered every year[4], ranging from the decrees of emperors to the epitaphs of enslaved individuals, and preserving precious information on the cultural and linguistic life of an empire that spanned 5 million square kilometres and more than 2,000 years[5,6]. The study of these inscriptions, known as the discipline of epigraphy, faces numerous challenges: letters, words or entire sections of an inscription may be lost over time, with the full extent of the missing text often being unknown. High levels of human mobility, absence of explicit dates and the frequent use of abbreviations, a hallmark of the Latin epigraphic habit, further complicate our interpretation of these inscribed artefacts[7,8].

The resulting tasks of textual restoration, geographical and chronological attribution (Fig. 1) depend on specialist historians situating inscriptions within their wider linguistic and historical setting. A key method for this process of contextualization involves identifying parallels—inscriptions that share similar words, phrases, formulae or broader social, linguistic and cultural analogies[1]. Given the far-reaching communication networks of ancient societies, such connections often span vast geographical and temporal distances. By linking an inscription to this network of parallel texts and embedding it within its broader epigraphic culture, historians can refine its interpretation, reducing reliance on speculative, subjective hypotheses and isolated readings[9,10]. However, this contextualization is time-consuming, labour-intensive and highly specialized, requiring scholars to compare inscriptions against potentially hundreds of parallels. This demands extraordinary erudition, long-term knowledge acquisition, access to extensive library and museum collections, and repeated consultation of reference works—often using laborious manual searches or string-matching techniques. Consequently, scholars tend to develop regional and chronological specializations, which can limit the identification of epigraphic and historical connections at scale. We address the critical challenge of contextualization in ancient history and support historians in grounding their work using generative artificial intelligence (AI).

In recent years, the study of ancient languages has increasingly benefited from machine learning systems, which have advanced a range of tasks from digitization to decipherment[11–13], with several breakthroughs in the epigraphic domain[14–17]. Building on this momentum, we formulate and address the challenge of contextualizing inscriptions as a machine learning problem. In addition, our work is expanded to include two key

[1]Google DeepMind, London, UK. [2]Department of Classics and Archaeology, University of Nottingham, Nottingham, UK. [3]Department of Classics and Ancient History, University of Warwick, Warwick, UK. [4]Department of Informatics, Athens University of Economics and Business, Athens, Greece. [5]Google, Mountain View, CA, USA. [6]Sint-Lievenscollege, Ghent, Belgium. [7]Faculty of Classics, University of Oxford, Oxford, UK. [8]These authors contributed equally: Yannis Assael, Thea Sommerschield. ✉e-mail: yannisassael@google.com; thea.sommerschield@durham.ac.uk

Fig. 1 | Restoration of a damaged inscription. Fragment of a bronze military diploma from Sardinia, issued by the emperor Trajan to a sailor on a warship. 113/14 CE (*CIL* XVI, 60, The Metropolitan Museum of Art, Public Domain).

complementary functions. First, whereas modern epigraphic practice incorporates physical characteristics—shape, iconography and material—alongside textual content, AI approaches remain largely text-centric. Integrating multimodal models that combine textual and visual data is essential to fully situate inscriptions within their broader epigraphic landscape[11,18]. Second, whereas current machine learning methods have been successful in restoring gaps whose length is known[14], the challenge of arbitrary-length restoration—restoring gaps where the length of the missing text is uncertain (Fig. 1)—has not yet been addressed for ancient languages[19]. By prioritizing contextualization, integrating multimodality and advanced text restoration techniques, we demonstrate how AI can transform the study of inscriptions, advancing our understanding of the written cultures of the Roman world.

Contextualizing the past

This work presents Aeneas, a multimodal generative neural network for contextualizing Latin inscriptions, and sets the state of the art in the three key epigraphic tasks of restoration and geographical and chronological attribution. Aeneas incorporates a contextualization mechanism, which provides historians with a list of historically grounded textual and contextual epigraphic parallels to support their research. To capture a broader spectrum of information concerning the material dimension of inscriptions, Aeneas integrates both images and transcribed text as input, and is the first model to generate ancient text restorations of arbitrary length.

The name Aeneas is inspired by the wandering hero of Graeco-Roman mythology: like Aeneas, who journeyed from Troy across the Mediterranean seeking guidance on where to found the future city of Rome, our model seeks to uncover epigraphic parallels to ground historical research and link the past to the present. Our work demonstrates how AI can help historians detect previously unidentified parallels, and increase their confidence in tackling epigraphic tasks. By conducting an extensive collaborative historian–AI evaluation, we further showcase the model's cooperative performance compared with previous approaches. Finally, we illustrate the model's real-world impact as a research tool that is fully integrated in historical workflows by developing a case study in which Aeneas is applied to the study of the renowned Roman monumental inscription *Res Gestae Divi Augusti* (*RGDA*), authored by the emperor Augustus.

Integrating Latin epigraphic data

Latin is among the most extensively studied ancient languages, and frequently serves as a focal point for computational linguistics research[11] and related competitions[20]. Despite Latin inscriptions being the most abundant form of epigraphic evidence from the ancient world, relatively few machine learning studies[21–24] have focused specifically on them. We therefore focus on Latin inscriptions, as this gap offers a valuable opportunity for interdisciplinary research with broader scholarly impact.

To construct a comprehensive corpus for training Aeneas, we combine three of the most extensive Latin epigraphy databases: the Epigraphic Database Roma (EDR) (https://www.edr-edr.it), the Epigraphic Database Heidelberg (EDH) (https://edh.ub.uni-heidelberg.de) and the Epigraphik-Datenbank Clauss-Slaby ETL (EDCS_ETL) (https://github.com/sdam-au/EDCS_ETL). To harmonize these corpora, we developed a complex pipeline to standardize metadata, and disambiguate and process texts into a machine-actionable format using their unique Trismegistos identifiers (http://www.trismegistos.org). Additionally, we also source the images of inscriptions from these datasets, when available. We refer to this processed corpus as the Latin Epigraphic Dataset (LED), encompassing inscriptions from the seventh century BCE to the eighth century CE, with a geographical coverage ranging from the Roman provinces of Britannia (modern-day Britain) and Lusitania (Portugal) in the west, to Aegyptus (Egypt) and Mesopotamia (Iraq) in the east. The final LED comprises 176,861 inscriptions (totalling 16 million characters), most of which have damaged sections, and we were able to obtain corresponding images of 5% of inscriptions. LED was partitioned into training, validation and test sets on the basis of the last digit of the unique inscription identifiers, thereby ensuring an even distribution of images across the subsets (Extended Data Table 1). Further details on this process are provided in the Methods.

Contextualizing inscriptions with Aeneas

Aeneas takes as input the image of an inscription and its textual transcription (Fig. 2). Its efficient architecture operates exclusively on characters, avoiding the need for additional word-level representations implemented by previous approaches[14,15]. To guide the model, two special characters are utilized—'-' indicates that the number of characters to restore is known, whereas '#' signals that it is unknown—and the input image is processed through a shallow vision neural network. The input text is handled by the model's core, referred to as the 'torso'. The torso is a deep narrow T5 transformer[25] decoder that is augmented with relative positional rotary embeddings[26] to effectively capture textual information. The outputs of the torso and the vision network are then directed to specialized neural networks, referred to as 'heads', each tailored to address the three key epigraphic tasks. For handling the restoration of lacunae of unknown character length, we introduce an auxiliary head to predict whether more than one character is missing at any given decoding step. For the geographical attribution task, Aeneas outputs a predictive distribution that classifies the target inscription among 62 Roman provinces. For the chronological attribution task, Aeneas assigns a dating estimate in decades. For the text restoration task, Aeneas produces multiple possible restoration hypotheses, generated using beam search and ranked jointly by probability and length. Aeneas' predictions for each task are accompanied by saliency maps[27], which identify the textual and image features that most influenced the model's output.

It should be noted that only the geographical attribution head incorporates the additional inputs from the vision network—the restoration and chronological attribution tasks do not use the visual modality. The visual input was excluded for the restoration task to prevent unintended information 'leakage'; as parts of the text are artificially masked without their exact locations in the image being unknown, the model would exploit visual cues to infer and restore the hidden characters, compromising the integrity of the task. The visual modality was also omitted for the dating task because experiments showed no significant performance gains, probably owing to the model already achieving near-optimal results.

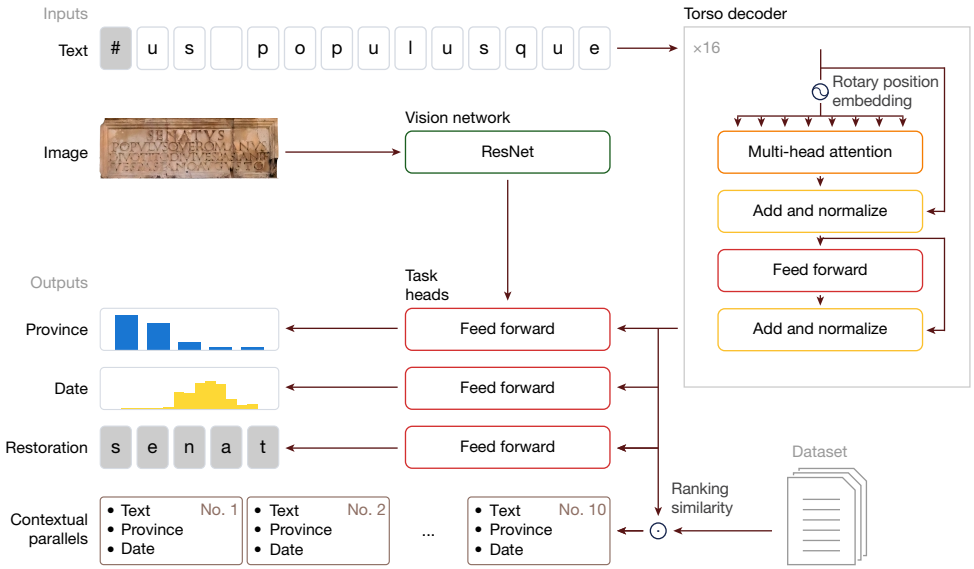


Fig. 2 | Processing of a textual transcription by the Aeneas architecture. Processing of the phrase *Senatus populusque Romanus* ('The Senate and the people of Rome') by Aeneas. Given the image and textual transcription of an inscription (with damaged sections of unknown length marked with '#'), Aeneas uses a transformer-based decoder (the torso) to process the text. Specialized networks (heads) handle character restoration, date attribution and geographical attribution (which also incorporates visual features). The torso's intermediate representations are merged into a unified, historically enriched embedding to retrieve similar inscriptions from the LED, ranked by relevance. Photograph of the arch of Titus by T.S.

As for the process of contextualizing inscriptions, Aeneas retrieves a list of the most relevant epigraphic parallels from the training set of LED. This process relies on historically rich embeddings—mathematical representations that capture the historical and linguistic patterns of the text, enabling comparisons based on both meaning and context. To produce these embeddings, Aeneas integrates the intermediate representations generated between the torso and the heads into a unified embedding vector. Unlike traditional text embeddings, this representation is enriched with historical context derived from the three key epigraphic tasks. This design enables the model to surpass traditional fuzzy string-matching methods, to include a wealth of epigraphic parallels from relevant places and periods, related concepts, synonymous terms, formulaic variations and analogous epigraphic practices. Finally, Aeneas scores all potential parallels against the input text using cosine similarity, ranking them by relevance. This ranked list, presented to experts alongside geographical and chronological metadata, provides a valuable starting point for historical research.

Contributing to historical research

To evaluate the potential of Aeneas as a foundational tool for historical research, we conduct a comprehensive human-centric evaluation involving the largest 'ancient historian and AI' collaborative study to date. We incorporate established metrics from previous research on restoring, placing and dating inscriptions, and also introduce a new measure to assess the impact of Aeneas' contextualization mechanism. This approach allows us to quantify the co-performance of Aeneas alongside human participants, and assesses historians' subjective experiences using Aeneas' parallels to support their predictions across the three epigraphic tasks. Additionally, we measure the cooperative performance of human specialists and AI in geographical and chronological attribution tasks. Finally, we show how Aeneas outperforms the previous state of the art using automated metrics.

Metrics and synergistic evaluation

To measure the effectiveness of different approaches, we adopt the evaluation metrics introduced by the previous state-of-the-art model, Ithaca[15]. For the restoration task, we simulate the damage suffered by an inscription by artificially corrupting text segments. Historians were asked to restore 1–10 characters (with the target restoration length disclosed), a practical range given the experimental setting and imposed time constraints. By contrast, Aeneas was challenged to restore 1–20 characters without knowing the target length. Aeneas' performance was then compared to that of Ithaca, which was retrained on the LED to support Latin and the restoration of lacunae of arbitrary length. Historians and models were also tasked with dating and placing target inscriptions.

Restoration performance was measured using the character error rate (CER) and the top-20 accuracy. For geographical attribution, we measured the top-1 and top-3 accuracy among 62 possible Roman provinces. For dating, we calculated the distance between the predictive mean date and the ground-truth ranges. To assess the impact of Aeneas' contextualization mechanism on historical research methods, we evaluate how many of Aeneas' retrieved parallels were accepted by the evaluated historians as relevant and useful to the key tasks. Finally, as an estimate of the inherent difficulty of epigraphic tasks using traditional historical methods, we introduce an automated 'onomastics' baseline. This baseline simulates how historians infer geographical and chronological indicia from personal names across known inscriptions.

Evaluating contextualization impact

The 'ancient historian and AI' study involved 23 participants with epigraphic expertise, ranging from masters students to professors, who engaged with Aeneas in an experimental simulation of real-world research workflows under a time constraint.

The evaluation is split into three stages. In each stage, participants are tasked with restoring, dating and placing a set of inscriptions drawn from a subset of 60 inscriptions in the LED test set. To assist with their workflow, historians are given access to the LED training set, which includes 141,000 inscriptions and associated metadata (date and place of writing), enabling them to manually search for parallel texts and record any inscriptions they found useful for completing the tasks. In stage 1, each historian is assigned five target inscriptions to assess their solo performance on the three epigraphic tasks and establish a baseline. In stage 2, for each target inscription, historians are provided with ten parallels retrieved by Aeneas from the LED training set and asked to

Table 1 | Historians' performance on epigraphic tasks with varying levels of Aeneas support

Method	Restoration 10-character CER ↓	Province		Date distance ↓ (years)	Confidence ↑	Research start ↑	Parallels added ↑
		Top-1 ↑	Top-3 ↑				
Onomastics	-	13.7%	23.5%	30.4	-	-	-
Historian	39.0%	27.0%	42.0%	31.3	49.5%	-	-
Historian with Aeneas parallels	33.9%	36.7%	56.7%	21.1	61.1%	75.0%	1.48
Historian with Aeneas parallels and prediction	**21.4%**	**68.3%**	**78.3%**	14.1	**70.0%**	**90.0%**	**1.58**
Aeneas	23.1%	66.7%	73.3%	**12.8**	-	-	-

Historians' performance on three epigraphic tasks (restoration, geographical attribution and dating) using 60 inscriptions from the LED test set. Tasks were performed independently, then assisted by Aeneas' parallels (historian with Aeneas parallels) or by its parallels and predictions (historian with Aeneas parallels and prediction). Metrics include restoration (CER, lower is better), geographical attribution (top-1 and top-3 accuracy), dating (distance in years), historian's confidence, use of Aeneas' parallels as research starting points, and the number of parallels used. Arrows (↑ and ↓) indicate the direction of optimal performance for each metric. The highlighted values indicate the top performing method.

complete the tasks again. This stage is crucial for measuring how contextual parallels influence historians' working hypotheses. Historians are invited to revise their list of manually retrieved parallels as needed, incorporating any Aeneas-retrieved parallels they found useful. In stage 3, historians are given Aeneas' restoration and attribution predictions for the target inscriptions to isolate their impact on the historians' predictions. All experts completed stage 1, and subsequently for each inscription they were assigned to stage 2 or stage 3 in an alternating sequence: the rationale behind this design is that Aeneas' predictions alone would significantly influence human performance metrics, but they would lack the relevant contextual grounding provided by epigraphical parallels. To explore this further, we also ask historians to score their confidence and report on their experience using Aeneas in a final survey at each evaluation stage.

Our evaluation showed that historians incorporated an average of 1.5 additional parallel inscriptions provided by Aeneas to their own manual selection of parallels (1.48 for 'historian with Aeneas parallels' and 1.58 for 'historian with Aeneas parallels and prediction'; values ranged from 0 to 6, median: 1; interquartile range: 0–2.5). In the survey, historians agreed that parallels generated by Aeneas enhanced their contextual understanding and interpretation of the target inscriptions. More specifically, when provided with Aeneas' parallels, historians reported that these could serve as a starting point for historical inquiry 75% of the time. This increased to 90% when Aeneas' predictions for the three epigraphic tasks were also included. Moreover, Aeneas' parallels boosted historians' confidence by an average of 23%, with an additional 21% increase when Aeneas' predictions were also available. These figures effectively demonstrate the significant role of Aeneas' retrieved parallels in the historian's workflow.

We also gathered qualitative feedback from participants on their subjective experience using Aeneas, with historians consistently emphasizing the value of Aeneas' contextualization mechanism in accelerating research and expanding the range of relevant parallels for the epigraphic tasks. For instance, one evaluated historian noted that "The parallels retrieved by Aeneas completely changed my perception of the (evaluated) inscription. I did not notice details that made all the difference in both restoring and chronologically attributing the text."

Similarly, another reported: "The help of parallel inscriptions is great for understanding the type of inscription of fellow soldiers setting up inscriptions, whereas my own search became more narrow zoning in on a set of inscriptions from Noricum. [Aeneas is] a nice parallel tool."

Finally, the impact of Aeneas on the speed and efficiency of research was repeatedly highlighted—for example: "The parallels retrieved by Aeneas completely changed my historical focus. […] it would have taken me a couple of days rather than 15 min [to find these texts]. Were I to base historical interpretations on these inscriptions' readings, now I would have days to write and frame the research questions rather than finding parallels."

Extended feedback by historians is available in the Methods.

In terms of overall accuracy across the three epigraphic tasks, Table 1 highlights the synergy between evaluated historians and the Aeneas model. For the restoration task, historians alone achieved a CER of 39%, which improved to 33% with the aid of Aeneas' parallels, and further dropped to 21% with Aeneas' predictive input, outperforming the solo performance of the model. These results suggest that the historians' performance was significantly enhanced when they used Aeneas' parallels and predictions. For the geographical attribution task, historians independently achieved 27% top-1 accuracy and 42% top-3 accuracy. With Aeneas' retrieved parallels, both metrics saw a 35% improvement. When Aeneas' predictions were also available, top-1 accuracy improved to 68%, a 152% increase, thereby surpassing the performance of Aeneas alone. For dating accuracy, historians averaged 31.3 years from ground truth date ranges, closely aligning with the onomastics baseline (established as a proxy for experts' traditional methods). Performance improved by 32% with Aeneas' contextualization support and by another 32% when Aeneas' predictions were provided, reducing the distance from ground truth date ranges to 14.1 years, approaching Aeneas' 12.8-year performance. All comparisons between historians working solo and with Aeneas' parallels and predictions show statistical significance (permutation test; restoration: $P < 0.01$, geographical attribution top-1: $P < 0.0009$ and top-3: $P < 0.001$, dating: $P < 0.014$). These findings demonstrate that Aeneas' contextualization mechanism could serve as a starting point for historical inquiry, boosts historians' confidence, and allows them to focus on evaluating predictions rather than manually compiling lists of potential textual analogies. The greatest synergistic impact is observed when contextual information from retrieved parallels is provided alongside Aeneas' predictive hypotheses.

Breaking new ground in epigraphy

For the comparison in performance between Ithaca and Aeneas: in all evaluations described in Table 2, Aeneas consistently outperforms both the onomastics baseline and Ithaca, thereby demonstrating the effectiveness of Aeneas' architecture. When the restoration character length was provided, Aeneas achieved a CER of 40.5% and a top-20 prediction accuracy of 46.5%. However, Aeneas is designed to handle restorations of unknown length, which is crucial for real-world applications, where the damage suffered by an inscription may be extensive and the number of missing characters is unknown. Even with this added complexity, the CER for segments of unknown length was only 15% higher. For geographical attribution, Aeneas achieves 72% accuracy in predicting the correct Roman province of origin. This task includes a multimodal aspect, incorporating images as input, thus highlighting the importance of the visual modality, which outperformed the text-only modality. In the chronological attribution task, Aeneas dates texts within an average of 13 and a median of zero years from the ground-truth ranges provided by historians. A detailed performance analysis per decade and per region is available in Extended Data Figs. 3–6. The reported results between Aeneas and Ithaca are also statistically significant (permutation test,

Table 2 | Automated performance comparison with previous literature on epigraphic tasks

Method	Restoration fixed length	Restoration 20-character CER ↓	Restoration 20-character top-20 ↑	Province (Vision)		Date distance ↓ (years)
				top-1 ↑	top-3 ↑	
Onomastics	-	-	-	17.9%	30.9%	51.3
Ithaca[15]	✓	43.5%	44.2%	61.3%	75.3%	14.1
Aeneas	✓	**40.5%**	**46.5%**	72.3%	83.9%	**13.4**
Aeneas	×	66.1%	32.7%	72.3%	83.9%	**13.4**

Comparative performance of Aeneas and baseline models on the LED test set across three key epigraphic tasks. Unlike the previous state of the art, which operates under fixed restoration lengths, Aeneas demonstrates flexibility by handling both fixed and unknown restoration lengths. Arrows (↑ and ↓) indicate the direction of optimal performance for each metric. The highlighted values indicate the top performance method.

restoration: $P < 0.0001$, geographical attribution top-1: $P < 0.0001$ and top-3: $P < 0.0001$, dating: $P < 0.0005$).

Evaluating Aeneas in the real world

Grounding the *Res Gestae Divi Augusti*

To illustrate how Aeneas may be integrated into historical workflows, bridging traditional methods with state-of-the-art generative models, we used it to analyse the *Monumentum Ancyranum*, one of the most important inscriptions of the ancient world. It is inscribed on the walls of the Temple of Rome and Augustus in Ancyra (modern Ankara), and preserves the text of the *RGDA*. Famously described as the "queen of Latin inscriptions"[28], the *RGDA* records the account of his life composed by Augustus, first Roman emperor. Authored in Rome and copied across the Empire (epigraphic copies survive in modern Turkey, among which the *Monumentum Ancyranum* is the most complete), it details Augustus' achievements, his impact on the Empire and beyond, and the monumental transformations that he led. The *RGDA* is a fundamental source for understanding imperial ideology in Augustan Rome[29]. An expert historian on our team examined Aeneas' predictions and parallels for the *RGDA*, working together with our model. Our resulting study focuses on analysing Aeneas' textual and contextual parallels, saliency maps and predictions for the attribution of this text. Our aim was to examine how Aeneas would cope with the complexities around this inscription's dating and provenance (the idiosyncrasies of the text are discussed further in the Methods).

Our first finding was that Aeneas' chronological attribution of the full text of the *RGDA* reflects prevailing scholarly hypotheses, with a distribution exhibiting a strong bimodal pattern (Fig. 3a). The distribution shows a modest peak around 10–1 BCE and a higher, more confident peak spanning 10–20 CE. A closer examination of the model's predictions for each chapter of the *RGDA* reveals that Aeneas is not misled by the many consular dates mentioned in the text, which are unrelated to its date of composition. Instead, the model's dating predictions appear to be driven by granular linguistic information. Indeed, closer scrutiny of Aeneas' saliency maps for each *RGDA* chapter reveals that Aeneas is highly receptive to chronologically significant features, such as archaizing Latin orthography, linguistic formulae, references to historically specific institutions, monuments and personal names. For example, in the Heading paragraph, Aeneas' saliency map highlights the spelling of the word *aheneis*, which generally shifts to *aeneis* only in the first century CE. Aeneas also picks up on historically specific Latin institutions: for example, the title *princeps iuventutis* (chapter 14) was first awarded in 5 BCE to Gaius Caesar (Augustus' grandson). Monuments also function as chronological indicia, as illustrated by the Altar of Augustan Peace, commissioned by the Roman Senate in 13 BCE to honour Augustus' return to Rome after an absence of 3 years. This monument appears as an area of interest on Aeneas' saliency maps for chapter 12. Most notably, the saliency map highlights many words in chapter 32, in which many distinctive non-Roman personal names appear, which belong to specific chronological contexts. Just as a trained historian would note these features, Aeneas' saliency maps indicate the model's attention to such markers (a full list of orthographic shifts and historical references is made available in Extended Data Table 2).

These observations are further supported by an analysis of Aeneas' parallels for this inscription. The top-five parallels are all texts composed in Rome, even though the geographical findspots of the actual inscriptions are diverse. Among them are two inscribed copies of the Valerian Aurelian law of 19 CE, issued by the Senate to honour Germanicus (Tiberius' heir). The surviving fragments found in Rome and the Spanish copy of the decree appear among Aeneas' complete list of parallels for the *RGDA* (*Corpus Inscriptionum Latinarum* (*CIL*) VI, 40348–TM 262102; *AE* 1984, 508–TM 224627). The language used by the Senate in decrees honouring members of Augustus' family adopted features

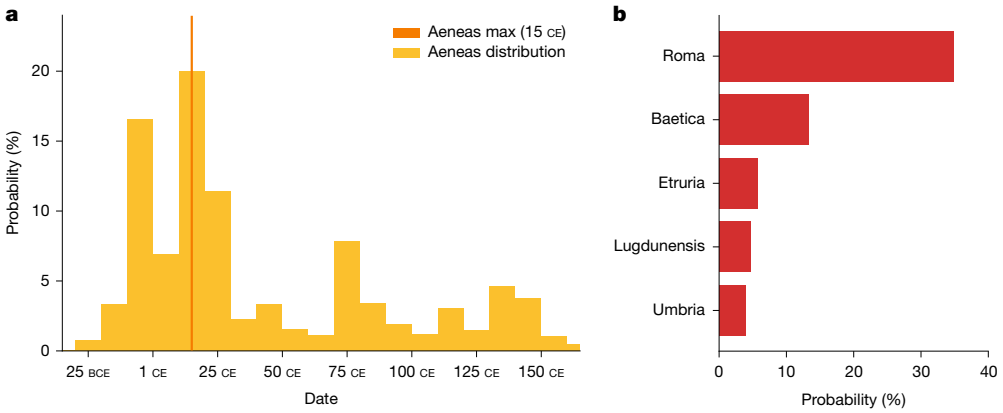


Fig. 3 | Aeneas' hypotheses for attribution of the *RGDA*, aggregated across its 35 chapters. The top-5 parallels retrieved by Aeneas were TM 262102, TM 558342, TM 224699, TM 535818 and TM 273657. Owing to length limitations, each chapter was processed individually. The resulting distributions were then averaged across all chapters. We report the maximum value from this averaged distribution, as it is less susceptible to noise arising from inter-chapter variance.

of Augustan imperial ideology, displaying strong verbal and contextual similarities to the *RGDA*. Other texts identified by Aeneas show a tendency to use archaizing orthography (a feature of Roman public legal documents), similarly to the *RGDA*, and just under half are public texts issued by Senate or emperor. This suggests that the geographical provenances of parallel inscriptions identified by Aeneas (Rome, Trento, Baetica and Ercolano) are secondary to their shared function as expressions of imperial political discourse. This commonality, captured by Aeneas' parallels, explains their similar textual and contextual features. The epigraphic dissemination of these texts illustrates the spread of Roman imperial ideology beyond Rome.

In sum, this case study demonstrates the capabilities of Aeneas as an assistive tool in historical workflows. Its results align well with the insights of a world-class expert on the *RGDA*, who noted the parallels, attribution and granularity of Aeneas' saliency maps. By systematically analysing diachronic and linguistic patterns, Aeneas not only supports but also complements traditional historical dating and parallel-finding methods, providing a transformative tool for in-depth historical analyses.

Retrieving parallels across the Roman Empire

We also tested the effectiveness of Aeneas' contextualization mechanism on a representative inscription of a well-attested type, selecting as a case study a votive altar from Mogontiacum (Mainz), *CIL* XIII, 6665 (TM 211813, HD54789). Dedicated in 211 CE by the *beneficiarius consularis* Lucius Maiorius Cogitatus, the altar honours the *Deae Aufaniae* and *Tutelae loci*, reflecting common military devotional practices in the Western provinces[30]. Aeneas' predictions successfully situate the inscription within this broader epigraphic habit, with a dating estimate (214 CE, within the expected range) and geographical attribution (correctly placing the stone in Germania Superior, with related alternatives in Germania Inferior and Pannonia) that align well with historical expectations. The saliency maps highlight the consular dating formula and the invocation of the *Deae Aufaniae*, showing that Aeneas is homing in on the details that a historian would recognize as diagnostic markers of date and provenance. The model also effectively restores damaged text sequences with contextually appropriate predictions, demonstrating its capacity for nuanced epigraphic reasoning.

Even more striking is Aeneas' top parallel identification for this text: another votive altar, dedicated in 197 CE by *beneficiarius* Iulius Bellator and found near the same location in Mainz (*FM* 07-055 no. 16). This altar shares rare textual formulas and an identical iconographical type with Cogitatus' dedication, supporting the hypothesis that the later inscription was directly influenced by the earlier one. Aeneas retrieves this parallel alongside other inscriptions from Germania and Pannonia, all of which reflect interconnected epigraphic, historical and linguistic traditions. Although Aeneas does not have previous knowledge of the archaeological context or the spatial connection between these stones (this information is absent from the LED training data), it is nonetheless capable of recognizing the subtle yet meaningful contextual relationships between them, whereas traditional text-matching approaches may miss indirect linguistic or historical links. When used as an assistive tool by historians, who can integrate archaeological knowledge, Aeneas therefore supports a more robust and expansive analysis of religious, linguistic and historical dynamics across the Roman provinces. An extended discussion of this case study is available in the Methods, and a full visualization of Aeneas' outputs for this inscription is provided in Extended Data Fig. 2.

Conclusions

Aeneas represents a leap forward in the integration of AI within the study of ancient texts. It introduces a carefully designed mechanism for the crucial process of contextualization, enabling historians to capture large-scale, in-depth epigraphic and historical connections that might otherwise remain obscured. Aeneas' architecture outperforms the previous state-of-the-art model, offers multimodal capabilities, enables restoration of text sequences of unknown lengths, and can also be adapted to any ancient language and written medium (such as papyri, manuscripts or coinage). These features highlight its potential for augmenting datasets with textual and contextual parallels, or providing hypotheses for missing values, as well as serving as a modular component for enhancing larger dialogue-based language models.

The case studies examined demonstrate the reliability of Aeneas as a specialized AI aid for epigraphic research. The examination of the *RGDA* tested Aeneas' capacity to handle the compositional complexities of this inscription, and the analysis of Cogitatus' votive altar from Mainz illustrated its ability to systematically track granular diachronic and linguistic patterns. In both instances, Aeneas was able to leverage relevant epigraphic parallels and produce accurate predictions, aligning and representing scholarly hypotheses in a quantitative way. Together, these case studies highlight Aeneas' versatility across diverse epigraphic contexts. Whether applied to an imperial monument or a provincial votive inscription, Aeneas mirrors the analytical process of an epigrapher, complementing traditional historical methodologies and generating accurate, meaningful insights. These findings are supported by the results of an extensive historian–AI evaluation, in which historians confirmed that Aeneas can seamlessly integrate into research workflows and provide a transformative aid for historical inquiry. The public interface that we have released for historians to use Aeneas in their research is available at https://predictingthepast.com. In conclusion, Aeneas tangibly enhances the collaborative capabilities between human experts and AI in a mutually enriching intertwining of the sciences and the humanities.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-025-09292-5.

1. Robert, L. in *Les Épigraphies et l'Épigraphie Grecque et Romaine* (ed. Samaran, C.) 453–497 (Gallimard, 1961).
2. Panciera, S. What is an inscription? Problems of definition and identity of an historical source. *Z. Papyrol. Epigr.* **183**, 1–10 (2012).
3. Bodel, J. in *Epigraphic Culture and the Epigraphic Mode* (eds Benefiel, R. & Keesling, C.) 11–44 (Brill, 2023).
4. Alföldy, G. Il futuro dell'*epigrafia*. *In XI Congresso Internazionale di Epigrafia Greca e Latina* 87–102 (Edizioni Quasar, 1999).
5. Cooley, A. *The Cambridge Manual of Latin Epigraphy* (Cambridge Univ. Press, 2012).
6. Mattingly, D. J. *Imperialism, Power, and Identity: Experiencing the Roman Empire* (Princeton Univ. Press, 2013).
7. Adams, J. N. *The Regional Diversification of Latin 200 BC–AD 600* (Cambridge Univ. Press, 2007)
8. Clackson, J., James, P., McDonald, K., Tagliapietra, L. & Zair, N. (eds). *Migration, Mobility and Language Contact in and around the Ancient Mediterranean* (Cambridge Univ. Press, 2020).
9. Bodel, J. (ed.). *Epigraphic Evidence: Ancient History from Inscriptions* (Routledge, 2001).
10. MacMullen, R. The epigraphic habit in the Roman Empire. *Am. J. Philol.* **103**, 233–246 (1982).
11. Sommerschield, T. et al. Machine learning for ancient languages: a survey. *Comput. Linguist.* **49**, 703–747 (2023).
12. Fiorucci, M. et al. Machine learning for cultural heritage: a survey. *Pattern Recognit. Lett.* **133**, 102–108 (2020).
13. Narang, S. R., Jindal, M. K. & Kumar, M. Ancient text recognition: a review. *Artif. Intell. Rev.* **53**, 5517–5558 (2020).
14. Assael, Y., Sommerschield, T. & Prag, J. in *Empirical Methods in Natural Language Processing* (eds Inui, K. et al.) 6368–6375 (ACL, 2019).
15. Assael, Y. et al. Restoring and attributing ancient texts using deep neural networks. *Nature* **603**, 280–283 (2022).
16. Huang, H. et al. AGTGAN: Unpaired image translation for photographic ancient character generation. In *ACM International Conference on Multimedia* 5456–5467 (ACM, 2022).
17. Fetaya, E., Lifshitz, Y., Aaron, E. & Gordin, S. Restoration of fragmentary Babylonian texts using recurrent neural networks. *Proc. Natl Acad. Sci. USA* **117**, 22743–22751 (2020).

18. Petrovic, A., Petrovic, I. & Thomas, E. (eds). *The Materiality of Text: Placement, Perception, and Presence of Inscribed Texts in Classical Antiquity* (Brill, 2019).
19. Shen, T., Quach, V., Barzilay, R. & Jaakkola, T. Blank language models. In *Proc. 2020 Conference Empirical Methods in Natural Language Processing* (eds Webber, B. et al.) 5186–5198 (ACL, 2020).
20. Sprugnoli, R., Passarotti, M., Cecchini, F. M. & Pellegrini, M. Overview of the EvaLatin 2020 evaluation campaign. In *Proc. Workshop on Language Technologies for Historical and Ancient Languages* (eds Sprugnoli, R. & Passarotti, M.) 105–110 (ACL, 2020).
21. Molton, N. et al. Visual enhancement of incised text. *Pattern Recognit.* **36**, 1031–1043 (2003).
22. Terras, M. *Image to Interpretation: an Intelligent System to Aid Historians in Reading the Vindolanda Texts* (Oxford Univ. Press, 2006).
23. Tupman, C., Kangin, D. & Christmas, J. Reconsidering the Roman workshop: using computer vision to analyse the making of ancient inscriptions. *Umanistica Digitale* **10**, 461–473 (2021).
24. Kase, V., Heřmánková, P. & Sobotková, A. Classifying Latin inscriptions of the Roman Empire: a machine-learning approach. In *Proc. Conference on Computational Humanities Research 2021* (eds Ehrmann, M. et al.) 123–135 (2021)
25. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**, 1–67 (2020).
26. Su, J. et al. Roformer: enhanced transformer with rotary position embedding. *Neurocomputing* **568**, 127063 (2024).
27. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations* (2014).
28. Mommsen, T. Der Rechenschaftsbericht des Augustus. *Historische Zeitschrift* **57**, 385–397 (1887).
29. Cooley, A. *Res Gestae Divi Augusti: Text, Translation, and Commentary* (Cambridge Univ. Press, 2009).
30. Ehmig, U. & Haensch, R. Serie und individuum. Neue Benefiziarieraltäre aus Mainz. *Chiron* **53**, 107–152 (2023).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025

Methods

Previous work

In recent years, the field of machine learning for ancient languages has gained remarkable momentum, driven by increased digitization efforts (creating standardized datasets of texts, metadata and images of ancient written evidence), by advances in machine learning architectures (for example, the Transformer[31]) and increased computational power. This progress, spanning numerous languages, scripts and tasks, has been extensively documented in works such as refs. 11,32 as well as in task-specific studies[13,33,34].

Work on restoration (including the tasks of reassembling fragments, restoring text and enhancing quality) encompasses a number of machine learning methods, modalities and ancient written evidence, including inscriptions in cuneiform[17,35,36], ancient Greek[14,15], Linear B[37], Hebrew[38], old Chinese[39], Indus[40–42], old Cham[43], Oracle Bone[16,44]; as well as papyri in Coptic[45], Hebrew[46] ancient Greek[47,48]; and manuscripts in old Korean[49], ancient Shui[50] and Tamil[51]. One of the most closely related efforts to Aeneas is work on multimodal old Chinese ideograph restoration[52]. However, replicating this approach for Latin inscriptions is limited by the quality and consistency in annotation of existing datasets. Moreover, this method is confined to the reconstruction of single ideographs, and does not extend to broader epigraphic tasks. In terms of evaluating human performance on the task of restoration, Assael et al.[15] was the first to establish a measure of joint human–AI performance in a real-world setting, an evaluation framework that has since been adopted by subsequent studies[53].

As for the challenge of unknown text restoration, this has been approached primarily by Shen et al.[19], but their application to ancient languages uses a known restoration length benchmark.

Work on ancient text attribution, both geographical and chronological, is less common, and to our knowledge, only Assael et al.[15] has attempted to tackle together the three tasks of restoring, dating and placing ancient Greek inscriptions. Other notable efforts on dating include those on Kannada inscriptions[54], on Arabic manuscripts[55], on Coptic papyri[56], on old Chinese manuscripts[57], on Cuneiform tablets[58], on Oracle Bone inscriptions[59], on Korean Hanja[60], and on Greek papyri[61]. The only other work on geographical attribution is on Greek literary texts[62]. Although the findspot of an inscription often indicates its place of writing, geographical attribution becomes important in cases of objects that have been moved around during the ancient or medieval periods[63], or in light of early modern collecting habits, as well as the illicit trade in antiquities.

With regard to Latin, recent efforts have focussed on Latin literary evidence to tackle a range of tasks, from intertextuality[64], part-of-speech tagging[65], translation[66], authorship attribution[67,68] and literary text restoration[69]. But despite the existence of large-scale Latin epigraphic datasets, many of which use the EpiDoc XML encoding gold-standard[70–73] and include images of inscriptions, very little work has attempted to apply machine learning techniques to Latin epigraphy—although quantitative approaches to Latin epigraphy using statistical techniques are continuously breaking new ground[74–77]. Early efforts include work on the Vindolanda stylus tablets[21,22], attempting to develop an image processing and pattern recognition pipeline for character recognition. More recently[24], there has been work to develop a classifier to automate the identification and labelling types of Latin inscriptions from the poorly standardized EDCS_ETL dataset using patterns learnt from the more richly annotated EDH dataset; and applied text detection methods to segment characters and analyse letters across a large dataset of Latin inscription images to isolate letter-cutting workshops[23].

Latin Epigraphic Dataset

Dataset generation. To create the LED, we processed the EDR, EDH and EDCS_ETL databases, resulting in the largest machine-actionable Latin inscription dataset to date (Extended Data Table 1). These databases collect inscriptions from various Roman provinces and historical periods, enhancing the diversity and temporal scope of LED. All databases were available under a Creative Commons Attribution 4.0 license via *Zenodo* (the open repository for EU-funded research outputs). To ensure consistency across the LED dataset, we standardized all metadata relating to dates and historical periods, converting them to numerals within the range of 800 BCE to 800 CE. Inscriptions outside this range were excluded. Province names obtained from EDR, EDH and EDCS_ETL were also standardized and merged.

To render the text machine-actionable, we applied a filtering ruleset to systematically process human annotations. Historians' epigraphic annotations (the Leiden conventions) were either stripped or normalized to preserve the closest version of the original inscribed text. Latin abbreviations were left unresolved, whereas word forms showing alternative spellings for diachronic, diatopic or diastratic reasons (for example, *bixit* for *vixit*) were preserved to enable the model to learn their epigraphic, geographical or chronological specific variations. Missing characters restored by editors (conventionally annotated within square brackets, and typically restored on the basis of grammatical and syntactical patterns and the reconstructed physical layout of an inscription) were retained. Missing characters that cannot be definitively restored by editors (conventionally represented using hyphens as placeholders, with each hyphen corresponding to one missing character) were also retained. When the exact number of missing characters was indeterminate, we used the hash (#) symbol as a placeholder to denote this uncertainty. Extra spaces were collapsed to ensure clean and concise outputs. Non-Latin characters were stripped using an accent removal function, leaving only Latin characters, predefined punctuation and placeholders. Duplicate inscriptions were excluded using their unique Trismegistos identifiers when available, and supplemented by additional deduplication using fuzzy string matching and MinHash locality-sensitive hashing[78]: texts exceeding a 90% content similarity threshold were considered duplicates, resulting in the removal of one text from each identified pair. Inscriptions under 25 characters in length were filtered out to focus on substantial textual content, essential for the model's learning and generalization capacities. For dataset partitioning, inscriptions whose numerical Trismegistos (or in alternative EDCS_ETL) identifiers ended in 3 or 4 were held out and allocated to the test and validation sets respectively, following previous work[15].

Images were sourced exclusively from EDR and EDH. To maintain high data quality and ensure standardization across the dataset, we implemented an automated filtering process. This process removed drawings, squeezes and other non-photographic artefacts by applying thresholds to colour histograms, specifically targeting and eliminating images composed primarily of a single solid colour. Additionally, we utilized the variance of the Laplacian matrix to identify and discard blurry images, leveraging the principle that blurry images have lower variance in their colour continuity. The cleaned images were then converted to greyscale, as this was the predominant format in the original dataset. For each inscription, only a single representative image was kept, excluding non-inscribed surfaces.

Dataset limitations. Despite representing the largest machine-actionable corpus of Latin inscriptions compiled to date, the size of LED (16 million characters) remains a significant limitation compared with the scale of datasets typically used in state-of-the-art natural language processing research. This relative scarcity of data inevitably constrains the model's capacity to generalize and may limit its performance on rarer epigraphic phenomena or under-represented regions and periods. Crucially, the available corpus is also subject to inherent biases, most significantly inscription survival bias, potentially skewing the data towards certain materials, locations, or historical contexts. This limitation is even more pronounced for the image modality, where only approximately 5% of the textual inscriptions have corresponding images. As a result, although saliency maps provided valuable insights

for the textual modality in the geographical attribution task, the artefacts highlighted by the image saliency maps were often less interpretable by domain experts. Moreover, the task of chronological attribution could also potentially benefit from additional images, which might allow for better alignment with palaeographic arguments.

Extended Data Figs. 3–6 provide an extended performance analysis broken down by decade and province, revealing that performance often tends to be weaker where data is limited. We emphasize therefore that large, open, linked, standardized multimodal datasets are key for advancing the field, and hope that initiatives such as ours might demonstrate the impact of digital epigraphic publication and catalyse further efforts.

The question of data circularity. As was acknowledged in Assael et al.[15] (see 'Data circularity'), the dataset contains within it an element of circularity. Editors of inscriptions traditionally restore two elements: they expand symbols and abbreviations, identified using (), and they attempt to restore missing text, using []. Aeneas similarly offers hypothetical restorations for missing text. In preparing the dataset we removed expansions, notwithstanding that their expansion is normally almost certain, as these letters did not appear on the stone originally; however, we retained previous editors' restorations of text originally carved but now lost. Restorations are based on parallels, and contextual knowledge, and best practice is only to offer such restorations when they have a high level of confidence (as stated in a recent manual, 'one must not forget that the task is to restore the document, and not to remake it' (our translation from ref. 79, page 67)). Nonetheless, it may be objected that by including such previous restorations in the training set there is a risk of confirmation bias, especially as not all scholars are consistently rigorous. As the available datasets do not provide information on editorial responsibility, and do not provide consistent or documented access to alternative editions, alternative approaches such as controlling editorial quality, or even increasing the size of the dataset by including alternative editions, could not be adopted.

The primary motivation for inclusion of this material was the limited availability of data. In preparing the I.PHI dataset for Ithaca, we computed that, by excluding the text within square brackets, we would lose 20% of the total texts available. Because deep learning models can greatly benefit from vast amounts of data, and our dataset is multiple orders of magnitude smaller than recent NLP datasets, we wanted to harness all available information to avoid overfitting and assist generalization. To assess the impact of this decision, we conducted additional experiments to evaluate the reliability of outputs when retaining the conjectured textual restorations in square brackets. Specifically, we trained a new model excluding previous hypothetical conjectures and evaluated both models' performance on the test set without conjectures.

The differences between the models trained with and without conjectured restorations were less than 5%; with the model trained excluding the conjectured restorations underperforming in all tasks compared to our original manuscript model in the given evaluation setup. We concluded that the benefit of improved performance outweighed the risk of bias, and the given evaluation setup (that is, using a model which included conjectures) was selected in this case because the same approach was adopted in Assael et al.[15] and therefore allows us to compare with previous work and estimate the baseline performance. Epigraphers commonly refer to the phenomenon of "history from square brackets"[80], which describes the reliance for historical reconstruction on the conjectural restoration of specific information in individual texts. This particular risk is arguably much lower, as the model works as an information 'compressor', creating multiple levels of abstraction of the raw data, thereby vastly reducing the influence of any particular unwarranted and historically specific conjecture.

Nonetheless, the risk of a broader bias must be acknowledged, and future work might seek to address this, as the quality and quantity of the available data improve. Using a model trained on data excluding conjectural restorations, one might seek to test existing editorial restorations and so identify existing biases in previous editorial work, utilizing the model to identify outliers. Going a step further, such a model might even serve to identify more or less reliable editors among past epigraphers, and has a substantial role in the ongoing work of revising existing epigraphic editions.

Aeneas' architecture

Aeneas is trained to perform four primary tasks: the restoration of a set character length, the restoration of an unknown lacuna length, geographical attribution and chronological attribution.

The input provided to Aeneas' architecture for each inscription consists of a character sequence (including spaces) and a corresponding greyscale image of size 224 × 224. The maximum sequence length is 768 characters. Two special symbols are included in the input to annotate missing information: '-' for a single missing character and '#' for a missing segment of unknown length. Additionally, the sequence is padded with a start-of-sentence token '<'. The textual inputs are processed through the model's torso, which is based on a large-scale transformer architecture derived from the T5 (ref. 25) model and adapted to use rotary embeddings. The T5 model features an embedding dimension of 384, query-key-value dimensions of 32, and a multi-layer perceptron (MLP) size of 1,536. It consists of 16 layers, each with 8 attention heads. The torso outputs a sequence of embeddings with a length equal to the input sequence. Each embedding is a 1,536-dimensional vector. These embeddings are passed to four task-specific heads: restoration, unknown-length restoration prediction, geographical attribution, and chronological attribution. Each task head consists of a two-layer MLP followed by a softmax function. The model was trained for one week using 64 Tensor Processing Unit v5e chips on the Google Cloud platform, with batch size of 1,024 text–image pairs, using the LAMB[81] optimizer. The learning rate follows a schedule with a peak value of 3×10^{-3}, a warm-up phase of 4,000 steps and a total of 1 million steps. Bayesian optimization is used to fine-tune the loss (L) for each task, combining them as follows:

$$L = 3L_{\text{restoration}} + L_{\text{unknown}} + 2L_{\text{region}} + 1.25L_{\text{date}}$$

To mitigate overfitting, especially given the limited dataset size, several data augmentation techniques are applied during training. These techniques include up to 75% text masking, text clipping, word deletion, punctuation dropping, and image augmentations such as zooming, rotation, and adjustments to brightness and contrast. A dropout of 10% and label smoothing are also used, with smoothing rates of 5% for the restoration task and 10% for geographical attribution. This multi-task setup, combined with the training and augmentation strategies, allows Aeneas to achieve robust performance across all four epigraphic tasks.

Training Aeneas. To better understand the underlying processes during Aeneas' training, this section provides a detailed overview of the inputs and outputs involved in the model's restoration and attribution tasks.

For the restoration task, ground truths are obtained by artificially corrupting the inscription's text, masking up to 75% of their characters. Some of these masks are deliberately grouped into continuous segments to better simulate real-world damage. When the corruption length is known, Aeneas predicts the missing characters directly. For unknown-length restoration, an additional neural network head is incorporated, using binary cross-entropy to predict whether one or more characters are missing whenever the unknown-length symbol (#) is encountered. Furthermore, the model's architecture maintains alignment between input characters and task outputs. Aeneas' torso embeddings, corresponding to input text characters, are directly mapped to their positions in the sequence. For each missing character (each annotated with a '?'), the corresponding embedding is fed to the restoration task head, which predicts the missing character(s). For

unknown-length restoration, the additional task head is activated whenever the '#' symbol appears in the input sequence, determining whether a single or multiple characters are missing. This architecture enables the model to handle the restoration and attribution tasks efficiently, while maintaining alignment between input characters and task outputs.

To generate Aeneas' textual restoration predictions, we use a beam search with a beam width of 100. Additionally, we implement a non-sequential beam search that incorporates the unknown-length prediction. Each beam starts with the restoration candidate with the highest confidence score and proceeds iteratively, restoring the characters with the highest certainty at each time-step. If an unknown-length restoration character is found, a missing character is prepended, and two entries are appended to the beam: the first keeps the unknown-length symbol, while the other removes it. This approach accounts for both scenarios: whether more than one character needs to be restored or only a single character is missing. Geographical and chronological attribution tasks use the first output embedding of the torso (at $t = 1$), which is passed to their respective task heads. Geographical attribution predicts one of 62 Roman provinces using categorical cross-entropy with ground-truth labels, when available. Chronological attribution maps historical dates between 800 BCE and 800 CE into 160 discrete decades using binarized bins. Kullback–Leibler divergence is used to match predicted distributions with the ground-truth ranges provided by historians. The visual inputs are processed using a ResNet-8 (ref. 82) neural network. The resulting outputs are concatenated with the relevant textual embeddings and jointly processed by the geographical attribution head.

Finally, the effectiveness of saliency maps remains an active topic of discussion[83]. However, the historians on our team have generally found them to be a valuable explainability tool, particularly for textual inputs, and for this reason we decided to include them among the outputs.

Aeneas' contextualization mechanism. Aeneas' contextualization mechanism can be framed as an embedding within a multidimensional space, where each inscription is positioned so that the closest neighbours correspond to the parallels a historian would use to ground their research. In the absence of ground-truth data for contextualization, we construct this embedding space using the epigraphic tasks as proxies. This approach aligns textual and contextually relevant parallel inscriptions by bringing them closer within the space.

We measure proximity in this embedding space using cosine similarity to retrieve a list of parallel inscriptions, which was identified by our interdisciplinary team as an effective metric during preliminary evaluations. To construct the historically rich embedding space, we combine the output embeddings (emb) of Aeneas' torso with the following formulation:

$$\mathrm{emb}_{\mathrm{context}} = \left(\mathrm{emb}^{t=1}_{\mathrm{torso}} + \frac{1}{N} \sum_{n=2}^{N} \mathrm{emb}^{t=n}_{\mathrm{torso}} \right) \div 2,$$

where the $\mathrm{emb}^{t=1}_{\mathrm{torso}}$ represents the torso's first output ($t = 1$) which aligns with the sentence prefix. This embedding is critical for the chronological and geographical attribution task heads. The subsequent outputs ($t = 2 \ldots N$, where N is the length of the input string, including the prefix symbol) align with textual inputs and are used for restoration task heads.

To demonstrate the potential of Aeneas' historically rich embeddings in the contextualization task, we compare their performance against textual embeddings derived from a multilingual T5 model that includes Latin in its training set. Specifically, we focus on the chronological and geographical attribution tasks. For chronological attribution, we use a colour scale transitioning from blue (earliest dates in the dataset) to red (latest dates). For geographical attribution, we apply a colour scale based on the geographical coordinates of 62 provinces, with yellow representing the north, red representing the west, green representing the east and blue representing the south.

Extended Data Fig. 1 presents a visualization of the embedding spaces using uniform manifold approximation and projection (UMAP) dimensionality reduction[84]. Although it is important to acknowledge the inherent limitations of directly interpreting UMAP projections, the embeddings derived from Aeneas appear to exhibit smoother distributions and greater alignment with chronological and geographical labels. These observations suggest that Aeneas' embeddings may better capture the underlying structure of historical context, as evidenced by the clearer separation of clusters. By comparison, the embeddings generated by T5 display a greater overlap, thereby indicating potential challenges in distinguishing contextual attributes. This highlights the effectiveness of Aeneas' embeddings in capturing historical information and suggesting relevant parallel texts from similar epigraphic contexts[85–88].

Our interdisciplinary team further evaluated various trained retrieval methods, including embedding the texts and their metadata or using them as raw inputs. However, owing to the limited dataset size, our preliminary evaluation revealed that similarity scoring with Aeneas' embeddings yielded the most relevant inscriptions, and this intuition was supported by the evaluation of expert historians.

Evaluating Aeneas

Task metrics. We adopt the evaluation framework proposed by Assael et al.[15] for the tasks of restoration, geographical attribution and chronological attribution, while further refining it to enhance consistency and interpretability.

For textual restoration, the difficulty increases with the number of characters to be reconstructed. As described above, our evaluation pipeline artificially corrupts arbitrary spans of text to produce targets for restoration. To ensure a fair comparison of this stochastic pipeline across different levels of difficulty, we calculate performance metrics based on sequence length. Specifically, we compute the CER for each sequence length (ranging from 1 to 20 characters) as follows:

$$\mathrm{CER}_l = \frac{1}{\sum_{i=1}^{N} I_{\mathrm{len}_i = l}} \sum_{i=1}^{N} I_{\mathrm{len}_i = l} \times \frac{\mathrm{edit\ distance}(\mathrm{pred}_i, \mathrm{target}_i)}{l},$$

where I is the indicator function, len_i denotes the length of the ith sample, N is the total number of samples, pred_i represents the predicted sequence and target_i corresponds to the ground truth. We then average the CER values across all sequence lengths:

$$\mathrm{CER}_{\mathrm{score}} = \frac{1}{L} \sum_{l=1}^{L} \mathrm{CER}_l.$$

where $L = 20$ represents the maximum sequence length used in the evaluation. Additionally, we calculate the top-20 accuracy following the same stratified approach.

For geographical attribution, we evaluate performance using standard top-1 and top-3 accuracy metrics. While top-1 accuracy measures the model's ability to pinpoint the correct province out of 62, top-3 accuracy provides additional insights by assessing its capacity to offer plausible alternative suggestions, aiding historians in their analysis. Finally, for chronological attribution, the model generates a predictive distribution over possible dates. We use an interpretable metric to evaluate the temporal proximity between predictions and ground truth. The distance is computed based on the relationship between the predicted mean $\mathrm{pred}_{\mathrm{avg}}$ and the ground-truth interval defined by its minimum ($\mathrm{gt}_{\mathrm{min}}$) and maximum ($\mathrm{gt}_{\mathrm{max}}$) boundaries:

$$\mathrm{Years} = \begin{cases} 0, & \text{if } \mathrm{gt}_{\mathrm{max}} \geq \mathrm{pred}_{\mathrm{avg}} \geq \mathrm{gt}_{\mathrm{min}}, \\ \left| \mathrm{pred}_{\mathrm{avg}} - \mathrm{gt}_{\mathrm{max}} \right|, & \text{if } \mathrm{pred}_{\mathrm{avg}} > \mathrm{gt}_{\mathrm{max}}, \\ \left| \mathrm{pred}_{\mathrm{avg}} - \mathrm{gt}_{\mathrm{min}} \right|, & \text{if } \mathrm{pred}_{\mathrm{avg}} < \mathrm{gt}_{\mathrm{min}}. \end{cases}$$

Onomastics baseline. Personal names provide valuable insights for epigraphers, often serving as key indicators in attribution predictions[89]. Building on their significance within the broader epigraphic workflow, we introduce an onomastics baseline that exclusively leverages metadata derived from these personal names. Unlike earlier studies[15], which apply this method to a limited subset of data using human evaluators, our approach fully automates the process, enabling its application across the entire evaluation dataset and improving scalability. In the absence of a digital pre-compiled list of Roman onomastic components, we adapt the repository of proper names provided by the Classical Language Toolkit (https://cltk.org/). From this list, we manually removed 350 items that did not represent proper names, excluded shorter entries (one or two characters) due to their ambiguous usage, and eliminated those containing non-Latin characters, resulting in a curated list of approximately 38,000 proper names. The resulting list is available on our GitHub repository. To enhance the robustness of our method, we identify the most frequent word unigrams, bigrams and trigrams within the dataset (to capture *tria nomina* and other Roman onomastic features), retaining only those appearing more than five times. We further filter these n-grams to include only those composed entirely—or as a combination—of entries from the curated proper name list. For each identified n-gram, we compute the average chronological and geographical distributions across the training dataset, based on the ground truths of the texts in which they appear. Finally, when analysing a new inscription, we check which of these n-grams occur, aggregate their associated statistics, and use them to predict both the date and provenance of the inscription.

Historian–AI evaluation ethics protocol. One of the central components of this research was the historian–AI evaluation, the largest conducted to date. The goal was to assess the effectiveness of Aeneas' contextualization mechanism as a foundational tool in historical research. Our specially developed ethics protocol received a favourable ethical opinion by the Faculty of Arts Research Ethics Committee of the University of Nottingham. The evaluation involved 23 epigraphers who responded to our call for participants. All responses were anonymized. Each participant was assigned five target inscriptions, presented as text transcriptions without metadata or images. The evaluation consisted in three consecutive stages per inscription, conducted via an online Google Form which was programmatically generated and populated for each participant.

In stage 1, experts performed the three epigraphic tasks (textual restoration, geographical and chronological attribution) independently, without AI assistance. In stage 2, they were provided with 10 parallels retrieved by Aeneas from the LED training set of 141,000 inscriptions, and repeated the same tasks on each inscription. In stage 3, experts also received Aeneas' predictions and saliency maps to complete the same epigraphic tasks a final time. All experts completed stage 1, and subsequently for each inscription they were assigned to stage 2 or stage 3 in an alternating sequence. At the end of each stage, participants completed a brief survey to assess their confidence in their predictions for the three tasks and their subjective experience using Aeneas' contextualization aid. In this paired evaluation, during stage 3 two historians analysed the same inscription under different configurations (that is, one with parallels, the other with parallels and predictions). Thus, variations observed in the initial solo evaluation reflect the participants' diverse backgrounds, which ranged from masters students to professors, with a roughly equal split between early career and senior researchers. Participants also differed in their experience of working with inscriptions: while some regularly edit newly discovered texts for publication, others engage primarily in historical analysis of already-published material. This distinction between 'primary' and 'secondary' epigraphic work is important, as it highlights the broader relevance of Aeneas for scholars working with established corpora, where restorations, datings, or provenances may be taken for granted but still warrant critical reassessment.

The evaluation had a maximum time limit of 2 h. To adhere as closely as possible to traditional epigraphic workflows (where scholars consult encyclopaedic resources to find relevant parallels), while acknowledging the artificial constraints of the experimental evaluation, participants were allowed to manually search for parallels using the provided 'Parallel Searching Dataset'. This online spreadsheet, extrapolated from the LED training set, comprised 141,000 texts with associated metadata (place and date of writing), excluding the evaluated inscriptions. Participants were required to note the unique identifiers of all the manually retrieved parallels they used in a designated field within the evaluation form. To ensure impartiality and prevent inadvertent exposure to the evaluated inscriptions, participants were barred from accessing online epigraphic datasets (such as EDR, EDH and EDCS), print editions, search engines or generative AI tools during the evaluation.

Evaluating contextualization. To assess the effectiveness of Aeneas' contextualization mechanism, we counted how many of its suggested parallel inscriptions historians independently incorporated into their manually retrieved list of parallels during stage 2. Historians incorporated an average of 1.5 parallel inscriptions suggested by Aeneas into their own list of parallels (values ranged from 0 to 6; median: 1; interquartile range: 0–2.5).

We further measure the historians' confidence in their predictions across the three stages: it increases by an average of 23% when Aeneas' parallels are provided (restoration from 60.4% to 68.7%, geographical attribution from 46.6% to 57.0%, chronological attribution from 43.7% to 57.5%). Historians' confidence increases by an additional 21% when Aeneas' predictions for the three tasks are also shared (restoration from 53.3% to 75.4%, geographical attribution from 48.7% to 67.0%, chronological attribution from 44.1% to 67.5%).

Finally, we solicited feedback from historians on whether they found that the parallel texts provided by Aeneas served as effective starting points for historical inquiry. When only Aeneas' parallels were provided, 75% of historians agreed (38.3% to a great extent, 36.7% somewhat, 20% very little, 5% not at all). When Aeneas' predictions for the three epigraphic tasks were also included, agreement increased to 90% (45% to a great extent, 45% somewhat, 6.7% very little, 3.3% not at all).

Historians' qualitative feedback. As part of the historian–AI evaluation, we sought qualitative feedback from participants on their subjective experience of using Aeneas in their evaluation. Historians consistently emphasized the value of Aeneas' contextualization mechanism in providing relevant textual and contextual parallels for carrying out the epigraphic tasks on the target inscriptions. A selection is included below:

- "The parallels retrieved by Aeneas completely changed my historical focus. […] it would have taken me a couple of days rather than 15 min [to find these texts]. Were I to base historical interpretations on these inscriptions' readings, now I would have days to write and frame the research questions rather than finding parallels."
- "The help of parallel inscriptions is great for understanding the type of inscription, […] whereas my own search became more narrow."
- "The predictions are very good - as are the preponderance of [parallels for] freed person inscriptions that Aeneas produced. The Statilii Tauri being a prominent family would mean that rabbit holes may be easy to fall down."
- "The help of more parallel inscriptions is great for understanding the type of inscription of fellow soldiers setting up inscriptions, whereas my own search became more narrow on training in on a set of inscriptions from Noricum. [Aeneas offers] a nice parallel tool."
- "The parallels retrieved by Aeneas completely changed my perception of the inscription from stage 1. I did not notice details that made all the difference in both restoring and chronologically attributing the text."

- "Each task was made qualitatively more doable thank to Aeneas' retrieved texts, some of which I had completely missed by solo searching."
- "Aeneas retrieved a very useful parallel (a formula) that I had not found in the dataset."
- "The top parallel [for this inscription] was found independently by both me and Aeneas."
- "[Aeneas shows an] impressive capacity to broaden and, at the same time, refine my [parallel] search results."

Three key themes emerged from the historians' feedback. First, historians highlighted how Aeneas significantly reduced the time required to find relevant parallels, allowing them to focus on deeper historical interpretation and framing research questions. This efficiency also enabled them to explore broader and more refined sets of parallels that traditional historical methods might have missed. Second, they confirmed that Aeneas' retrieved parallels provided valuable insights into the type and context of inscriptions, aiding them in the three epigraphic tasks. Finally, they emphasized Aeneas' ability to broaden searches by identifying significant but previously unnoticed parallels and overlooked textual features, while simultaneously refining results to avoid overly narrow or irrelevant findings.

Some contributors noted challenges with the experimental conditions of the evaluations. First, the imposed time limit, although necessary, acted as a constraint, as historians typically have weeks or months to access materials in standard research settings. Second, the 'Parallel Searching Dataset' online spreadsheet was less easily searchable than specialized corpora (such as *Roman Inscriptions of Britain*[90] and *I.Sicily*[91], which offer refined filtering and cross-searching functionalities for identifying exact textual parallels, as well as a range of additional contextual data regarding form, iconography and archaeological setting). Such artificial limitations were, regrettably, unavoidable due to the constraints inherent in simulating real-world research workflows under experimental conditions. A further observation advanced by some contributors concerned Aeneas' suitability for extremely short, fragmentary, or formulaic inscriptions—particularly those involving abbreviated names—where any guess, whether made by a human expert or an AI model, is inherently risky:

- "None of the parallels really help in this case. The gap precedes a fragmentary *gentilicium* in nominative, so once you restore the nomen, what remains is most likely an abbreviated praenomen. […] It is particularly difficult to use with personal names. Any option would still be very risky."
- "This was an extremely short and vague funerary text, it's impossible to restore with high certainty. It would seem […] that Aeneas retrieves texts which are thematically or stylistically similar to the target text (as one would hope!), however, in the cases of funerary epigraphy, these parallels are as of little use to the epigrapher are manually retrieved parallels! One simply wouldn't use Aeneas for such a text."

On the other hand, Aeneas' ability to retrieve parallels for these short, standardized texts was praised, as it went beyond basic string matching to identify salient formulaic features, even from the limited text available.

- "The retrieved parallels focussed on the formulaic contents of the inscription, not just on the word matching."

In sum, the evaluated historians' qualitative feedback underscores Aeneas' strengths as a research tool: its speed and the historically enriched depth of the parallels it retrieves enables it to not only accelerate research, but also open new avenues of historical inquiry.

Aeneas' limitations. Despite the overall positive feedback from historians, we acknowledge that Aeneas' performance may vary across the entire geographical and chronological scope of the LED dataset. While we see the model's abilities to learn representative patterns for regions and periods, a number of additional factors underlie this variability beyond, for example, changes in language over time and space. To provide a quantitative assessment of Aeneas' limitations and performance variations we conduct an error analysis for geographical and chronological attribution across all provinces and decades using LED's test set. Furthermore, to put that into perspective we plotted the number of inscriptions available for each province and decade in LED's training set. A detailed analysis of these metrics for individual provinces and periods can be found in Extended Data Figs. 3–6.

Explaining this observed variance is challenging, and would serve as a research project by itself. Within the scope of this work, two principal sources can be assumed. The first of these is the availability of data. On the one hand, rates of publication of inscriptions vary from region to region and also by period within regions (due to resources available for study, specific focuses of interest, and so on). On the other hand, even when a region is well published, it does not automatically follow that the data has been systematically incorporated into the existing digital resources (the principal online databases such as EDR have specific geographical focuses, and not all regions are equally well covered). The second is the inherent variability in the cultural practice of inscribing texts in Latin across the Roman Empire in both time and space, meaning that even where a region has been well studied and documented, the quantity of material may well still be very limited compared to other regions. A subsidiary consideration, which may be implied by the variability in performance, is the extent to which that cultural practice actually varies from one region or period to another; but to approach that question would require substantial further work. Assessment of the representativeness of the data remains somewhat impressionistic. Some high-level patterns can however be identified, to illustrate the variation and possible contributing factors.

Perhaps most obviously, we see that Aeneas exhibits the highest performance in chronological attribution around 200 CE. This can be seen to correlate directly to the period for which we have the most inscriptions; this peak in the Latin 'epigraphic habit' has been frequently observed. It can be tentatively argued, however, that this is also the period for which we have the highest number of closely dated inscriptions, meaning that it is not simply the period for which we have the most data, but also the period for which we have the best data. The increase in accuracy for the later third century BCE on the other hand does not correlate so directly to the number of inscriptions. Arguably, this reflects the relatively rapid evolution of the written Latin language in this period, in combination with a relatively rapid increase in the practice of inscribing texts (almost entirely restricted to Italy at this date), such that this is a period to which texts can be dated with some accuracy. By contrast, the earlier texts are both very few in number and traditionally difficult to assign to a narrow window in time.

When considering geographical variation, although there is positive correlation between high availability of texts and high accuracy of attribution (for example, Roma and Africa Proconsularis), two particular sets of variation can perhaps be highlighted. First, several regions of ancient Italy (such as Apulia et Calabria, Aemilia, Etruria and Samnium) offer large numbers of inscriptions, but poor accuracy. A possible explanation for this is presumably the division of Italy into its ancient regions, in contrast to the division of the rest of the dataset into the larger provincial divisions of the Empire, loosely equivalent to modern countries. It is not unlikely that the level of linguistic and cultural variation within ancient Italy is insufficient to permit the model to differentiate so finely; were all the data from the Italian regions to be amalgamated, the accuracy of attribution to 'Italia' would probably be very high. However, the apparent distinctiveness of the city of Rome in comparison to the rest of Italia is notable. Second, and in direct contrast, several more remote parts of the Empire (such as Aegyptus, Cappadocia, Arabia and Cyrenaica), which produce fewer Latin inscriptions (both in terms of data recording and in terms of the original epigraphic production), nonetheless show a higher level of accuracy of

attribution. This can be assumed to reflect greater regional linguistic and cultural distinctiveness in the content of the inscriptions. Finally, two contrasting examples illustrate the underlying problem of data representativeness. Sicily and Sardinia are traditionally associated with a rather weak epigraphic culture (that is, under production), but also are relatively poorly documented in the datasets: this is reflected both in relatively low numbers and particularly poor accuracy. By contrast, Roman Britain is also traditionally described as having a very weak epigraphic culture; however, it is one of the best documented epigraphic traditions in modern studies, and consequently shows a relatively high number of texts; but it also shows a high level of accuracy in the model, suggesting significant regional variation.

Given the space constraints of this Article and the extensive scope of potential analysis, we have limited our discussion here to identifying illustrative high-level patterns of failure cases. Preliminary observations indicate a positive correlation between the number of available inscriptions from a given historical period or Roman province and the model's accuracy in dating or attributing them. However, further investigation is required to disentangle the effects of data availability from other contributing factors, such as the linguistic or epigraphic distinctiveness of certain regions or periods. A more in-depth examination of these nuances and potential mitigation strategies will be the focus of future work.

Modelling epigraphic networks with Aeneas

Parallels, patterns and provincial cult. To showcase the effectiveness of Aeneas' contextualization mechanism for the retrieval of relevant epigraphic parallels, we chose a representative inscription of a well-attested type as a case study. The target inscription (*CIL* XIII, 6665) is an inscribed limestone votive altar from the Roman province of Germania superior, found in the city of Mogontiacum (modern-day Mainz) in 1895 during excavations of a city centre road. The altar can be dated precisely thanks to the internal dating cues: the Ides of July (15 July) of the year of the consulship of Gentianus and Bassus in Rome (211 CE) is explicitly mentioned as the year the altar was dedicated. The inscription records a dedication to the *Deae Aufaniae* (the Aufaniae goddesses) and *Tutela loci* (the local divine patron) by a *beneficiarius consularis* named Lucius Maiorius Cogitatus. *Beneficiarii consulares* were part of the Roman military staff (usually legionaries close to retirement) at the service of provincial governors across the Empire, and are well-attested in the epigraphic evidence of main cities, outposts, frontiers and major communication routes of the Western military provinces[92]. The *beneficiarius* Cogitatus will have been posted at Mogontiacum to assist the provincial governor in administrative, judicial and military duties. It was customary for *beneficiarii* to dedicate a votive altar, such as the one in question: more than 650 such inscriptions are known today, found especially in the provinces on the Rhine and Danube[93–95]. Some of these altars were dedicated to the *Matronae Aufaniae* (as they are more commonly referred to in the epigraphic evidence, the title *Deae Aufaniae* being quite rare), local goddesses whose cult was particularly well-attested in the Rhineland under Roman occupation[96,97].

Aeneas' performance across the three epigraphic tasks for this inscription effectively demonstrates its receptiveness to distinctive geographical, chronological, linguistic, and cultural features (Extended Data Fig. 2). Aeneas' dating average for this altar is 214 CE, which is well within the 10-year range the model is trained on, and its top-3 geographical attributions are Germania superior (correct), Germania inferior and Pannonia superior. Looking at Aeneas' attribution saliency maps, there is a clear focus on the historically specific personal names of two consuls serving that particular year (*Gentiano et Basso cosulibus*) and the worshipped goddesses (*Deab(us) Aufan(iabus)*) whose cult is particularly well-attested in the regions identified by Aeneas. The saliency map of the image of this inscription also shows interesting results, highlighting a focus on the altar's shape, layout and architectural-iconographical elements. This is a sound choice: the *beneficiarii* altars tend to have standardized designs (this particular altar corresponds to type D described by Frenz[98] in *CSIR* De II 4). Finally, wishing to test Aeneas' capacity to restore arbitrary text lengths, we artificially damaged 8 characters ('*loci pro*'): Aeneas' top-5 predictions for this unknown character restoration sequence are all contextually and linguistically accurate, with its first restoration hypothesis (*pro*) being the more commonly attested version of the formula '*Tutelae pro salute*', while the second hypothesis captures the more uncommon version of the formula '*Tutelae loci pro salute*'—the correct restoration for this inscription.

But the story of Cogitatus' altar is far from over. 112 years after the discovery of this altar, 12 similar altars were discovered in 2007 during excavations of the State Chancellery in Mainz, less than 100 m from where the target text was found. The first of these new altars (*FM* 07-055 No. 16 - EDCS-71100087) was published in 2017 (ref. 99), and is included in the LED training data. A complete publication of the 11 other altars was only completed in 2023 (ref. 30), and they do not appear in LED given their recency. This second altar was also dedicated to the *Deae Aufaniae* by the *beneficiarius* Iulius Bellator on the Ides of July in the year of Lateranus and Rufinus' consulship (197 CE). This was the year when emperor Septimius Severus defeated the imperial pretender Clodius Albinus at Lugdunum (Lyon) in a bloody battle, and the phrasing *pro salute et incolumitate sua suorumq(ue) omnium* appearing in this inscription could even be related to Bellator's gratefulness to the goddesses for having survived the battle unscathed[30]. The textual formula is extremely rare, and has only one known parallel in Aeneas' training dataset: Cogitatus' altar from 211 CE. This observation led Haensch[99] in his first edition of the text to note: "*Die Ähnlichkeiten im Formular zur Stiftung des Bellator sind zu groß, um zufällig zu sein*" (the similarities in the formula for Bellator's donation are too great to be coincidental). Haensch believes that the two altars stood together in a ritual space, and that the *beneficiarius* Cogitatus who dedicated *CIL* XIII, 6665 might actually have copied the earlier text by Bellator (both altars belong to the D type figurative design).

Aeneas was also able to identify this crucial parallel for Cogitatus' target text, precisely as an expert historian with knowledge of the archaeological and historical context would have: the first parallel retrieved by Aeneas' contextualization mechanism is Bellator's inscribed altar from 197 CE. Moreover, Aeneas retrieved additional parallels across Germania superior (HD024937, HD017045, HD072700, HD072701 and HD042511), Germania inferior (HD080071) and Pannonia (HD072564, HD033669 and HD051735)—not through exact text-string matches but by recognizing historical, linguistic and epigraphic affinities to the target text. This is a crucial distinction: although historians often use resources such as EDCS_ETL for speedy searches of exact text strings or formulae to yield results, such tools depend on the user formulating precise queries, anticipating what variations might exist. This limits the discovery of related formulae or similar onomastic patterns that fall outside expected templates. Aeneas, by contrast, navigates these constraints by identifying subtle and meaningful historical connections beyond literal matches in ways that mirror expert-level reasoning, despite not having previous knowledge of the archaeological context or spatial relation between inscriptions—features the historian typically uses to guide their interpretation.

These findings underscore the reliability and power of Aeneas as a tool for reconstructing epigraphic networks. Its capacity to emulate and extend historical inquiry highlights its potential as a transformative tool for epigraphic scholarship, producing predictions and associations that consistently align with those a domain expert might draw, despite operating without access to archaeological or spatial context.

Compositional complexities of the *RGDA*. The predictions, parallels, and saliency maps produced by Aeneas for the *RGDA* mirror the complexity of this inscription. While the details of Aeneas' analysis have been addressed above (and integrated by Extended Data Table 2, which summarizes the main historical features identified as

chronologically specific through Aeneas' saliency maps), the debates around the *RGDA* will now be expounded to add nuance and background to Aeneas' analysis.

Whereas the version of the text inscribed at Ancyra (known as the *Monumentum Ancyranum*) was created in c.19 CE[29], the text itself of the *RGDA* was first 'published' when read out to the Senate in 14 CE, shortly after Augustus' death, and was then inscribed outside his Mausoleum in Rome. Scholars debate, however, when the *RGDA* was composed, whether it was drafted as an evolving document in various stages throughout Augustus' reign (starting as early as 2 BCE) or as a unified, retrospective account of his achievements composed near the end of his life (13–14 CE). In contrast to other inscriptions, therefore, the date of inscribing in this case is different from the date of composition.

At the end of the *RGDA* (35.2), Augustus concludes with the statement "When I wrote this I was in my seventy-sixth year" (*cum scripsi haec, annum agebam septuagensumum sextum*). This 'seventy-sixth year' should be understood as the last year of his life, that is, the period between the celebration of his birthday on 23 September 13 CE to his death on 19 August 14 CE. Despite this clear statement, most scholars have assumed that this is a misleading statement, supporting the view that this only refers to a moment of final revision of the text rather than to its complete composition. Some have also argued[100,101] that Tiberius carried out final revisions to the text after Augustus' death, updating information relating to the years 13 CE and 14 CE, such as adding into chapter 4.4 the thirty-seventh grant of tribunician power to Augustus on 26 June 14 CE. Various proposals have also been put forward in support of the idea that it is possible to trace compositional layers in the text, with the most popular argument being that the *RGDA* was substantially completed by 2 BCE, with other drafts and emendations perhaps occurring in 23 BCE, 12 BCE, 4 BCE, 1 CE, 6 CE and 14 CE. This debate about how to identify compositional layers in the *RGDA* was summarized by Kornemann[102] and further evaluated by Gagé[103]. The idea remained influential in the recent commentary by Scheid[104], who also argues against the idea of Augustus as the author of the text in any meaningful sense. Both Ramage[105] and Cooley[29], however, have proposed a more straightforward approach that sees Augustus as composing the whole text during the last year of his life, and so taking at face value what he writes at the end of his account, cited above. In particular, it is suggested that the text was essentially put together during the summer of 14 CE, perhaps even between 26 June and his final departure from Rome on 24 July.

The question is not just of 'academic interest', as Augustus' approach to composing his account of his life's achievements has implications for his understanding of his contribution to Roman history. Did he feel, by 2 BCE, that he had essentially achieved the pinnacle of his career, at the moment when he was hailed as 'father of the fatherland', and was this something that prompted him to compose his *RGDA*? Or was it only in 14 CE that he felt the need to compose a partisan account of his lifetime's achievements, justifying his position in Roman society and wishing to influence the way in which he was to be remembered by future generations[106]?

These long-standing debates about the composition of the *RGDA* highlight the interpretive complexities that Aeneas was designed to engage with, as shown in the Grounding the *Res Gestae Divi Augusti* section. This case study demonstrates how Aeneas can support historical workflows by testing existing hypotheses against linguistic patterns in the text, and by complementing expert-led interpretation with quantitative historical analysis.

Teaching with Aeneas in the classroom

To maximize Aeneas' impact, we partnered with the teacher training programme at the University of Ghent and Sint-Lievenscollege Ghent to co-design a course for educators and high school students. Building on previous work on the Ithaca model for ancient Greek inscriptions[107] (https://www.robbewulgaert.be/education/ithaca-teaching-history-journal, AI & Greek Ithaca—syllabus), which was recognized by the *Teaching History Journal* and twice honoured at the European AI for Education Awards, this new curriculum bridges AI and ancient history, with a pedagogical focus that centres Aeneas in the learning process (https://www.robbewulgaert.be/education/predicting-the-past-aeneas, AI & Latin Aeneas—syllabus). The course shifts the focus to Latin inscriptions and their contextualization, allowing students to engage directly with primary sources in classical studies while exploring novel AI methods. Currently, incorporated into the in-service teacher training programme at the University of Antwerp, it showcases practical applications of AI in the humanities, while promoting digital literacy.

This curriculum aligns with the European Union's *Digital Competence Framework for Citizens* (DigComp) and UNESCO's *AI Competency Framework for Students*, addressing key competencies such as critically evaluating AI-generated outputs, adopting a human-centred approach, and applying human oversight in interdisciplinary contexts.

Future directions

Aeneas demonstrates the transformative potential of AI in augmenting historical research, yet several avenues offer promising prospects for future development. One key direction involves integrating Aeneas's capabilities into large-scale dialogue models. This could enable more natural and interactive research workflows, allowing historians to query the system, probe the model's answers, and receive better explanations. Addressing the inherent uncertainty in historical data, particularly concerning chronological attribution, remains a critical challenge. Future work could focus on developing better methods for representing and evaluating wide dating brackets, both within the model's architecture and through refined evaluation metrics that better capture the nuance of historical dating practices beyond distance from estimated ranges. A further opportunity lies in conducting additional ablation studies to quantify the contribution of different components (such as the impact of visual inputs on different tasks); as well as exploring how contextual parallels change with different textual inputs and how sensitive the system is to variations in input formatting (and across different types of inscriptions). Improving the multimodal capabilities with larger, highly standardized and FAIR datasets (those adhering with FAIR principles; https://www.go-fair.org/fair-principles/), while broadening the scope beyond Latin inscriptions, are also rewarding research directions. This would allow a deeper exploration of the visual modality's potential beyond geographical attribution, potentially informing chronological dating through iconographic or otherwise archaeologically informed analysis. Finally, we believe that deepening interdisciplinary collaborations is paramount: we hope that future projects continue to build along the path of bridging the humanities and the sciences.

Ethics and inclusion statement

This study was developed through a collaborative, interdisciplinary approach, bringing together ancient historians, computer scientists and educational experts, ensuring diverse perspectives throughout the research process. Capacity-building was central to this effort, enabling effective communication between disciplines and the exploration of meaningful research questions, while leveraging state-of-the-art technology to advance our understanding of ancient history. Central to this project was the recognition that epigraphy serves as a key source of direct evidence for understanding a wide range of social groups in the ancient world, including not only emperors and élites, but also marginalized and subaltern groups such as enslaved individuals, women, and other voiceless communities. This focus on the diversity of ancient social identities highlights the crucial role of epigraphic data in challenging dominant narratives and fostering a more inclusive understanding of the Roman world.

Although our methods hold great promise in advancing historical research, we are mindful of the risks associated with the misuse of AI.

The potential for misclassification or misrepresentation of historical data is a notable concern, particularly in the Roman world, where AI models could inadvertently reinforce biased or inaccurate readings of the past. It is essential that AI tools are deployed with human oversight, as blind reliance on the comprehensiveness of automated methods risks distorting historical interpretations. We also emphasize that AI should complement rather than substitute human expertise in the humanities. Our approach aims to alleviate the enormous effort and time-consuming nature of processing and analysing large datasets, allowing historians to focus on the critical interpretation and contextual analysis of ancient texts. This collaboration between AI and human scholarship is crucial for advancing responsible and ethical practices in digital humanities research. Finally, we are committed to promoting the responsible use of AI in humanities research, with an emphasis on ensuring that AI tools are used thoughtfully and transparently. By integrating interdisciplinary expertise, we aim to foster a more responsible and inclusive approach to the application of AI in the study of ancient history.

Ethics approval statement

Ethical approval for the historian–AI evaluation protocol was granted by the ethics board of the School of Humanities, University of Nottingham. The protocol adhered to rigorous ethical standards and received a favourable ethical opinion from the Faculty of Arts Research Ethics Committee at the University of Nottingham. In accordance with the approved protocol, all participants were provided with a participant information sheet, a participant consent form and a General Data Protection Regulation (GDPR) privacy notice. A comprehensive data management plan was developed, and an awareness of ethical behaviour for data collection form was completed. Evaluation designers also completed two mandatory online courses—Research Integrity and Human Subjects Protections—offered by the University of Nottingham's Researcher Academy, ensuring alignment with the highest standards of research ethics and integrity. To ensure inclusivity, the 23 expert researchers involved in the human evaluations reflected gender diversity (11 male and 12 female) and career-stage representation (early career researchers working alongside full professors).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The LED was developed by processing and integrating data from the EDR, EDH and EDCS_ETL databases, all of which are openly accessible under a Creative Commons Attribution 4.0 license. The original data sources used in this work can be accessed directly at their respective repositories: EDR (https://zenodo.org/records/3575495 (ref. 108)), EDH (https://zenodo.org/records/3575155 (ref. 109)), and EDCS_ETL (https://zenodo.org/records/7072337 (ref. 110)). The processed LED dataset is publicly available online at https://github.com/google-deepmind/predictingthepast. The list of Roman personal names used for evaluating the onomastics baseline was derived from the Classical Language Toolkit (https://cltk.org/), is available under the MIT License, and the processed proper names are available at https://github.com/google-deepmind/predictingthepast. The new curriculum co-designed as part of this research's wider knowledge exchange and impact strategy is freely available online at https://predictingthepast.com.

Code availability

The Aeneas source code is available at https://github.com/google-deepmind/predictingthepast under Apache License 2.0, along with the trained weights, licensed under a Creative Commons Attribution-ShareAlike 4.0 International license (CC BY-SA 4.0). A public interface for historians using Aeneas for their research (that is, contextualization, restoration and attribution of Latin inscriptions and use of all visualization tools discussed in the Article) is available at https://predictingthepast.com. Neural networks were developed with JAX v.0.4.37 (https://github.com/jax-ml/jax) and Flax v.0.10.2 (https://github.com/google/flax). The XLA compiler is bundled with JAX and does not have a separate version number. Dataset processing and analysis used Python v.3.9 (https://www.python.org), NumPy v.2.1.3 (https://github.com/numpy/numpy), SciPy v.1.13.1 (https://www.scipy.org), pandas v.2.3.3 (https://github.com/pandas-dev/pandas), beautifulsoup4 v.4.12.3 (https://www.crummy.com/software/BeautifulSoup) and Google Colab (https://research.google.com/colaboratory). Visualizations were generated using matplotlib v.3.10.0 (https://matplotlib.org), plotly v.5.24.1 (https://plotly.com/python), seaborn v.0.13.2 (https://seaborn.pydata.org) and GeoPandas v.1.0.1 (https://geopandas.org) and CartoDB basemaps (https://github.com/CartoDB/CartoDB-basemaps) for the map tiles.

31. Vaswani, A. et al. in *Advances in Neural Information Processing Systems 30* (eds Guyon, I. et al.) 5998–6008 (Curran Associates, 2017).
32. Pavlopoulos, J. et al. (eds). *Proc. 1st Workshop on Machine Learning for Ancient Languages* (ACL, 2024).
33. Bhurke, S. S., Lomte, V. M., Kolhe, P. M. & Pednekar, A. U. Survey on Sanskrit script recognition. In *International Conference on Mobile Computing and Sustainable Informatics* 771–782 (Springer, 2020).
34. Faigenbaum-Golovin, S., Shaus, A. & Sober, B. Computational handwriting analysis of ancient Hebrew inscriptions—a survey. *IEEE BITS* **2**, 90–101 (2022).
35. Tyndall, S. Toward automatically assembling Hittite-language cuneiform tablet fragments into larger texts. In *Proc. 50th Annual Meeting of the Association for Computational Linguistics* (eds Li, H. et al.) 243–247 (2012).
36. Lazar, K., Saret, B., Yehudai, A., Horowitz, W., Wasserman, N. & Stanovsky, G. Filling the gaps in ancient Akkadian texts: a masked language modelling approach. In *Proc. of Empirical Methods in Natural Language Processing* 4682–4691 (ACL, 2021).
37. Papavassileiou, K., Kosmopoulos, D. I. & Owens, G. A generative model for the Mycenaean Linear B script and its application in infilling text from ancient tablets. *ACM J. Comput. Cult. He.* https://doi.org/10.1145/3593431 (2023).
38. Faigenbaum-Golovin, S. et al. Algorithmic handwriting analysis of Judah's military correspondence sheds light on composition of biblical texts. *Proc. Natl Acad. Sci. USA* **113**, 4664–4669 (2016).
39. Yu, T. et al. Artificial intelligence for Dunhuang cultural heritage protection: The project and the dataset. *Int. J. Comput. Vision* **130**, 2646–2673 (2022).
40. Rao, R. P. et al. A Markov model of the Indus script. *Proc. Natl Acad. Sci. USA* **106**, 13685–13690 (2009).
41. Yadav, N. et al. Statistical analysis of the Indus script using n-grams. *PLoS ONE* **5**, e9506 (2010).
42. Sproat, R. A statistical comparison of written language and nonlinguistic symbol systems. *Language* **90**, 457–481 (2014).
43. Nguyen, T.-N., Burie, J.-C., Le, T.-L. & Schweyer, A.-V. On the use of attention in deep learning based denoising method for ancient Cham inscription images. In *Proc. 16th International Conference on Document Analysis and Recognition* (eds Lladós, J., Lopresti, D. & Uchida, S.) 400–415 (ACM, 2021).
44. Zhang, C. et al. Data-driven Oracle Bone rejoining: a dataset and practical self-supervised learning scheme. In *Proc. 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 4482–4492 (ACM, 2022).
45. Pirrone, A., Aimar, M. B. & Journet, N. Papy-S-Net: A siamese network to match papyrus fragments. In *Proc. 5th International Workshop on Historical Document Imaging and Processing* 78–83 (ACM, 2019).
46. Abitbol, R., Shimshoni, I. & Ben-Dov, J. Machine learning based assembly of fragments of ancient papyrus. *J. Comput. Cult. He.* **14**, 1–21 (2021).
47. Parker, C. S. et al. From invisibility to readability: recovering the ink of Herculaneum. *PLoS ONE* **14**, e0215775 (2019).
48. Swindall, M. I. et al. Dataset augmentation in papyrology with generative models: a study of synthetic ancient Greek character images. In *Proc. 31st International Joint Conference on Artificial Intelligence* 4973–4979 (IJCAI, 2022).
49. Kang, K. et al. Restoring and mining the records of the Joseon dynasty via neural language modeling and machine translation. In *Proc. 2021 Conference North American Chapter of the Association for Computational Linguistics* (eds Toutanova, K. et al.) 4031–4042 (ACL, 2021).
50. Zhao, H., Chu, H., Zhang, Y. & Jia, Y. Improvement of ancient Shui character recognition model based on convolutional neural network. *IEEE Access* **8**, 33080–33087 (2020).
51. Dinesh, P., Sruthi, A. L., Praveen, S. & Manjunathan, A. Word prediction using CNN for ancient manuscripts. *AIP Conf. Proc.* https://doi.org/10.1063/5.0119383 (2023).
52. Duan, S., Wang, J. & Su, Q. Restoring ancient ideograph: a multimodal multitask neural network approach. In *Proc. 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (eds Calzolari, N. et al.) 14005–14015 (ACL, 2024).
53. Cowen-Breen, C., Brooks, C., Haubold, J. & Graziosi, B. Logion: machine learning based detection and correction of textual errors in Greek philology. In *Proc. of the Ancient Language Processing Workshop* 170–178 (ACL, 2023).

54. Soumya, A. & Kumar, G. H. Classification of ancient epigraphs into different periods using random forests. In *International Conference on Signal and Image Processing* 171–178 (ACM, 2014).
55. Adam, K., Baig, A., Al-Maadeed, S., Bouridane, A. & El-Menshawy, S. Kertas: dataset for automatic dating of ancient Arabic manuscripts. *Int. J. Doc. Anal. Recognit.* **21**, 283–290 (2018).
56. Goler, S. et al. Dating ancient Egyptian papyri through Raman spectroscopy: concept and application to the fragments of the Gospel of Jesus' wife and the Gospel of John. *J. Study New Testam.* **42**, 98–133 (2019).
57. Yu, X. & Huangfu, W. A machine learning model for the dating of ancient Chinese texts. In *International Conference on Asian Language Processing* 115–120 (IEEE, 2019).
58. Bogacz, B. & Mara, H. Period classification of 3D cuneiform tablets with geometric neural networks. In *17th International Conference on Frontiers in Handwriting Recognition* 246–251 (2020).
59. Chang, X., Chao, F., Shang, C. & Shen, Q. Sundial-GAN: a cascade generative adversarial networks framework for deciphering Oracle Bone inscriptions. In *ACM International Conference on Multimedia* 1195–1203 (ACM, 2022).
60. Yoo, H. et al. HUE: pretrained model and dataset for understanding Hanja documents of ancient Korea. In *North American Chapter of the Association for Computational Linguistics* 1832–1844 (ACL, 2022).
61. Paparigopoulou, A., Pavlopoulos, J. & Konstantinidou, M. Dating Greek papyri images with machine learning. In *ICDAR Workshop on Computational Paleography* https://doi.org/10.21203/rs.3.rs-2272076/v1 (2023).
62. Yamshchikov, I. P., Tikhonov, A., Pantis, Y., Schubert, C. & Jost, J. BERT in Plutarch's shadows. In *Proc. 2022 Conference Empirical Methods in Natural Language Processing* 6071–6080 (ACL, 2022).
63. Greenhalgh, M. *Marble Past, Monumental Present: Building with Antiquities in the Mediaeval Mediterranean* (Brill, 2009).
64. Burns, P. J., Brofos, J., Li, K., Chaudhuri, P. & Dexter, J. P. Profiling of intertextuality in Latin literature using word embeddings. In *Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics* (eds, Toutanova, K. et al.) 4900–4907 (ACL, 2021).
65. Sprugnoli, R., Passarotti, M., Flavio Massimiliano, C., Fantoli, M. & Moretti, G. Overview of the EvaLatin 2022 evaluation campaign. In *Proc. 2nd Workshop on Language Technologies for Historical and Ancient Languages* (eds Sprugnoli, R. et al.) 183–188 (ACL, 2022).
66. Yousef, T., Palladino, C., Wright, D. J. & Berti, M. Automatic translation alignment for ancient Greek and Latin. In *Proc. 2nd Workshop on Language Technologies for Historical and Ancient Languages* (eds Sprugnoli, R. et al.) 101–107 (ACL, 2022).
67. Martins, A. et al. Historia Augusta authorship: an approach based on measurements of complex networks. *Appl. Netw. Sci.* **6**, 50 (2021).
68. Corbara, S., Moreo, A. & Sebastiani, F. Syllabic quantity patterns as rhythmic features for Latin authorship attribution. *J. Assoc. Inf. Sci. Technol.* **74**, 128–141 (2022).
69. Bamman, D. & Burns, P. J. Latin BERT: a contextual language model for classical philology. Preprint at https://doi.org/10.48550/arXiv.2009.10053 (2020).
70. Bodel, J. in *Latin Epigraphy and the IT Revolution* (eds Davies, J. & Wilkes, J.) 275–296 (British Academy, 2012).
71. Elliott, T. in *Epigraphy and Digital Resources* (eds Bruun, C. & Edmondson, J.) 78–85 (Oxford Univ. Press, 2014).
72. Espinosa Espinosa, D. & Velázquez Soriano, I. (eds). *Epigraphy in the Digital Age: Opportunities and Challenges in the Recording, Analysis and Dissemination of Inscriptions* (Archaeopress, 2021).
73. Bodel, J., Prag, J. R. W. & Roueché, C. in *L'Épigraphie au XXIe Siècle. Actes du XVIe Congrès International d'Épigraphie Grecque et Latine* (eds Fröhlich, P. & Cabellero, M. N.) 91–117 (Ausonius, 2024).
74. Lavan, M. The army and the spread of Roman citizenship. *J. Rom. Stud.* **109**, 27–69 (2019).
75. Prag, J. R. W. in *Becoming Roman, Writing Latin* (ed. Cooley, A.) 15–31 (Cambridge Univ. Press, 2002).
76. Varga, R. A quantitative approach on Latin occupational epigraphy. *J. Anc. Hist. Archaeol.* **6**, 78–109 (2019).
77. Kaše, V., Heřmánková, P. & Sobotková, A. Division of labor, specialization and diversity in the ancient Roman cities: a quantitative approach to Latin epigraphy. *PLoS ONE* **17**, e0269869 (2022).
78. Broder, A. Z. On the resemblance and containment of documents. In *Proc. Compression and Complexity of Sequences* 21–29 (IEEE, 1997).
79. Lassère, J.-M. *Manuel d'Épigraphie Romaine* (Picard, 2007).
80. Badian, E. History from "square brackets". *Z. Papyr. Epigr.* **79**, 59–70 (1989).
81. You, Y. et al. Large batch optimization for deep learning: training BERT in 76 minutes. In *International Conference on Learning Representations* (ICLR, 2020).
82. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
83. Szczepankiewicz, K. et al. Ground truth based comparison of saliency maps algorithms. *Sci. Rep.* **13**, 16887 (2023).
84. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).
85. Beltrán Lloris, F. in *The Epigraphic Habit in the Roman World* (eds Bruun, C. & Edmonson, J.) 131–148 (Oxford Univ. Press, 2015).
86. Haeussler, R. in Signes de la "Romanisation" à Travers l'Épigraphie: Possibilités d'Interprétations et Problèmes Méthodologiques (ed. Haeussler, R.) 9–30. (Éditions Monique Mergoil, 2001).
87. Corbier, M. L'urbs: espace urbain et histoire (Ier siècle av. J.-C.–IIIe siècle ap. J.-C.). In *Actes du Colloque International de Rome (8–12 Mai 1985)* 27–60 (1987).
88. Meyer, E. Explaining the epigraphic habit in the Roman Empire: the evidence of epitaphs. *J. Rom. Stud.* **80**, 74–96 (1990).
89. Salway, B. What's in a name? A survey of Roman onomastic practice from c. 700 B.C. to A.D. 700. *J. Rom. Stud.* **84**, 124–145 (1994).
90. Vanderbilt, S. Roman Inscriptions of Britain. https://romaninscriptionsofbritain.org/ (accessed 1 December 2024).
91. Prag, J. R. W. Sicily: Inscriptions of Ancient Sicily. http://sicily.classics.ox.ac.uk (accessed 1 December 2024).
92. Rankov, N. B. *The Beneficiarii Consularis in the Western Provinces of the Roman Empire*. D.Phil thesis, Univ. of Oxford (1987).
93. Nelis-Clément, J. *Les Beneficiarii: Militaires et Administrateurs au Service de l'Empire (Ier S. a.C.-VIe S. p.C)* (Ausonius, 2000).
94. Haensch, R. Beneficiarii consularis in Sirmium. *Chiron* **24**, 345–404 (1994).
95. Steidl, B. Die Station der beneficiarii consularis in Obernburg am Main: vorbericht über die Ausgrabungen 2000/2002. *Germania* **83**, 67–94 (2005).
96. Spickermann, W. *Germania Superior. Religion der Römischen Germanien I (Religion der Römischen Provinzen)* (Mohr Siebeck, 2003).
97. Garman, A. G. *The Cult of the Matronae in the Roman Rhineland: an Historical Evaluation of the Archaeological Evidence* (Edwin Mellen Press, 2008).
98. Frenz, H. G. *Denkmäler Römischen Götterkultes aus Mainz und Umgebung (Corpus Signorum Imperii Romani Deutschland II 4: Germania Superior)* (Habelt, 1992).
99. Haensch, R. Steine sind. Benefiziarier, Matronae Aufaniae und die Topographie des Römischen Mainz. *Z. Papyr. Epigr.* **203**, 293–296 (2017).
100. Mommsen, T. in *Der Rechenschaftsbericht des Augustus. Gesammelte Schriften IV* 247–258 (1906).
101. Brunt, P. A. & Moore, J. M. *Res Gestae Divi Augusti. The Achievements of the Divine Augustus* (Oxford Univ. Press, 1967).
102. Kornemann, E. in *Monumentum Ancyranum* (ed. Wissowa, G.) 211–231 (J. B. Metzler Verlag, 1933).
103. Gagé, J. *Res Gestae Divi Augusti: Ex Monumentis Ancyrano et Antiocheno Latinis, Ancyrano et Apolloniensi Graecis* (Les Belles Lettres, 1935).
104. Scheid, J. *Res Gestae Divi Augusti. Hauts Faits du Divin Auguste* (Les Belles Lettres, 2007).
105. Ramage, E. S. The date of Augustus' Res Gestae. *Chiron* **18**, 71–82 (1988).
106. Horster, M. *Bauinschriften Römischer Kaiser. Untersuchungen zu Inschriftenpraxis und Bautätigkeit in Städten des Westlichen Imperium Romanum in der Zeit des Prinzipats* (Franz Steiner Verlag, 2001).
107. Wulgaert, R. Ithaca AI meets ancient Greek: Muses and robots in the classroom. *Teach. Hist.* **57**, 16–20 (2023).
108. Panciera, S. et al. EDR—Epigraphic Database Roma EpiDoc files [Dataset]. *Zenodo* https://doi.org/10.5281/zenodo.3575495 (2019).
109. Cowey, J. M. S. et al. Epigraphic Database Heidelberg EpiDoc files [Dataset]. *Zenodo* https://doi.org/10.5281/zenodo.3575155 (2019).
110. Heřmánková, P. EDCS (2.0) [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.7072337 (2022).

Acknowledgements We acknowledge N. de Freitas, S. Reed, M. Hoffman, K. McKee, I. Androutsopoulos, Z. Ghahramani for their scientific insights; A. Senoner for supporting the outreach of this project; M. Beard and G. Woolf for their endorsements; the Epigraphic Database Roma, Epigraphic Database Heidelberg, EDCS_ETL, Epigraphik-Datenbank Clauss Slaby and Trismegistos for their invaluable resources; the GDKE-Landesmuseum Mainz for permission to use the photograph of inv. no. S 553 (*CIL* XIII, 6665); M. C. Georgiadou and M. Puddu for their encouragement. We thank the expert historian evaluators: A. Meyer, B. Kolbeck, B. Benedetta, C. Cenati, F. Ugolini, F. Lentini, F. Feraudi-Gruénais, K. Sekita, L. Calvelli, M. J. E. Tolosa, M. Puddu, P. Christoforou, P. Heřmánková, S. Vanderbilt, S. España-Chamorro, S. Orlandi, S. Zoumbaki, T. Tommasi, V. Olivero and the participants who preferred to remain anonymous. T.S. acknowledges that this project has received funding from the Leverhulme Trust.

Author contributions Y.A. and T.S. co-led the project, jointly overseeing its design, execution and overall coordination, and wrote the article together: they are co-first authors and the order of the names is alphabetical. A.C. supported the extended analysis of the *Res Gestae Divi Augusti*. J. Pa. and B.H. contributed to the human evaluations. B.S., B.M., J.G. and N.D undertook the design of the online public interface and the open-sourcing process. P.S. contributed to data analysis. R.W. in collaboration with the University of Ghent (UGent) and Sint-Lievenscollege Ghent (K12) co-designed a course tailored for educators and high school students. A.M., J. Pr. and S.M. served in an advisory capacity during the project's development.

Competing interests The authors declare no competing interests.

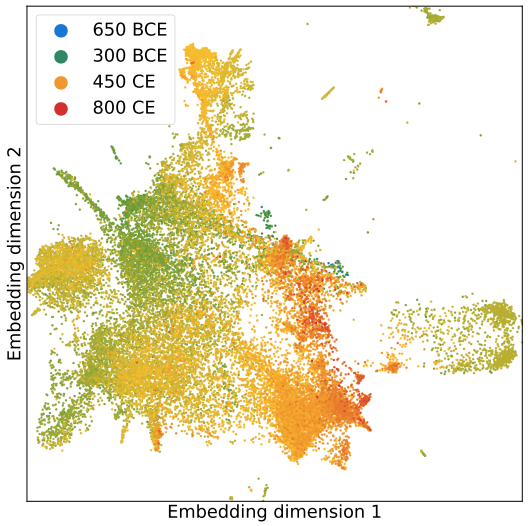
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-025-09292-5.

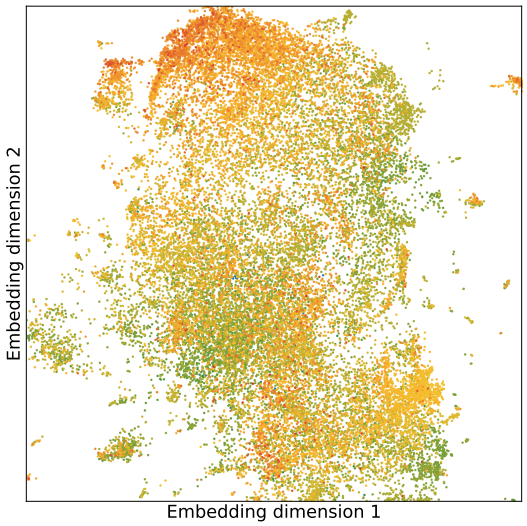
Correspondence and requests for materials should be addressed to Yannis Assael or Thea Sommerschield.

Peer review information *Nature* thanks John Bodel, Shai Gordin, Myles Lavan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

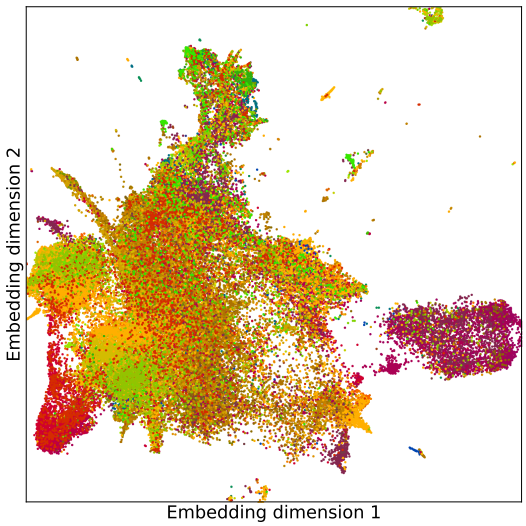
Reprints and permissions information is available at http://www.nature.com/reprints.



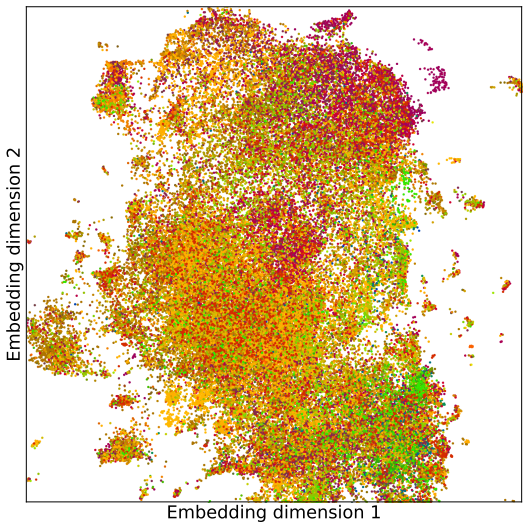
(a) Chronological attribution - Aeneas



(b) Chronological attribution - T5



(c) Geographical attribution - Aeneas



(d) Geographical attribution - T5

Extended Data Fig. 1 | Aeneas vs. T5 embeddings using UMAP. UMAP visualisation illustrating the chronological and geographical attribution of Aeneas' historically rich embeddings in comparison to traditional T5 textual embeddings. Geographical labels were excluded due to their length; instead, a colour gradient was employed to encode geographical coordinates: yellow representing northern locations, red for western, green for eastern, and blue for southern Roman provinces. (a) Chronological Attribution – Aeneas; (b) Chronological Attribution - T5; (c) Geographical Attribution – Aeneas; (d) Geographical Attribution - T5.

Article

Inscription: HD054789 (CIL 13, 6665)

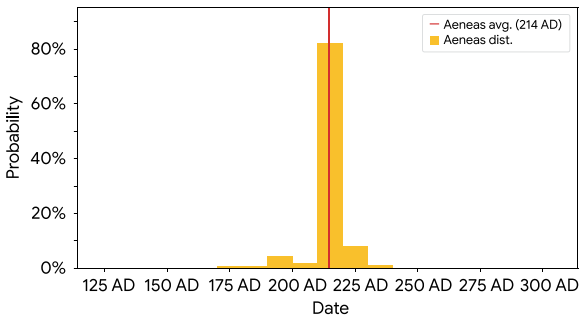
Date: 15 July 211 CE
Province: Germania superior, Mogontiacum (Mainz)

Text: Deab(us) Aufan(iabus) / et Tutelae loci / pro salute et in/col(u)mitate sua / suorumq(ue) om/nium L(ucius) Maiori/us Cogitatus b(ene)f(iciarius) / co(n)s(ularis) vot(um) sol(vit) l(ibens) l(aetus) m(erito) / Idibus Iuli(i)s / Gentiano et / Basso co(n)s(ulibus)

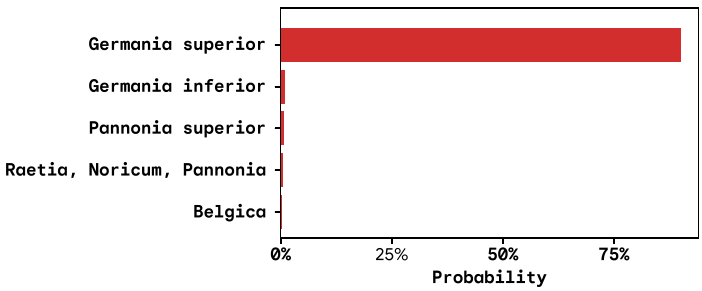
(a) Aeneas contextual parallels

1: EDCS-71100087
2: HD024937
3: HD017045
4: HD072700
5: HD042511
6: HD072701
7: HD080071
8: HD072564
9: HD033669
10: HD051735

(b) Chronological attribution



(c) Geographical attribution



(d) Geographical attribution visual saliency map

(e) Geographical attribution textual saliency map

deab aufan et tutelae loci pro salute et incolmitate sua suorumq omnium l maiorius cogitatus
bficiarius cosularis vot sol l l m idibus iulis gentiano et basso cosulibus

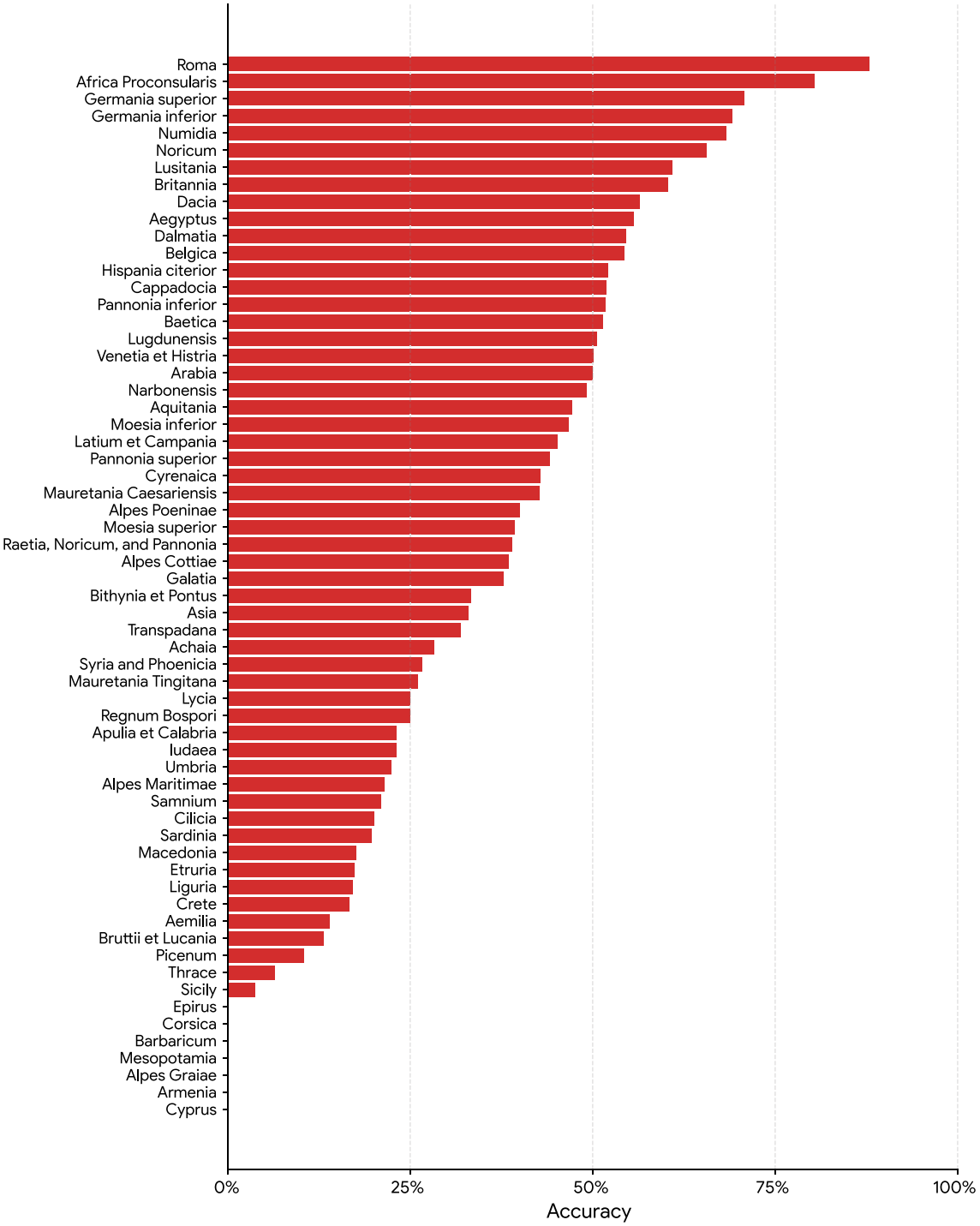
(f) Restoration - unknown length

deab aufan et tutelae ___#___ salute et incolmitate sua suorumq omnium l maiorius cogitatus
bficiarius cosularis vot sol l l m idibus iulis gentiano et basso cosulibus

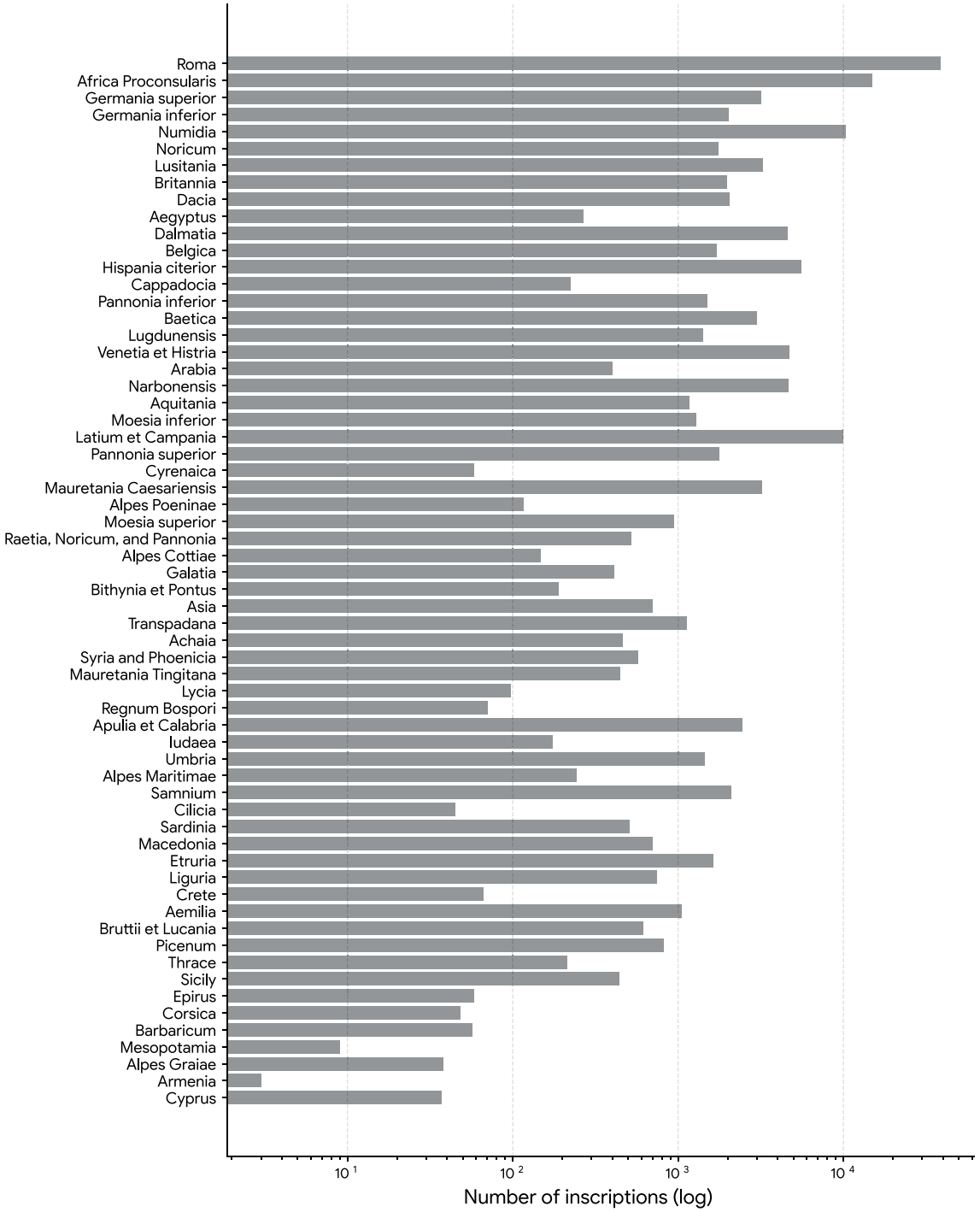
Aeneas hypotheses

1) "pro" 2) "loci pro" 3) "aug pro" 4) "sacr pro" 5) "sacrum" 6) ...

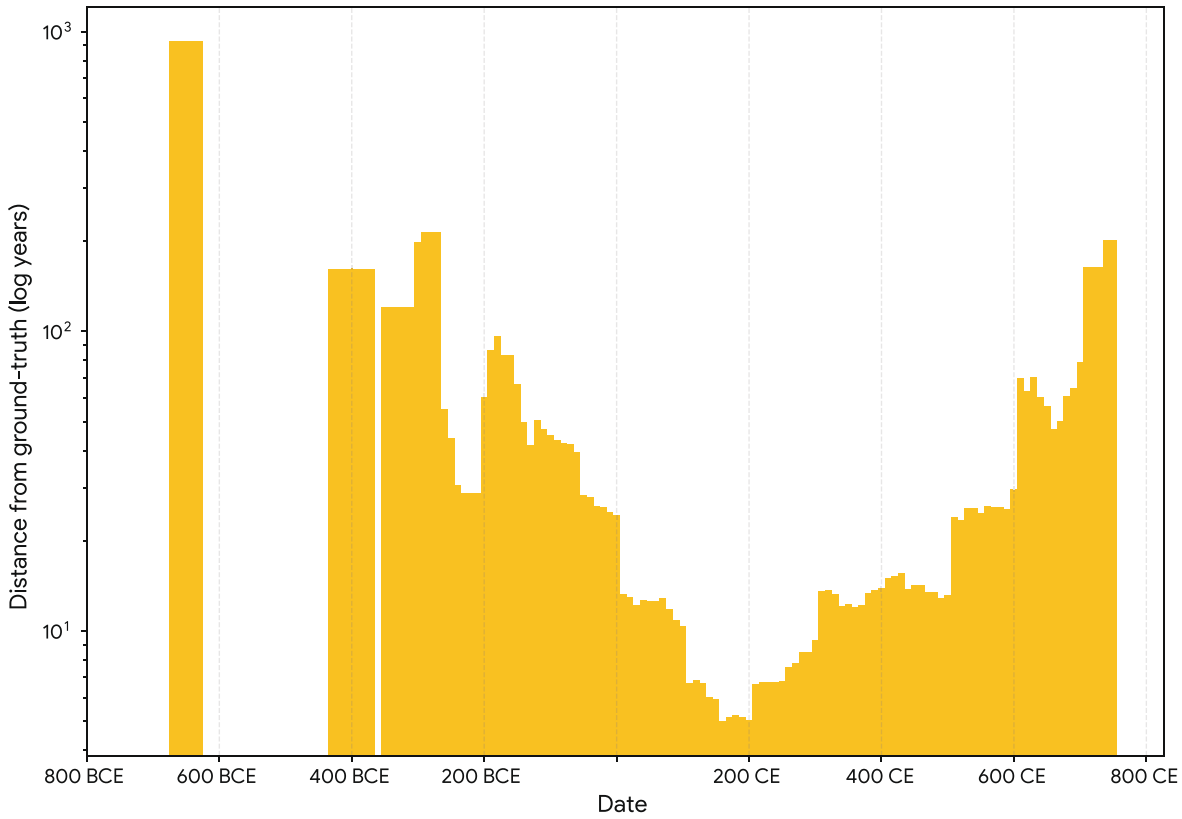
Extended Data Fig. 2 | Aeneas' outputs for contextualising the altar *CIL* XIII 6665 (TM 211813 = HD54789), a votive altar from Mainz (ancient Mongontiacum, in the province of Germania superior), dating 15 July 211 CE. For this inscription, Aeneas provides: (a) the retrieved textual and contextual parallels; (b) the chronological attribution predictions; (c) the geographical attribution predictions; (d) the image saliency map for geographical attribution; (e) the textual saliency map for geographical attribution; (f) the restoration hypotheses for a lacuna of unknown length. Photograph of inv. no. S 553, courtesy of GDKE-Landesmuseum Mainz (ph. Ursula Rudischer). Map reproduced from CartoCB basemaps under a CC BY 3.0 Attribution 3.0 Unported license.



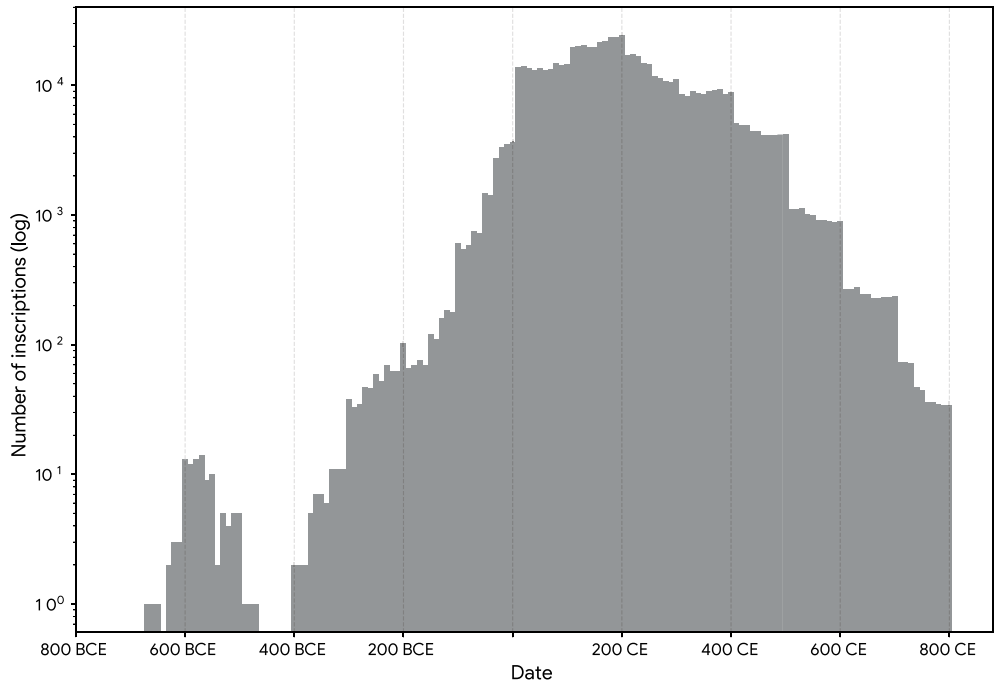
Extended Data Fig. 3 | Geographical attribution performance analysis (LED testing set). Geographical attribution accuracy per Roman province (LED test set). Some provinces may be empty as no inscriptions exist in the test set.



Extended Data Fig. 4 | Geographical attribution inscriptions per province (LED training set). Geographical attribution inscriptions per province (LED training set).



Extended Data Fig. 5 | Chronological attribution performance analysis (LED test set). Chronological attribution date loss per decade (LED test set). Some decades may be empty as no inscriptions exist in the test set.



Extended Data Fig. 6 | Chronological attribution inscriptions per province (LED training set). Chronological attribution inscriptions per decade (LED training set).

Extended Data Table 1 | LED dataset statistics

	LED *(our dataset)*			EDR			EDH			EDCS_ETL	
	Text	Char.	Img.	Text	Char.	Img.	Text	Char.	Img.	Text	Char.
Train	141k	12,900k	6,524	26k	3,258k	585	40k	4,188k	5k	74k	5,454k
Valid	17k	1,560k	778	3k	361k	71	5k	524k	734	9k	674k
Test	17k	1,571k	810	3k	384k	63	5k	501k	747	9k	684k

Extended Data Table 2 | Aeneas' *RGDA* analysis

RGDA archaising form	Standard form
impensa	*inpensa*
terra et mari	*terra marique*
aheneis	*aeneis*
apsenti	*absenti*
conlega	*collega*
claussum	*clausum*
caussa	*causa*
plebei	*plebi*
sexsiens	*sexiens*
emeriteis	*emeritis*
argenteis	*argentis*
adsignavi	*assignavi*
conlegio	*collegio*
Dalmateis	*Dalmatis*
quadrigeis	*quadrigis*

Historical feature	Contextual reference
Triumviri rei publicae constituendae (Ch. 1)	Political alliance ruling from 43–33 BCE
Augustalia (Ch. 11)	Religious festival instituted in the Augustan era
Ara Pacis Augustae (Ch. 12)	Monument commissioned in 13 BCE
Princeps iuventutis (Ch. 14)	Title assumed in 5 BCE
Aerarium militare (Ch. 17)	Treasury established in 6 CE
Forum Augustum (Ch. 21)	Forum dedicated in 2 BCE

Summary of the main salient historical features and their chronological reference identified by Aeneas throughout the *RGDA*.

nature portfolio

Corresponding author(s): Yannis Assael, Thea Sommerschield

Last updated by author(s): 29/05/2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Aeneas' source code is available at https://github.com/google-deepmind/predictingthepast under Apache License 2.0, along with the trained weights, licensed under Creative Commons Attribution-ShareAlike 4.0 International (CC-BY 4.0). Neural networks were developed with JAX v0.4.37 (https://github.com/jax-ml/jax), Flax v0.10.2 (https://github.com/google/flax). The XLA compiler is bundled with JAX and does not have a separate version number.
Data analysis	Dataset processing and analysis used Python v3.9 (https://www.python.org), NumPy v2.1.3 (https://github.com/numpy/numpy), SciPy v1.13.1 (https://www.scipy.org), pandas v2.3.3 (https://github.com/pandas-dev/pandas), beautifulsoup4 v4.12.3 (https://www.crummy.com/software/BeautifulSoup), and Google Colab (https://research.google.com/colaboratory), which is an online service and does not have a version number. Visualizations were generated using matplotlib v3.10.0 (https://matplotlib.org), plotly v5.24.1 (https://plotly.com/python), seaborn v0.13.2 (https://seaborn.pydata.org), and GeoPandas v1.0.1 (https://geopandas.org) and CartoDB basemaps (https://github.com/CartoDB/CartoDB-basemaps) for the map tiles.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The Latin Epigraphic Dataset (LED) was developed by processing and integrating data from the EDR, EDH, and EDCS_ETL databases, all of which are openly accessible under Creative Commons Attribution 4.0 licenses. The original data sources used in this work can be accessed directly at their respective repositories: EDR (https://zenodo.org/records/3575495), EDH (https://zenodo.org/records/3575155), and EDCS_ETL (https://zenodo.org/records/7072337). The processed LED dataset is publicly available online at https://github.com/google-deepmind/predictingthepast. The list of Roman personal names used for evaluating the Onomastics baseline was derived from the Classical Language Toolkit (CLTK) (https://cltk.org/), is available under the MIT License, and the processed proper names are available at https://github.com/google-deepmind/predictingthepast. The new curriculum co-designed as part of this research's wider knowledge exchange and impact strategy is freely available online at https://predictingthepast.com.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Findings do not apply to only one sex or gender; sex and gender were not considered in study design; sex and gender data was not collected in this study. This is because the purpose of the study was to exclusively collect epigraphic annotations for a set of Roman inscriptions, and no participant data was recorded for this purpose.
Reporting on race, ethnicity, or other socially relevant groupings	No socially constructed or socially relevant categorisation variable(s) were used in this manuscript.
Population characteristics	No socially constructed or socially relevant categorisation variable(s) were used in this manuscript.
Recruitment	Potential participants were identified among key contributors in the field of Digital Epigraphy (with 1-4+ publications over the last 20 years) and/or among specialists in the study of Roman inscriptions. The participants were early, mid, and senior career academics in the field of Roman history, specialising in Latin epigraphy, currently employed in the Higher Education sector. In total, 41 experts were contacted via email with an invitation to participate. The 23 that accepted the call were provided with a link to the anonymous online form. Self-selection bias is possible, as participants volunteered to participate. Given the specialized nature of the study, participants' expertise mitigates potential biases affecting the study results.
Ethics oversight	Ethical approval for the AI-Historian evaluation protocol was granted by the Ethics Board of the School of Humanities, University of Nottingham. The protocol adhered to rigorous ethical standards and received a Favourable Ethical Opinion from the Faculty of Arts Research Ethics Committee at the University of Nottingham. In accordance with the approved protocol, all participants were provided with a Participant Information Sheet, a Participant Consent Form, and a GDPR Privacy Notice. A comprehensive Data Management Plan was developed, and an Awareness of Ethical Behaviour for Data Collection form was completed. Evaluation designers also completed two mandatory online courses - Research Integrity and Human Subjects Protections - offered by the University of Nottingham Researchers Academy, ensuring alignment with the highest standards of research ethics and integrity.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study was conducted in three phases: - Phase 1: Participants were assigned a set of Latin inscriptions (five randomized texts per participant) and asked to submit their predictions for the three tasks described above using traditional methods without digital support. - Phase 2: Participants were given another set of inscriptions (five randomized texts per participant) to restore, attribute, and date, with support from an assistive digital tool that provided its own predictions and examples. - Phase 3: Participants completed a brief survey consisting of multiple-choice and free-form responses, assessing their experience

	using the digital aids and commenting on the usefulness of these aids to their workflow. The objective of Phase 1 was to assess the difficulty of the epigraphic tasks for domain experts using traditional methods. The objective of Phase 2 was to evaluate the extent to which digital tools could assist domain experts in completing these complex and time-consuming epigraphic tasks. The objective of Phase 3 was to gather feedback from the domain experts to evaluate their experience using digital aids for the three epigraphic tasks. The answers from each phase were scored using quantitative metrics against ground-truths, with additional qualitative feedback gathered in Phases 2 and 3.
Research sample	Potential participants were identified among key contributors in the field of Digital Epigraphy (with 1–4 or more publications over the past 20 years) and/or among specialists in the study of Roman inscriptions. To ensure inclusivity, the 23 expert researchers involved in the human evaluations reflected: 1) gender diversity (11 male, 12 female); 2) career-stage representation (early-career researchers working alongside full professors, all currently employed in the Higher Education sector); 3) age diversity (ranging from 30 to 67 years old at the time of evaluation); 4) academic specialism (all were key contributors to Roman History, particularly Latin Epigraphy, with 1 to 4 or more publications in the field over the past 20 years). In total 41 participants were contacted via email with an invitation to participate and the 23 that accepted were provided with a link to the anonymous online form.
Sampling strategy	Participants were selected through convenience sampling, specifically targeting individuals with expertise in epigraphy. Since world-wide there is a limited number of experts that could contribute to this evaluation, the sample size was defined by the number of volunteers that were willing to participate.
Data collection	The data collection took place online via an anonymous form. The responses from the participants were collected in an anonymised online form (Google Forms), and final statistics were drawn from the performance of all participants across both phases of the study. The form included the Participant Information Sheet, the GDPR Privacy Notice, and the Participant Consent Form. Participants who consented to take part proceeded to the evaluation section of the form, which was entirely anonymised. To compare the results of Phase 2 and Phase 3, participants were randomly divided into two groups, with one group conducting Phase 2 on a given inscription while the other conducted Phase 3 on the same inscription. The allocation was uniformly randomized. Each participant was assigned a minimum of five inscriptions, with the option to evaluate additional inscriptions if desired. Given the additional information provided between all Phases participants knew the Phase they were in, but didn't have visibility on the group they belonged, or answers from other participants. In total, the 23 participants evaluated 120 inscriptions, representing 60 unique texts.
Timing	The evaluations took place between 2/8/2024 and 25/09/2024. Each evaluation lasted a maximum of 2 hours.
Data exclusions	No data was excluded from the analyses.
Non-participation	No participants dropped out or denied participation. Out of the 41 participants contacted, 23 responded to our call.
Randomization	Participants were not assigned to predefined experimental groups. Instead, the study design followed a paired evaluation framework where all participants first completed tasks independently in Phase 1. In Phases 2 and 3, participants were provided with AI assistance. Each participant was presented with a set of inscriptions, and the selection of either Phase 2 or Phase 3 for each inscription was determined through uniform random sampling.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	n/a
Novel plant genotypes	n/a
Authentication	n/a