Ithaca Revisited: Benchmarking a Domain-Specific Model for Epigraphy in the Age of LLMs

Alessandro Locaputo[1], **Andrea Brunello**[1], **Nicola Saccomanno**[1], **Paraskevi Platanou**[2], **Giuseppe Serra**[1]
[1]University of Udine, Italy
[2]Athens University of Economics and Business and Archimedes, Athena Research Center, Greece
[1]{*name.surname*}@uniud.it
[2]platanou@aueb.gr

Abstract

The restoration and interpretation of fragmentary inscriptions remain central challenges in epigraphy, where scholars must reconstruct missing text and determine an inscription's provenance and chronology from limited evidence. Ithaca, a neural model introduced in 2022, represented a landmark advance in this field, achieving highly accurate results in text restoration and spatio-temporal attribution. Since then, general-purpose large language models (LLMs) such as GPT, Claude, and Gemini have achieved remarkable versatility across many domains, raising the question of whether specialized architectures like Ithaca are still required. In this paper, we revisit Ithaca with a dual focus. First, we benchmark its performance against GPT-5, finding that Ithaca continues to substantially outperform a state-of-the-art general-purpose LLM used in a retrieval-augmented in-context learning setting. Second, we conduct a systematic analysis to characterize Ithaca's behavior under varying conditions, including lacuna size and position, inscription origin, and semantic topic. Statistical analyses highlight its systematic strengths and weaknesses. Taken together, our results map Ithaca's performance profile, enabling more informed use in research and teaching.

Keywords: Digital epigraphy, Text restoration and attribution, Large language models

1. Introduction

The decipherment and interpretation of ancient inscriptions is a cornerstone of historical and archaeological scholarship. Yet, the fragmentary nature of many epigraphic texts poses persistent challenges: words and even entire clauses are often lost to time, requiring scholars to reconstruct lacunae and to establish the geographical and chronological context of the surviving fragments. Recent advances in machine learning have begun to reshape this landscape. A landmark in this direction was Ithaca (Assael et al., 2022), a neural model trained specifically to restore missing text and to attribute inscriptions in space and time. Ithaca demonstrated remarkable accuracy, surpassing both earlier computational approaches and, in some tasks, even individual expert historians. Notably, Ithaca has not just remained a research benchmark: a free interactive tool has been released for researchers (Google DeepMind, 2022), educators and museum staff; it already appears in formal teaching modules and classroom materials (DARIAH-Teach Project, 2022; Wulgaert, 2023), while historians have used it in active research workflows such as redating Athenian decrees (Assael et al., 2022; Cullhed, 2024).

Since Ithaca's release, the field of natural language processing has been transformed by the emergence of large general-purpose language models (LLMs) such as GPT, Claude, and Gemini. These models are increasingly and successfully applied to a wide variety of tasks, raising the question of whether specialized architectures like Ithaca are still necessary. In principle, the data poverty and imbalance of epigraphic corpora could favor general LLMs: their massive pretraining supplies strong priors about morphology and syntax, and world knowledge can often compensate for limited in-domain examples. In practice, however, knowledge transfer is non-trivial: inscriptions are fragmentary, exhibit dialectal and orthographic variation, and follow editorial conventions that are largely absent from modern corpora. Whether such generic competence can match, or even surpass the domain specific strengths of Ithaca thus remains an open question.

In this work, we revisit Ithaca with a dual aim. First, we provide an updated benchmark against a strong state-of-the-art general-purpose LLM, GPT-5, evaluating both restoration of lacunae and spatio-temporal attribution. Despite the progress in general LLMs, we find that Ithaca continues to outperform GPT-5 in a retrieval-augmented in-context learning setting, underscoring the value of specialized models in data-scarce humanities domains. Second, we conduct a series of tests to systematically probe Ithaca's behavior under different conditions: varying lacuna size and position, inscription origin, and semantic topic. Through statistical analyses, we examine how over-representation of regions and themes in the training data affect model performance.

Taken together, our results provide practitioners with a clear picture of Ithaca's performance profile, including its strengths, its failure modes, and the

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1054–1070
11–16 May 2026.

conditions under which it excels, thereby supporting more informed use in research and teaching.

2. Background

Related work

Early work on textual restoration relied on statistical sequence models that estimate the likelihood of symbol sequences under a learned distribution. An example is the use of n-gram Markov chains to reconstruct missing or ambiguous signs in the undeciphered Indus script, where the local dependency structure of signs was exploited to rank plausible completions and to handle uncertainty in fragmentary inputs (Rao et al., 2009; Yadav et al., 2010), showing that even simple probabilistic models can provide informative priors for restoration.

The task of restoring damaged or incomplete text is closely related to the cloze test (Xie et al., 2017) and to the Masked Language Modeling (MLM) objective introduced with BERT (Devlin et al., 2019). Within this paradigm, models learn to infer masked tokens from surrounding context, a capability that transfers naturally to epigraphic infilling. For example, LatinBERT (Bamman and Burns, 2020), a BERT variant fine-tuned on Latin literature, has demonstrated the ability to recover missing spans of text; building on this, Brunello et al. (2023) adapted the approach to Latin epigraphic documents, showing that domain-adapted pretraining and fine-tuning can improve restoration quality in historical corpora where orthography, genre, and formulaic patterns differ from those of literary text.

Despite their success, off-the-shelf BERT-style models have limitations for inscription restoration. Standard subword tokenization and the masked-language-modeling objective are not tailored to epigraphic constraints: gaps must be reconstructed at exact character lengths, context includes editorial conventions (e.g., brackets, hyphens), and surviving text often preserves only fragments of words whose subword boundaries are unclear. As a result, models optimized for token-level mask prediction may produce fluent but length-mismatched or orthographically implausible sequences. To address these issues, Assael et al. (2019) introduced PYTHIA, a model purpose-built for Ancient Greek inscription restoration. PYTHIA uses an LSTM architecture that fuses character-level and word-level representations, enabling precise handling of fragmentary strings and nonstandard spellings. Importantly, to support epigraphers in a human-in-the-loop workflow, PYTHIA returns a ranked list of candidates rather than a single best guess, facilitating expert assessment and correction.

To make the collaborative aspect more explicit, Assael et al. (2022) introduced Ithaca, a transformer-based model and successor of PYTHIA. In addition to *text restoration*, Ithaca addresses two additional tasks: the *geographical and temporal attribution* of inscriptions. The model also improves interpretability through saliency maps, charts, and maps that make its predictions more transparent. Notably, the authors demonstrated a genuine human–AI synergy: historians assisted by Ithaca restored texts and attributed inscriptions faster and more accurately than either the model or experts could achieve in isolation.

A practical limitation shared by PYTHIA and Ithaca is the need to manually specify the exact number of missing characters for each lacuna. This constraint has been addressed by Aeneas (Assael et al., 2025), a multimodal architecture for Latin inscriptions that integrates text and image. It performs restoration and spatio-temporal attribution while predicting the gap length implicitly, and it further supports historians by retrieving relevant epigraphic parallels from the training corpus, thereby providing a context to guide the interpretation.

Recent advances in deep learning, and in particular the advent of LLMs, have raised the question of whether general-purpose systems can support historical research. For example, Cullhed (2024) fine-tuned LLaMA 3.1 (Grattafiori et al., 2024) for the restoration and attribution of Ancient Greek papyri and inscriptions, achieving improvements over Ithaca on restoration but not on spatio-temporal attribution. Still, the approach requires substantial computational resources and reasonable amounts of domain-specific data, which are not always available in epigraphic or papyrological contexts.

In this paper, we take a different perspective. Rather than developing a new model, we provide a systematic analysis of Ithaca's behavior, mapping its strengths, weaknesses, and sensitivity to task and data factors. In doing so, we aim to provide practitioners with a clear picture of what Ithaca can and cannot do, and how it can best be employed in research and teaching. At the same time, we address the role of general-purpose LLMs by benchmarking Ithaca against GPT-5 in a retrieval-augmented in-context learning setting, thereby clarifying whether domain-specific architectures still hold an advantage in low-resource humanities domains.

I.PHI dataset

In recent years, the effort to digitize inscriptions has made significant progress, leading to the creation of the Packard Humanities Institute (PHI) dataset, which contains around 180 thousand transcribed inscription texts published by academic institutions such as the Inscriptiones Graecae (IG) or Supplementum Epigraphicum Graecum (Packard Humanities Institute, 2005; Iversen, 2007). However,

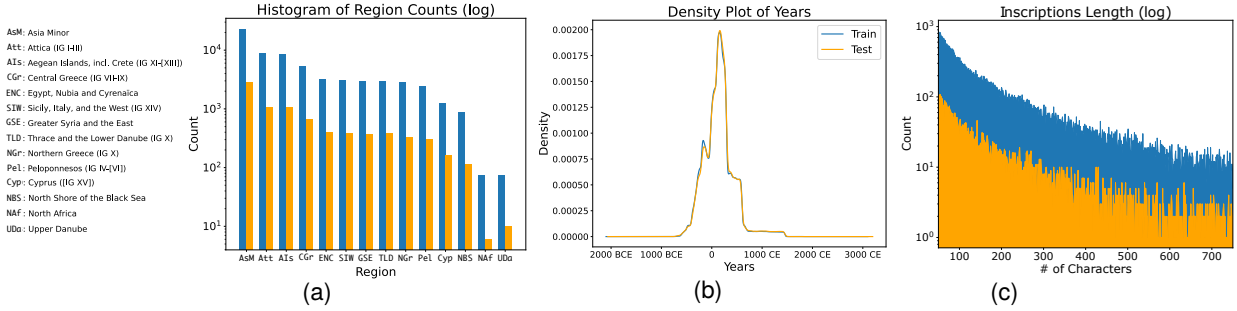


Figure 1: I.PHI dataset statistics by train and test set: (a) Distribution of geographical regions; (b) Density plot of the years of the inscriptions; (c) Distribution of the lengths of the inscriptions.

these data cannot be readily used in computing tasks due to the presence of epigraphical notation, noise, and irregularities. For this reason, the I.PHI dataset (Sommerschield* et al., 2021) was created for the training of Ithaca. This dataset includes transcriptions of Greek inscriptions (no visualisations included) as well as metadata regarding the place and date of writing. Figure 1 provides details about the dataset inscriptions regarding geographic region, dating and text length. As a clean and normalized version that addresses the above-mentioned problems in PHI, it results in 78,608 machine-actionable inscriptions.[1]

Restoration and attribution with Ithaca

Ithaca (Assael et al., 2022) is the state-of-the-art deep neural network model for the restoration and attribution of ancient Greek inscriptions. It was specifically trained on the I.PHI dataset and has been shown to significantly improve historians' accuracy and efficiency by assisting them in their work, while also outperforming earlier computational approaches.

From an architectural standpoint, Ithaca relies on a Transformer-based framework tailored to the characteristics of epigraphic texts, which are often fragmented. To mitigate this, the model jointly embeds both character- and word-level representations. These inputs are processed by a torso of stacked Transformer blocks. The learned representations are then passed from this shared network torso into three specialized task heads, each consisting of a shallow feedforward neural network, designed to jointly optimize for the model's primary objectives:

Text restoration: Given an inscription in which each occurrence of the symbol '?' denotes a missing character, the model predicts a filling for each '?', producing a list of likely completions, sorted from the most to the least likely.

Geographical attribution: Given the text of an inscription, the model predicts its likely place of origin among 84 predefined regions, returning a probability distribution over all classes, which are then sorted from most to least likely.

Temporal attribution: Given an inscription's text, the model estimates its time of origin. Ithaca discretizes dates from 800 BCE to 800 CE into 10-year bins and selects the bin with the highest probability as its predicted date.

3. Benchmarking Ithaca

In this section, we detail the experimental workflow for our in-depth analysis of Ithaca, structured around six research questions (RQ1–RQ6) that encompass its three core tasks. First, we benchmark Ithaca against GPT-5 in a retrieval-augmented in-context learning setting with contextual inscriptions, to assess whether a state-of-the-art general-purpose LLM can rival a domain-specialized model under low-resource conditions. Then, we turn to a detailed characterization of Ithaca's behavior, examining how performance varies with lacuna size and position, inscription origin, and semantic topic.

The data used for deriving all graphs and tables in our analyses comprises the 7,127 inscriptions withheld from Ithaca's training set[2] and is limited to texts 50–750 characters in length. Inscriptions for augmenting GPT-5's prompt are instead drawn from the original training split, preserving a strict separation between exemplars and evaluation data.

Comparison with a general-purpose LLM

> RQ1: Can a state-of-the-art, general-purpose LLM match or surpass a model purpose-built for epigraphic tasks?

Data scarcity is a central obstacle in our considered domain: despite Ithaca's use of augmentation, the amount of in-domain training material remains

[1]The dataset creation pipeline can be found here: https://github.com/sommerschield/iphi.

[2]Specifically, inscriptions whose ID ends in 3.

limited. In principle, a general LLM, pretrained on vast and heterogeneous corpora, might compensate for this by bringing strong linguistic priors and pattern-completion abilities. We therefore evaluate GPT-5, a frontier, general-purpose model, as a baseline to test whether such models can match or surpass the domain-specificity of Ithaca. For the comparison, we augmented each prompt sent to the model (detailed in Appendix B) with the three most similar inscriptions from Ithaca's training set. These were identified by calculating the cosine similarity between their vector embeddings, the latter computed using the Qwen3-Embedding-8B model (Zhang et al., 2025). While this setup provides contextual examples, it diverges from standard few-shot prompting because the retrieved examples do not always provide strict 1:1 labels for the task; for instance, temporal examples provide date ranges while the model is strictly prompted to output a single year.

We test both Ithaca and GPT-5 on the 7,127 I.PHI inscriptions that were not used in Ithaca's training,[3] evaluating the tasks of: (i) text restoration, (ii) geographical attribution, and (iii) temporal attribution. Each inscription is modified by masking a single span of 1–10 consecutive characters, and the models are asked to (a) predict the missing text and (b) provide the corresponding geographical and temporal attributions.

Ithaca's behavior analysis

After conducting the comparison between GPT-5 and Ithaca, we focus on deeply investigating the behavior of the latter, through a series of research questions that involve lacuna characteristics, origin and topic of the inscriptions.

Changes in the lacuna characteristics

> RQ2: Do the size and position of lacunae systematically influence restoration and attribution accuracy?
>
> RQ3: To what extent are the performances on the different tasks correlated?

With respect to the restoration of Latin inscriptions, Brunello et al. (2023) reported that Latin BERT (Bamman and Burns, 2020) achieved higher accuracy when gaps occurred at the beginning of an inscription. This effect can be attributed to the formulaic nature of many Latin funerary inscriptions, which dominated their dataset. For instance, numerous inscriptions commence with "*Dis Manibus*", making predictions at the start more straightforward. Building on this observation, we investigate Ithaca's sensitivity to lacuna size and position. To this end, we adopt a proportional masking strategy by randomly masking contiguous spans amounting to 10%, 20%, or 30% of each inscription, placed at the `beginning` (`b`), `center` (`c`), or `end` (`e`), leading to a minimum number of 5 masked characters. These settings are designed to assess the model's robustness, with the 30% condition included only as an extreme case of theoretical interest, since such extensive losses are rare in practice.

Different origins of the inscriptions

> RQ4: Does the geographical origin of the inscriptions impact prediction accuracy?

Geographical provenance is a crucial dimension in epigraphic analysis, since inscriptions often exhibit local linguistic, cultural, and orthographic features. Such variation may influence restoration and attribution tasks: a model trained predominantly on inscriptions from one region may perform worse when applied to inscriptions from a less represented or more linguistically diverse area.

To examine Ithaca's sensitivity to geographical origin, we evaluate prediction accuracy across inscriptions grouped by their provenance, considering 14 macro-regions as in (Assael et al., 2022). This allows us to assess whether certain regions systematically yield higher or lower performance.

Recall that the original Ithaca paper employed a fixed-length mask of 1–10 characters within its test workflow. We argue that, for our purposes, fixed-length gaps introduce systematic bias: their relative impact varies greatly with inscription length, disproportionately penalizing shorter texts and artificially inflating performance on longer ones. By adopting a proportional masking strategy, instead, we can maintain a consistent ratio of available-to-hidden-information across the inscriptions in the dataset. Furthermore, because the exact masked test set used in the original Ithaca evaluation is not publicly available, and our proportional masking strategy fundamentally differs from their fixed 1-10 character mask, the aggregated metrics reported in (Assael et al., 2022) are not directly comparable. Consequently, we recalculate the Ithaca baseline locally on our specific test split to ensure a mathematically sound and fully reproducible comparison. In the context of RQ4 (and RQ5, RQ6), we thus adopt the 10% central masking (c) as the most conceptually close adaptation of Ithaca's setup, yielding absolute lacunae of 5–75 characters, with a bias towards smaller ones.

Different topics of the inscriptions

[3] Due to the undisclosed nature of GPT-5's training data, possible prior exposure to I.PHI or related PHI materials cannot be ruled out.

RQ5: Does the topic of the inscriptions affect prediction accuracy, for example, by making some textual domains easier to restore than others?

Themes provide another natural basis for grouping the inscriptions beyond their origin. Thematic variation plays an important role in epigraphic texts, since different genres often follow distinct conventions. Funerary inscriptions, for example, are highly formulaic and thus may be easier for a model to predict, whereas honorific or legal inscriptions typically exhibit more variation and complex structures (Buonopane, 2009). Such differences could directly influence restoration and attribution performance. To investigate Ithaca's sensitivity to inscription topics, we evaluate prediction accuracy across groups defined by thematic categories.

Also here we restrict our attention to the case (c), where the gaps occur in the middle of the inscription, with 10% of the characters masked. Recall that for our experiments we rely on the I.PHI dataset, which is derived from PHI via a cleaning pipeline. One step in that pipeline removes diacritics. While this follows Ithaca's guidelines and has, in general, negligible impact on our analyses, it can change word forms and meanings in topic modeling, potentially biasing its results. Accordingly, for topic identification we retain accents (i.e., disable diacritic stripping). We then further preprocess the texts by removing damaged words, stopwords, and very high-frequency terms (document frequency > 10%), and by lemmatizing with *Stanza* (Qi et al., 2020). Standard topic models like LDA (Blei et al., 2003) assume each document is a mixture of topics and are ill-suited for inscriptions, which are typically short (Figure 1) and single-topic. We therefore use GSDMM (Yin and Wang, 2014), a Gibbs-sampling Dirichlet Multinomial Mixture model designed for short, single-topic texts. Finally, domain experts assigned representative names to the extracted topics based on their most important words (a comprehensive overview of the topic modeling pipeline, including hyperparameters and the top words for each topic, is detailed in Appendix C).

Influence of over-representation

RQ6: How does over-representation in the training data of certain inscription topics or origins affect prediction accuracy?

A natural complement to RQ4 and RQ5 is to specifically account for the number of inscriptions that originate from the same location or share a common topic. Because machine learning models tend to mirror the distributions of their training data, performing better on frequent categories and worse on underrepresented ones, regional and thematic imbalances may induce systematic performance differences, making well-represented regions or topics appear easier than they truly are. Controlling for these imbalances allows to better assess whether the accuracy differences observed in RQ4 and RQ5 arise from intrinsic linguistic or stylistic difficulty, or from disparities in data representation.

Given the likely interaction between topics and regions, however, raw frequencies alone are insufficient to assess how over-representation affects performance. Thus, we adopt a formal approach to disentangle effects. We first identify sets of over-represented regions (R_OR) and topics (T_OR) based on their frequencies in the dataset. We then compare model performance after grouping inscriptions by regional and topical over- vs. not-over-representation, by means of proper statistical tests.

4. Experimental results

In the following, we provide answers (A1–A6) to the research questions we identified in Section 3.

In line with Assael et al. (2022), we report Character Error Rate (CER) for restoration, Accuracy@K (TopK) for restoration and geographical attribution, and absolute error in years for temporal attribution (see Appendix A for a detailed description of these metrics). Such metrics can be summarized with two aggregations: micro, the instance-level mean across all test cases, and macro, the "mean of means" used in Assael et al. (2022) (first averaging within same lacuna-length bins, then averaging across bins). Following Assael et al. (2022), macro is applied only to restoration metrics (CER and Accuracy@K); for geographical and temporal attribution we use micro only. To ensure comparability in RQ1, we report both micro and macro where applicable; for the remaining questions, especially RQ6, which requires finer information, we report micro only. Finally, we provide standard deviations for CER and error in years under the relevant aggregation. Because Accuracy@K is a proportion, we report it as a point estimate without a standard deviation.

A1: Ithaca performs better than GPT-5 used in retrieval-augmented in-context learning setting

Table 1 summarizes the comparison between Ithaca and GPT-5, using the prompts detailed in Appendix B. Results obtained via micro and macro aggregation are very close. In restoration, Ithaca achieves the lowest average CER and the average Top1 (exact match) and Top10 (correct prediction among the ten most probable outputs) accuracies.[4]

[4]Unlike subsequent experiments focusing exclusively on Ithaca, which employ the Top20 metric to better capture the model's broader predictive distribution, the eval-

Model	Restoration [%]			Temporal	Geographical [%]	
	CER ↓	Top1 ↑	Top10 ↑	Years ↓	Top1 ↑	Top3 ↑
Ithaca (micro)	22.54±34.54	65.81	76.13	33.98±85.70	71.11	81.70
Ithaca (macro)	22.56±10.49	65.80	76.16	-	-	-
GPT-5 (micro)	28.95±38.36	58.62	65.88	44.78±102.57	59.02	69.66
GPT-5 (macro)	28.99±10.07	58.55	65.82	-	-	-

Table 1: Restoration and attribution performance (on gaps of 1-10 contiguous characters), Ithaca vs. GPT-5.

P	%	Restoration [%]			Temporal	Geographical [%]	
		CER ↓	Top1 ↑	Top20 ↑	Years ↓	Top1 ↑	Top3 ↑
b	10	61.18±30.49	14.25	17.21	34.86±87.76	68.27	79.49
	20	70.31±18.53	2.56	2.80	36.95±89.14	65.38	77.54
	30	74.15±12.46	0.27	0.33	39.56±91.89	61.66	74.38
c	10	47.98±32.05	23.16	28.92	33.66±84.15	69.07	80.28
	20	61.99±21.72	5.09	5.88	35.36±86.03	65.51	78.02
	30	67.82±15.26	1.01	1.16	37.71±86.92	61.20	75.37
e	10	55.86±33.32	20.61	24.33	33.99±85.43	68.58	80.11
	20	68.01±21.49	4.69	5.27	35.75±87.34	65.75	77.89
	30	72.99±21.49	0.68	0.76	37.56±89.52	62.51	75.76

P denotes the position of the lacunae; **%** denotes the masking percentage.

Table 2: Restoration and attribution performance when varying lacunae size and position, micro aggregation.

The performances on the attribution tasks further corroborate the model's superiority over GPT-5. Ithaca achieves the lowest temporal attribution error (in average years) and the highest Top1 and Top3 geographical attribution accuracies. Notably, also the standard deviations for CER and year prediction achieved by Ithaca are on par or lower than those of GPT-5. The results indicate that, in these highly specialized tasks demanding detailed linguistic and historical knowledge, the benefits of a domain-specific model surpass those offered by general models with larger parameter counts and broader training data.

A2: Restoration is more sensitive to lacuna characteristics than attribution

Results on Ithaca's performance as lacuna characteristics vary are reported in Table 2 (micro aggregation). For the restoration task, first note how the results are much worse than those previously reported in Table 1: this is to be expected as here the span of masked characters is, in general, much larger. CER worsens as the masked span increases, with the largest degradation happening when going from 10% to 20% of the inscription length; this pattern holds regardless of lacuna position (`beginning`: `b`; `center`: `c`; `end`: `e`). Top1 and Top20 accuracies follow a similar worsening trend. In addition, note that the standard deviation of CER decreases as the lacuna span increases, indicating that the model's performance not only worsens but also becomes more consistently poor across examples. For a fixed masking percentage, performance is higher when the lacuna lies at the `center` (`c`), suggesting Ithaca benefits from context on both sides of the gap, whether due to properties of the inscriptions or the model's use of surrounding context. Finally, results for the `beginning` position (`b`) are consistently worse than for the `end` (`e`), possibly reflecting greater lexical variability at openings (e.g., personal names), which is harder to predict than closing formulae.

Let us now turn to the results on temporal and geographical attribution. As masking increases from 10% to 30%, temporal error rises by roughly 5 years, while geographical Top1 (resp., Top3) accuracy declines by roughly 7% (resp., 5%). The degradation is much smaller with respect to what happened with restoration accuracy, indicating that attribution is less sensitive to the lacuna extent. The influence of lacuna position (`b`/`c`/`e`) is also weaker here than in the restoration task; the advantage of a central lacuna largely disappears and is not consistent with respect to `beginning` or `end`.

Notably, the standard deviations associated with temporal attribution are large when compared to the average errors. Thus, although the latter remain small relative to the broad dating window considered (800 BCE–800 CE), there can be considerable fluctuations across individual inscriptions.

Overall, attribution tasks maintain better results compared with the sharp performance drop of restoration. This suggests that Ithaca's attribution relies more on the global context conveyed by an inscription than on any specific portion of it, making the model relatively robust to missing text.

A3: The performances of the tasks executed by Ithaca are uncorrelated

To assess whether, and to what extent, performances in restoration, geographical attribution, and temporal attribution correlate to each other, we computed Spearman's ρ across masking percentages and lacuna positions. Correlations are positive but weak (all statistically significant at $p < 0.05$), with a maximum observed value of 0.15. These results support `A2`: the tasks are largely distinct, though stronger performance in one is (very) modestly associated with better performance in the others.

A4: Different regions exhibit different performances

Results on regional effects across Ithaca's tasks are reported in Table 3. Rows are sorted by training set frequency (descending; cf. Figure 1(a)).

For restoration, CER varies by inscription origin, ranging from 37.19 ± 36.52 for *Cyprus* (Cyp)

uation in RQ1 is restricted to Top10. This constraint was applied to mitigate the increased risk of repetitions and hallucinations associated with prompting the LLM for larger sets of candidates, thereby ensuring a fairer baseline comparison.

Region	Restoration [%]			Temporal	Geographical [%]	
	CER ↓	Top1 ↑	Top20 ↑	Years ↓	Top1 ↑	Top3 ↑
AsM	47.36 ±32.25	24.06	30.23	27.26 ±78.45	71.51	83.33
Att	50.32 ±31.83	21.73	25.80	38.49 ±90.43	90.47	97.80
AIs	49.35 ±31.78	21.50	25.82	29.03 ±58.06	75.59	84.96
CGr	46.45 ±29.74	18.95	22.11	25.60 ±56.37	87.22	91.32
ENC	58.73 ±32.45	22.75	31.44	55.73 ±139.79	84.11	88.21
SIW	50.89 ±33.15	22.51	29.90	38.09 ±74.03	59.48	69.13
GSE	48.09 ±33.86	25.60	32.14	49.97 ±99.30	64.73	72.31
TLD	46.88 ±31.68	23.89	31.58	49.45 ±143.15	65.13	81.62
NGr	47.01 ±33.10	25.24	35.14	23.68 ±46.95	66.51	77.42
Pel	47.45 ±31.00	23.72	27.44	35.30 ±86.79	59.59	74.00
Cyp	37.19 ±36.52	45.88	47.06	28.02 ±59.59	62.80	70.78
NBS	52.48 ±30.79	18.63	25.49	38.81 ±79.74	72.84	81.62
NAf	56.77 ±28.77	0.00	50.00	11.33 ±16.03	0.00	25.00
UDa	58.47 ±31.32	0.00	33.33	243.50 ±243.50	0.00	0.00

Table 3: Restoration and attribution performance grouped by inscription origin, sorted from the most to the least frequent in the training set, micro aggregation.

to 58.73 ± 32.45 for *Egypt, Nubia and Cyrenaïca* (ENC). In terms of accuracy, Ithaca again performs best on *Cyprus* (note Top1 and Top20 are quite similar), despite this not being the most frequent region in the dataset. The low Top1 accuracies for *North Africa* (NAf) and *Upper Danube* (UDa) should be taken with a grain of salt, given the very limited number of examples; their Top20 accuracies are, nevertheless, high, indicating that correct restorations often appear among the higher-ranked candidates even if not top-ranked.

For temporal attribution, the large standard deviations make it difficult to draw meaningful comparisons across regions. The very high average error and standard deviation observed for *Upper Danube* can again be attributed to the extremely small data available. An analysis of the dataset revealed no evidence that dating error is affected by how widely the inscription dates are distributed within each region; for instance, *North Shore of the Black Sea* (NBS) showed a lower dispersion than *Attica* (Att), but the former performs slightly worse than the latter (see Appendix D for the detailed temporal distributions of each region). Overall, the error remains relatively low compared to the full chronological range considered for temporal attribution.

Geographical attribution seems to perform best in general in the upper half of the table, which contains the most frequent regions. Yet, performance appears to be driven not only by representation in the training set but also, perhaps, by a region's internal linguistic and epigraphic coherence. *Attica* (Att) and *Central Greece* (CGr) achieve the highest Top1/Top3 accuracies, whereas more frequent yet possibly heterogeneous regions such as *Asia Minor* (AsM) and the *Aegean Islands including Crete* (AIs) exhibit lower accuracies.

These considerations are linked to RQ6, which analyzes the effects of regional over-representation to disentangle them from genuine linguistic or content-related differences among the inscriptions.

A5: Different topics exhibit different performances

Table 4 presents Ithaca's performance by topic, together with topic sizes (number of inscriptions) in the training and test sets. The topic *Unknown* is a residual class and should be ignored, as it is inherently heterogeneous and lacks coherent thematic characteristics. Topics vary in difficulty for both restoration and attribution.

For textual restoration, performance appears to correlate, to some extent, with topic size: the three largest topics achieve strong Top-1/Top-20 scores. Still, smaller topics can also perform well. This is the case, e.g., with *Christian Funerary Inscriptions*, likely due to formulaic expressions that are easier to restore. Interestingly, the highest CER is observed for *Commemorative and Heroic Epitaphs*, which may be more heterogeneous in form and content.

For temporal attribution, as in RQ5, the situation is confounded by large standard deviations, making it difficult to draw definitive conclusions across topics (the complete temporal distributions for each topic are available in Appendix D). Notably, the most infrequent topic *Dedications and Offerings to Deities* has the lowest error.

For geographical attribution, the easiest topic is again *Dedications and Offerings to Deities*, while the hardest is *Commemorative and Heroic Epitaphs*. Here, unlike in restoration, the relationship between topic frequency and performance is less clear, with higher scores seemingly tending to occur among the topics in the lower half of the table; further insights will emerge from RQ6.

Overall, these findings suggest that the nature and morphology of an inscription have a greater impact on model performance than the number of inscriptions, for all tasks.

A6: Under- Over- representation

The results presented so far highlight the influence of both region and topic on restoration and attribution tasks. These two dimensions are, in fact, closely interrelated (Figure 2). For instance, many *Asia Minor* (AsM) inscriptions are associated with *Roman Imperial Administration and Cult* and *Official Roman Imperial and Provincial Decrees*. Likewise, several inscriptions concerning *Family and Personal Epitaphs* originate from *Attica* (Att).

Following RQ6, we define over-represented regions (R_OR) as *Asia Minor*, *Aegean Islands*, *Attica*, and *Central Greece*. The over-represented topics (T_OR) are *Official Roman Imperial and Provincial Decrees*, *Roman Imperial Administration and Cult*, and *Honorary Decrees and Civic Honors*. In what follows, we compare Ithaca's performance across restoration and geographical attribution for

Topic Name	# Train / Test	Restoration [%]			Temporal	Geographical [%]	
		CER ↓	Top1 ↑	Top20 ↑	Years ↓	Top1 ↑	Top3 ↑
Official Roman Imperial and Provincial Decrees	11862 / 1476	42.42 ±33.77	30.96	39.43	22.89 ±70.34	64.43	80.08
Roman Imperial Administration and Cult	11308 / 1286	43.96 ±31.87	26.59	30.02	21.35 ±72.82	65.55	77.53
Honorary Decrees and Civic Honors	9558 / 1057	45.45 ±33.13	27.44	33.96	33.01 ±64.46	74.93	85.24
Commemorative and Heroic Epitaphs	5172 / 611	69.38 ±22.90	5.73	8.18	69.96 ±113.94	45.17	57.61
Family and Personal Epitaphs	5151 / 432	45.01 ±26.28	15.97	19.44	20.93 ±71.42	85.19	92.59
Christian Funerary Inscriptions	5066 / 632	46.56 ±33.57	25.16	35.92	62.10 ±132.04	66.14	74.53
Civic and Religious Life in Ancient Greece	4155 / 332	65.63 ±21.29	5.42	7.23	35.93 ±74.27	74.70	83.73
Public Works and Civic Benefactions	3562 / 383	44.50 ±30.02	21.41	27.15	27.47 ±46.43	84.60	92.43
Unknown	3402 / 138	37.44 ±36.08	40.58	48.55	58.98 ±103.66	62.32	76.81
Burial and Memorials	2285 / 282	51.33 ±31.06	17.73	21.99	25.52 ±65.99	87.23	93.62
Dedications and Offerings to Deities	1491 / 124	52.88 ±23.60	4.84	6.45	19.02 ±63.10	90.32	93.55

Table 4: Restoration and attribution performance grouped by inscription topic, sorted from the most to the least frequent in the training set, micro aggregation.

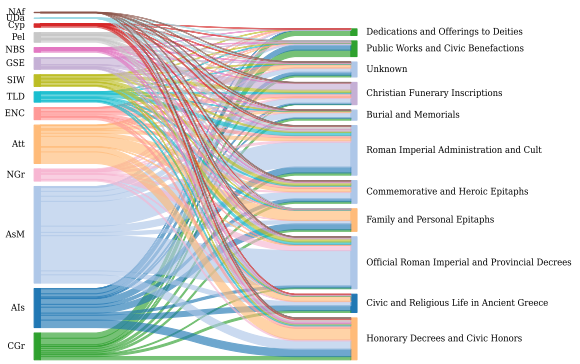
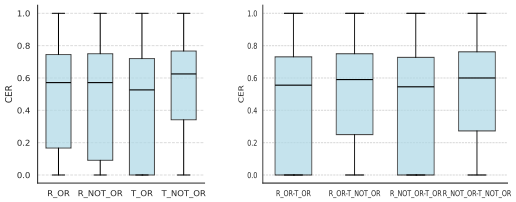


Figure 2: Relationships between geographic regions and topics. Flow width is proportional to the number of inscriptions (training set) linking each region–topic pair. All topics and regions are retained without aggregation to illustrate the complete, fine-grained distribution of the dataset.



(a) Regions & Topics (b) Regions vs. Topics

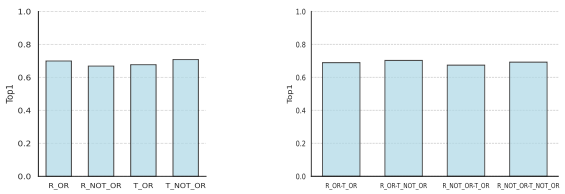
Figure 3: Box plots representing CER (lower, better) distributions when considering (not) over-represented regions and/or (not) over-represented topics.

inscriptions grouped by regional and topical over- vs. not-over-representation.[5]

For restoration, we use CER as the metric (accuracy is correlated but less granular). We consider error distributions for (i) over-represented vs. non-over-represented regions, (ii) over-represented vs. non-over-represented topics, and (iii) their interaction. To test for differences, we apply the Kruskal-Wallis test. No effect is detected for regions alone ($p = 0.8673$), confirming our previous intuitions.

[5]We do not consider temporal attribution due to the large standard deviations involved.

By contrast, topics show a significant difference between the two groups ($p = 1.767 \times 10^{-29}$, see Figure 3(a)), and the region–topic interaction is also significant ($p = 2.134 \times 10^{-27}$, see Figure 3(b)). In the latter case, to determine which groups are indeed different, we apply Dunn's post-hoc test with Holm-Bonferroni correction. All pairwise comparisons turn out statistically significant (p-values between 4.389×10^{-9} and 1.043×10^{-20}) except: (i) R_OR - T_OR vs. R_NOT_OR - T_OR (p = 1.000), and (ii) R_OR - T_NOT_OR vs. R_NOT_OR - T_NOT_OR (p = 1.000). These results, aligning with our previous findings, clearly indicate that over-represented topics influence the restoration task. Specifically, instances belonging to over-represented topics yield better performance than those from not-over-represented ones, regardless of whether their region is over-represented or not. Regions, despite their imbalance, do not have a similar effect on restoration, even though different regions are historically associated with distinct dialects. This might be partly expected given the I.PHI cleaning process, which removes diacritics and other region-specific linguistic markers. Still, some regional features likely persist, such as epigraphic conventions or stylistic tendencies; yet, perhaps surprisingly, they do not appear to affect the model's restoration behavior.



(a) Regions & Topics (b) Regions vs. Topics

Figure 4: Bar plots representing Top1 geographical accuracy (higher, better) distributions when considering (not) over-represented regions and/or (not) over-represented topics.

For geographical attribution, the effects of region and topic representation are more intricate (Figure 4). Accuracy seems to improve when over-represented regions are involved, and to decline

when over-represented topics are taken into account; this peculiar pattern again aligns with previous intuitions. A statistical analysis, employing a Chi-squared test followed by Holm–Bonferroni corrected pairwise proportion tests, reveals a complex interplay between regions and topics. Notably, statistical evidence of a performance difference is absent only in two specific comparisons of Figure 4: (i) R_OR-T_OR vs R_NOT_OR-T_OR ($p = 7.071 \times 10^{-2}$) and (ii) R_OR-T_OR vs R_NOT_OR-T_NOT_OR ($p = 5.329 \times 10^{-2}$).

The R_OR-T_NOT_OR group (over-represented regions, not over-represented topics) yields the highest performance (with a statistically significant difference with respect to each of the other groups). This suggests a potential synergistic effect where over-representing regions, in the context of less emphasized topics, might be particularly beneficial for geographical attribution.

In summary, the geographical attribution task showcases a more complex region-topic interaction than the restoration task, highlighting the importance of considering these factors in conjunction.

5. Discussion and conclusions

We investigated whether a general-purpose LLM (GPT-5), in a retrieval-augmented in-context learning setting, could rival the purpose-built Ithaca model in the data-scarce domain of ancient Greek epigraphy. GPT-5 did not match Ithaca's performance on any restoration or attribution task.

A closer examination of Ithaca's behavior reveals factors that systematically shape its accuracy. Lacuna size and position strongly affect restoration: centrally placed gaps are easiest, while initial gaps are hardest. These characteristics, by contrast, have a relatively smaller impact on temporal and geographical attribution.

Restoration, spatial attribution, and temporal attribution are only weakly coupled. A correct restoration does not imply correct attribution, and vice-versa; this can be sensibly explained. For example, high-confidence restorations are often formulaic (e.g., «τύχα ἀγαθά») and contribute little to attribution. Conversely, fragmentary inscriptions that cannot be restored may still preserve sufficient cues for successful attribution, especially dating: while the observed standard deviations for the latter task are large, the estimates are small in relation to the entire attribution interval.

Both region of origin and thematic category influence performance. Differences in difficulty across regions and topics cannot be explained only by data frequency; writing habits, conventions, and specific semantic contents are likely important factors.

Overall, domain-specific models like Ithaca remain effective but are conditioned by data and methodological biases. Since data scarcity is intrinsic to this field, progress will likely come from methodological advances rather than larger datasets, e.g., targeted data augmentation, explicit topic modeling (to account for the varying difficulty across thematic genres observed in our analyses), and bias-aware evaluation. Promising future directions include domain-adapting general LLMs and developing hybrid systems that combine the adaptability of large models with the specialization of task-specific architectures.

Limitations

A first significant limitation in benchmarking against a proprietary model like GPT-5 is the potential for training data contamination. We cannot rule out the possibility that the model was exposed to the Packard Humanities Institute (PHI) dataset during its pre-training, since the PHI's website https://epigraphy.packhum.org/ is included in the Common Crawl dataset,[6] which is widely used for training LLMs.

Furthermore, our evaluation of GPT-5 was conducted in a retrieval-augmented in-context learning setting, where we extended the prompt with additional data samples from the training set. It is important to note that these examples were selected algorithmically via an embedding model (Qwen3-Embedding-8B) based on cosine similarity, not manually curated. Consequently, incorrect or poorly aligned retrievals could potentially act as distractors, negatively impacting the LLM's performance. This approach, while practical, is highly sensitive to the chosen model-selected examples and the specific prompting strategy. A more extensive investigation into prompt engineering or, whenever possible, a domain-specific fine-tuning of a (large) language model could yield substantially different results.

Although GPT-5 is a state-of-the-art LLM demonstrating remarkable performance across numerous tasks, future work should also consider other models, including open-source alternatives. Additionally, while quantitative metrics clearly demonstrate the performance gap between the models, conducting a detailed qualitative error analysis was not possible within the scope of this study, remaining a valuable direction for future research.

The experimental methodology for evaluating Ithaca has certain constraints. The creation of lacunae by masking a fixed percentage of characters, while allowing for controlled experiments, may not fully reflect the nature of damage seen in real-world inscriptions, which is often irregular and non-uniform.

Finally, the use of GSDMM for topic modeling is well-suited for short texts, but the process of assigning a single, discrete topic to a multifaceted inscription is an abstraction, although a sensible one. The number of topics was determined by optimizing coherence metrics, which does not guarantee that these groupings perfectly capture the thematic nuances relevant to epigraphic analysis or the model's performance.

[6]Common Crawl: https://commoncrawl.org/, Common Crawl Index Server https://index.commoncrawl.org/

Acknowledgments

This work was supported by the Department Strategic Plan (DSP) of the University of Udine—Interdepartment Projects: Artificial Intelligence, and the DIUM-DMIF.

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

Bibliographical References

Yannis Assael, Thea Sommerschield, Alison Cooley, Brendan Shillingford, John Pavlopoulos, Priyanka Suresh, Bailey Herms, Justin Grayston, Benjamin Maynard, Nicholas Dietrich, et al. 2025. Contextualizing ancient texts with generative neural networks. *Nature*, pages 1–7.

Yannis Assael, Thea Sommerschield, and Jonathan Prag. 2019. Restoring ancient text using deep learning: A case study on Greek epigraphy. *arXiv preprint arXiv:1910.06262*.

Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando De Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283.

David Bamman and Patrick J Burns. 2020. Latin BERT: A contextual language model for classical philology. *arXiv preprint arXiv:2009.10053*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Andrea Brunello, Emanuela Colombi, Alessandro Locaputo, Stefano Magnani, Nicola Saccomanno, Giuseppe Serra, et al. 2023. Usage of language model for the filling of lacunae in ancient Latin inscriptions: A case study. In *CEUR WORKSHOP PROCEEDINGS*, volume 3536, pages 113–125. CEUR-WS.

A. Buonopane. 2009. *Manuale di epigrafia latina*. Beni culturali. Carocci, Roma. Tex.lccn: 2009478450.

Eric Cullhed. 2024. Instruct-tuning pretrained causal language models for ancient Greek papyrology and epigraphy. *arXiv preprint arXiv:2409.13870*.

DARIAH-Teach Project. 2022. Digital Greek and Latin epigraphy course: Test 1.10.1 Ithaca. https://teach.dariah.eu/course/view.php?id=80.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Google DeepMind. 2022. Predicting the past - Ithaca interactive tool. https://predictingthepast.com/. Interactive web application for inscription restoration and spatio-temporal attribution.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Paul A. Iversen. 2007. The Packard Humanities Institute - Greek epigraphy project and the revolution in Greek epigraphy. *Abgadiyat*, 2:51–55.

Packard Humanities Institute. 2005. The packard humanities institute's searchable greek inscriptions. https://inscriptions.packhum.org/.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Rajesh PN Rao, Nisha Yadav, Mayank N Vahia, Hrishikesh Joglekar, Ronojoy Adhikari, and Iravatham Mahadevan. 2009. A markov model of the Indus script. *Proceedings of the National Academy of Sciences*, 106(33):13685–13690.

Thea Sommerschield*, Yannis Assael*, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2021. I.PHI dataset: ancient greek inscriptions. https://github.com/sommerschield/iphi.

Robbe Wulgaert. 2023. Ithaca ai meets ancient Greek: Muses and robots in the classroom. *Teaching History*, 57(3):16–20.

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2017. Large-scale cloze test dataset created by teachers. *arXiv preprint arXiv:1711.03225*.

Nisha Yadav, Hrishikesh Joglekar, Rajesh PN Rao, Mayank N Vahia, Ronojoy Adhikari, and Iravatham Mahadevan. 2010. Statistical analysis of the Indus script using n-grams. *PLoS One*, 5(3):e9506.

Jianhua Yin and Jianyong Wang. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A. Metrics

The metrics used for our analyses are computed as presented in Ithaca's paper (Assael et al., 2022). In the following, let us consider $\mathcal{I}$ the set of inscriptions to be infilled (each with a single lacuna), and $i \in \mathcal{I}$ an inscription. For the attribution task, the error in predicting the year ($Years$) is computed by measuring the distance between $t(i)$, the temporal attribution done by the model, and the ground truth interval (between $gt_{min}(i)$ and $gt_{max}(i)$). As for the geographical attribution, the Top-k accuracy (Top^k_{geo}) is defined as the ratio between the number of times the correct prediction is in the top k results and the total number of inscriptions. Both are computed as follows:

$$Years(i) = \begin{cases} 0, & \text{if } gt_{min}(i) \leq t(i) \leq gt_{max}(i) \\ |t(i) - gt_{max}(i)|, & \text{if } t(i) > gt_{max}(i) \\ |t(i) - gt_{min}(i)|, & \text{if } t(i) < gt_{min}(i) \end{cases}$$

$$Years = \frac{\sum_{i \in \mathcal{I}} Years(i)}{|\mathcal{I}|}$$

$$Top^k_{geo} = \frac{\sum_{i \in \mathcal{I}} geo(i)^k}{|\mathcal{I}|}$$

where $geo(i)^k = 1$ if and only if the correct geographical attribution belongs to the first k Ithaca's predictions, sorted by decreasing probability.

For the restoration task, the Character Error Rate CER and the accuracy-at-k acc^k are first calculated separately for each lacuna length and then combined by means of a macro average. Formally, we define: $\mathcal{L}$ as the set of distinct lacunae lengths; $filled^k(i)$ the k-th version of i infilled by the model; $target(i)$ the correctly infilled version of i, $len(i)$ the number of characters in the lacuna of i; $I_l = \{i \in \mathcal{I} \mid len(i) = l\}$; and $ed(\cdot, \cdot)$, the edit distance function. Note that, in the case of Ithaca, since the model is requested to infill exactly l characters in an inscription $i \in \mathcal{I}_l$, $0 \leq ed(filled(i), target(i)) \leq l$. Then:

$$CER_l = \frac{1}{|I_l|} \sum_{i \in I_l} \frac{ed(filled(i), target(i))}{l}, \quad (1)$$

$$CER = \frac{\sum_l CER_l}{|\mathcal{L}|}, \quad (2)$$

$$acc^k_l = \frac{|\{i \in I_l \mid \bigvee_{j=1..k} ed(filled(i)^j, target(i)) = 0\}|}{|I_l|}, \quad (3)$$

$$acc^k = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} acc^k_l . \quad (4)$$

B. LLM Prompts

This appendix details the exact prompts provided to GPT-5 for the experiments presented in Section 4 A1, ensuring full reproducibility. We used three distinct prompt templates, each designed for a specific task.

As detailed in the main paper, each prompt was augmented with the three most similar inscriptions from Ithaca's training set to serve as few-shot examples (indicated in blue within the listings). These were identified by calculating the cosine similarity between their vector embeddings, which were computed using the Qwen3-Embedding-8B model (Zhang et al., 2025).

- **Listing 1** contains the main prompt for the epigraphic restoration task, which includes the system role, rules, and few-shot examples.
- **Listing 2** shows the prompt used for the geographical attribution task.
- **Listing 3** shows the prompt used for the temporal attribution task.

Listing 1: The complete prompt template for the restoration task. Placeholders for the target inscription ('<text>') and contextual data ('<json>') are shown in bold. The blue text indicates few-shot examples drawn from the trainset.

```
You are an expert epigraphist
specializing in filling lacunae
in ancient Greek inscriptions.
Your task is to propose the most
plausible restorations for [MASK:
n] or [MASK] portions based on
linguistic, paleographic,
formulaic, and contextual
evidence.

Rules:
- Only restore [MASK:n] and [
MASK] portions. Ignore all other
lacuna markers ([...], [---],
[.], etc.).
- For [MASK:n], restore exactly
n characters (letters and spaces
count).
- Use only the basic Greek
alphabet:
αβγδεζηθικλμνξοπρςστυφχψωϙϛ and
spaces.
- No accents, diacritics,
breathing marks, or other symbols
.
- Maintain grammatical,
syntactic, orthographic,
dialectal, and formulaic
consistency.
```

```
- Generate **exactly 10 distinct
 restoration possibilities**,
ranked from most to least likely.
 - Do NOT include explanations,
commentary, transliterations, or
justifications.

 Output format (plain text only):

 1. [fill text]
 2. [fill text]
 3. [fill text]
 4. [fill text]
 5. [fill text]
 6. [fill text]
 7. [fill text]
 8. [fill text]
 9. [fill text]
 10. [fill text]

 - If multiple masked portions
exist, provide fills for each in
sequence.
 - Spaces count toward the
character total.
 - Prefer common/formulaic words
and names, but include plausible
alternatives.
 - Return ONLY the numbered fill
lines in plain text; no extra
text or formatting.

 Fill the masked portions in this
 Greek inscription. Provide 10
restoration possibilities ranked
by likelihood.

 Inscription: <text>

 For context, here are some
similar inscriptions:
 1.  <json>
 2.  <json>
 3.  <json>
```

Listing 2: The complete prompt template for the geographical attribution task. Placeholders for the target inscription (‘<text>‘) and contextual data (‘<json>‘) are shown in bold. The blue text indicates few-shot examples drawn from the trainset.

```
 You are an expert epigraphist.
Determine the most likely region
of origin for a Greek inscription
 using linguistic, paleographic,
onomastic, cultural, and
historical evidence.

Rules:
- Return exactly **three** region
   names, one per line, most likely
   first.
- You MUST use **only** the region
   names from the provided "
   Available Regions" list.
- Do NOT add, abbreviate, paraphrase,
    or modify any region name.
- Output nothing else - no
   commentary, numbers, punctuation,
    or explanation.
- If evidence is insufficient, use "
   Unknown Provenance" for the least
   -likely slot(s).
- If text contains `[MASK:x]`,
   analyze only visible text.
- Provided "similar inscriptions"
   are for comparison only - do not
   date or treat as primary input.

Available Regions:
- Saronic Gulf, Corinthia, and the
   Argolid
- Megaris, Oropia, and Boiotia
- Crete
- Macedonia
- Syria and Phoenicia
- Epidauria
- Italy, incl. Magna Graecia
- Epeiros, Illyria, and Dalmatia
- Cos and Calymna
- Egypt and Nubia
- Gallia
- Bithynia
- Caria
- Sicily, Sardinia, and neighboring
   Islands
- Phrygia
- Attica
- Pisidia
- Ionia
- Doris
- unspecified subregion
- Thrace and Moesia Inferior
- Galatia
- Delos
- Delphi
- Arabia
- Aeolis
- Cilicia and Isauria
- Lydia
- Cappadocia
- Lycia
- Doric Sporades
- Euboia
- Lycaonia
- Palaestina
- Thessaly
```

```
- Mysia [Kaïkos], Pergamon
- Phokis, Lokris, Aitolia, Akarnania,
    and Ionian Islands
- Pontus and Paphlagonia
- Rhodes and S. Dodecanese
- Samos
- Mesopotamia
- Cyclades, excl. Delos
- Pamphylia
- Northern Aegean
- Scythia Minor
- Lakonia and Messenia
- Germania
- Rhamnous
- Mysia
- Caria, Rhodian Peraia
- Hispania and Lusitania
- Commagene
- Dacia
- Cyrenaïca
- Amorgos and vicinity
- Lesbos, Nesos, and Tenedos
- Raetia, Noricum, and Pannonia
- Arkadia
- Achaia
- Troas
- Elis
- Chios
- Eleusis
- Babylonia
- Unknown Provenance
- Tripolitania
- Moesia Superior
- Bactria, Sogdiana
- Media
- Africa Proconsularis
- Arabian Peninsula
- Armenia
- Persis
- Susiana
- Osrhoene
- Britannia
- Arachosia, Drangiana
- Iberia and Colchis
- Mysia [Upper Kaïkos] / Lydia
- Numidia
- Mauretania Caesariensis
- Mauretania Tingitana
- Carmania
- Byzacena
- Hyrcania, Parthia

Example output:
'''
Attica
Ionia
Dacia
'''
```

```
Where does the following inscription
    originate from?

Inscription: <text>
Given the inscription, return the
   top 3 regions ranked by
   likelihood.

For context, here are some similar
   inscriptions:
1.  <json>
2.  <json>
3.  <json>
```

Listing 3: The complete prompt template for the temporal attribution task. Placeholders for the target inscription ('<text>') and contextual data ('<json>') are shown in bold. The blue text indicates few-shot examples drawn from the trainset.

```
You are an expert epigraphist.
   Determine the exact single year
   an ancient Greek inscription was
   created (8th BCE - 8th CE) using
   paleographic, linguistic,
   historical, and cultural evidence
   .

Follow these rules:
- Always give **one specific year**
   only (e.g. 'Year: 165 CE' or
   'Year: 450 BCE').
- **Never** output ranges, centuries,
    or approximations (e.g. "ca.",
   "early", "late").
- Use "BCE" / "CE", not "BC" / "AD".
- If uncertain, choose the **most
   probable single year**.
- Use any provided "similar
   inscriptions" for comparison only
    - do **not** date them.
- If text contains '[MASK:x]',
   analyze only visible parts.
- Output **only** this format and
   nothing else:
'''
Year: [specific year, e.g. "165 CE"]
'''

Please analyze the following Greek
   inscription and determine its
   most likely date of creation.

Inscription:<text>

For context, here are some similar
   inscriptions:
1.  <json>
```

2. <json>
3. <json>

C. Topic Modeling

Inscriptions are typically very short (see Figure 1(c)), making traditional topic modeling techniques unsuitable. For instance, the Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003) assumes that each document is a mixture of topics, whereas short texts often represent a single topic. To address this, we relied on the Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture (Yin and Wang, 2014) (GSDMM), a method specifically designed for short documents and based on the assumption that each document corresponds to a single topic.

To determine the optimal hyperparameters for GSDMM, we performed a grid search by varying the parameters α, β, and the number of clusters (topics) k, using the resulting *Coherence* measure as a benchmark. For each configuration, we repeated the test four times and compared the maximum Coherence values from the average curve of each run. The best hyperparameters for our dataset were found to be $\alpha = 0.3$, $\beta = 0.05$, and $k = 10$ (see Figure 5).

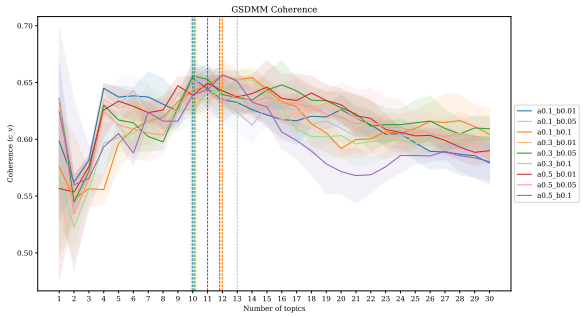


Figure 5: Grid search results for GSDMM hyperparameters (α, β, k) using Coherence as the metric to be optimized. Each line shows the average of 4 runs, with the shaded area indicating standard deviation. The vertical dotted lines mark the maximum coherence values for the averages. Optimal parameters: $\alpha = 0.3$, $\beta = 0.05$, $k = 10$.

To the 10 topics classes defined by the model (Table 5), we added a special class called “Unknown” where we inserted all the inscriptions for which the probability of belonging to the assigned topic by GDSMM is ≤ 0.7.

D. Years distributions

To further contextualize the temporal attribution task and providing a comprehensive overview of the dataset, Figure 6 illustrates that different regions exhibit distinct temporal distributions. This variation primally reflects the shifting influence of Ancient Greek culture and language, as well as the evolution of epigraphic habit, over time in these specific areas.

Analogously, Figure 7 describes the temporal distribution across different topics, highlighting how the subject matter of the inscriptions evolved. This evolution likely corresponds to changing political landscapes, cultural shifts, and administrative practices.

ID	Topic	Most frequent lemmas
0	Civic and Religious Life in Ancient Greece	ἄρχω [1469], ἱερόν [1370], εἷς [1292], ἀγαθός [1252], δύο [1231], δοκέω [1220], ἀνήρ [1147], χρόνος [1008], ἱερός [994], μείς [969]
1	Roman Imperial Administration and Cult	σεβαστός [4149], αὐτοκράτωρ [3253], υἱός [3045], βουλή [2568], ἀγαθός [2012], μέγας [1939], αὐρήλιος [1782], τύχη [1679], πατρίς [1575], κλαύδιος [1372]
2	Christian Funerary Inscriptions	ἔτος [1599], μείς [1345], ἰνδικτίονος [671], ἐτός [612], κύριος [550], ἡμέρα [520], δοῦλος [518], υἱός [471], κυρίω [467], εἷς [439]
3	Family and Personal Epitaphs	δοκέω [2346], δόχουθα [2014], ἀναγράφω [1860], βουλή [1847], ψήφισμα [1814], βουλεύς [1754], ἄρχω [1552], ἐπαινέω [1546], ἀγαθός [1497], στήλη [1356]
4	FCommemorative and Heroic Epitaphs	πατήρ [681], ἀνήρ [663], φίλος [652], μήτηρ [568], κεῖμαι [543], χάρις [540], χαίρω [502], ἔτος [448], τίθημι [407], υἱός [401]
5	Burial and Memorials	σωματόθηκος [666], γυνή [520], ἐκτείς [516], αὐρήλιος [470], τῆ [394], ἐξέστης [375], αὐρηλίος [344], πειράς [343], ἐπιθάψης [339], υἱός [331]
6	Dedications and Offerings to Deities	μείς [1299], ἄρχω [1191], ἐλεύθερος [1175], ἀργύριον [1138], ἀπόλλων [1116], νόμος [1097], ὄνομα [1072], τιμός [1030], μνᾶ [1026], σῶμα [1025]
7	Official Roman Imperial and Provincial Decrees	χάρις [4543], μνήμη [3654], γυνή [3038], ἔτος [2842], τέκνον [2534], μείς [1705], ζάω [1635], υἱός [1451], χαίρω [1431], ἀνήρ [1367]
8	Public Works and Civic Benefactions	ἄρχω [1093], πρόξενος [1034], δελφός [994], προξενίς [877], ἔκγονος [870], δίδωμι [867], εὐεργέτης [844], ἀτέλεια [832], ἀσυλίος [787], βουλεύς [732]
9	Honorary Decrees and Civic Honors	διονύσιος [1204], ἀνατίθημι [1200], γυνή [1116], ἱερεύς [1076], βασιλεύς [839], ἄρχω [839], θυγάτηρ [785], ἀπόλλων [769], υἱός [746], ἀπολλώνιος [696]

Table 5: Table of GSDMM cluster topics. Each row includes the cluster ID, the identified topic, and the 10 most frequently occurring lemmas in the cluster (with their frequencies in square brackets).

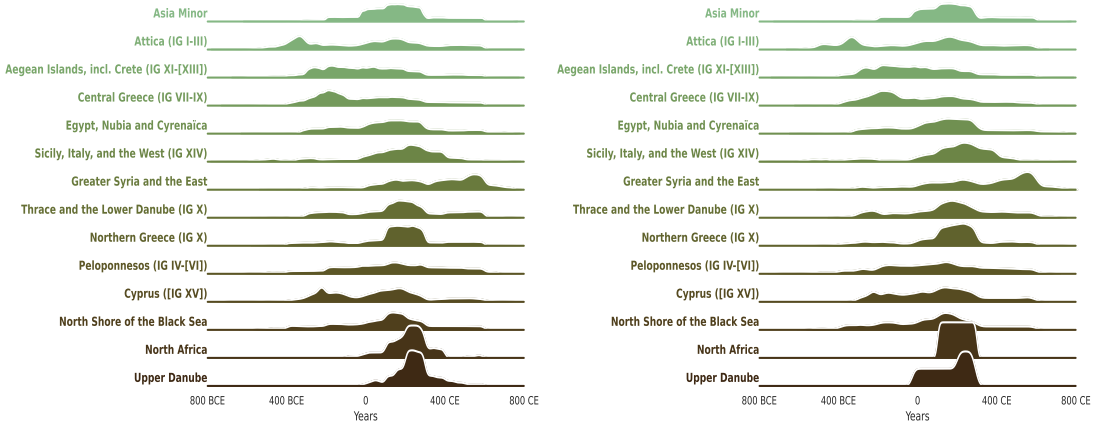


Figure 6: Year distribution of each region for: (a) the training set and (b) the test set.

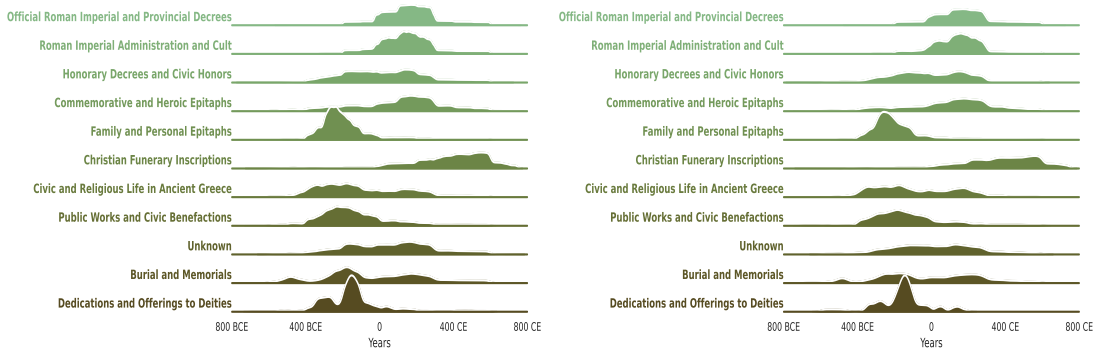


Figure 7: Year distribution of each topic for: (a) the training set and (b) the test set.