Language Models for the Restoration of Latin Legal Manuscripts

Shibingfeng Zhang[1], **Edoardo Caraffa**[1], **Annafelicia Zuffrano**[1], **Maddalena Modesti**[1], **Giovanni Colavizza**[1,2]
[1]Department of Classical Philology and Italian Studies, University of Bologna,
[2]Centre for Digital and Computational Humanities, University of Copenhagen
{shibingfeng.zhang, annafelicia.zuffran2, maddalena.modesti3, giovanni.colavizza}@unibo.it,
edoardo.caraffa2@studio.unibo.it

Abstract

The collection of historical notarial documentation from Bologna is a valuable source, providing deep insights into the city's institutional, legal, and socio-economic history. However, many of these manuscripts have sustained physical damage during centuries of conservation, rendering the text incomplete. To address this, we explored the restoration of these Latin notary documents using encoder-based pre-trained language models (PLMs) under the assumption that the length of missing text is known by estimation from the physical damage. We address the structural misalignment between the physical lacuna of the manuscript and the subword tokenization schemes of PLMs by designing an iterative decoding strategy to align model predictions with the known physical dimensions of lacuna. We also compared the efficacy of monolingual versus multilingual pre-training. Our strategy significantly outperforms baselines consist of standard decoding methods. Furthermore, stratified analysis across different text sections reveals that while monolingual models achieve better performance in general, multilingual models show a suggestive advantage in lexically dense segments, though this finding is not statistically significant. Overall, the best performance achieved by our method is a Hit@1 rate of 35.47% in the short-span setting and 18.75% in the long-span setting. While fully autonomous restoration remains an open challenge, our system provides a useful assistive tool for paleographers.

Keywords: Text Restoration, Diplomatic Documents, Pre-trained Language Models, Digital Paleography, Low-resource Languages

1. Introduction

The archival heritage of medieval Bologna represents a valuable resource to study the city's history. Since 1937, when Giorgio Cencetti proposed the creation of the Codice diplomatico bolognese, a comprehensive edition intended to systematically transcribe and preserve the city's medieval documentary heritage, research has progressed significantly, with the successful academic publication of 10th and 11th century documents (Cencetti, 1977; Feo, 2001) and the development of a 12th century notarial prosopography (Modesti, 2012). However, the physical preservation of these records is inconsistent. Many extant manuscripts suffer from lacuna that obscure critical historical data.

Text restoration, the challenge of inferring missing content within a damaged text, therefore becomes a potential helper for paleographers and archaeologists. In the context of historical manuscript restoration, the text length can be inferred from the size of the physical damage. While leveraging pre-trained language models (PLMs) has become a popular approach for this task (Lazar et al., 2021; Assael et al., 2025), applying them to digitized medieval notary documents presents unique challenges. This paper focuses on restoring the missing text of digitalized notary documents. We adopt encoder-based models pre-trained on the masked language modeling (MLM) objective, as this pre-training task is very similar in nature to text restoration.

When performing text infilling, the MLM head of encoder-based models outputs strictly one token per mask, which is ideal for text restoration with known text length, as it provide the full control over the length of predicted text. However, there are two primary challenges to address: First, while we can estimate how many characters the lacuna covers using its physical size, it maps to an unknown number of subword tokens. Second, physical damage often occurs at arbitrary positions that do not align with the PLM's tokenization scheme. This "slicing" of words often creates fragments that differ from the standard subwords seen during pre-training, leading to a structural misalignment that the MLM head is not natively designed to handle.

We raise the first research question:

- **RQ1:** When using encoder-based PLMs for the text restoration task, how do we effectively resolve the misalignment between the lacuna and the tokenization scheme of the PLMs?

For low-resource languages, multilingual training sometimes demonstrates benefits for system performance (Conneau et al., 2020). In this paper, we adopt a monolingual PLM pre-trained only on Latin and its multilingual counterpart pre-trained on Latin, Ancient Greek, and English. This setting allows us to raise the second research question:

- **RQ2:** Does the multilingual PLM have an advantage over its monolingual Latin counterpart

Proceedings of LT4HALA 2026 @ LREC 2026, pages 361–367.
11 May 2026.

on the text restoration task?

For RQ1, we designed an iterative decoding strategy in order to resolve the misalignment issue. Experiments demonstrate the effectiveness of our method, and we also conducted an ablation study to investigate further. For RQ2, we annotated the test set by dividing it into different functional components, then conducted a thorough analysis to study the performance of both PLMs on each component. Our contributions in this paper are threefold:

1. We proposed a decoding strategy to adapt encoder-based PLMs to scenarios of text restoration with known character length.
2. We conduct experiment to analyze the impact of multilinguality of PLMs in text restoration of Latin.
3. We conduct a study on text restoration task of different text components of diplomatic text. As far as we know this is the first time this technology is applied onto this genre of text.

2. Related Works

Several studies explore settings similar to ours, adopting PLMs where the system input includes both the textual context and the estimated character length of the missing text. For example, Lazar et al. (Lazar et al., 2021) investigated the restoration of Ancient Akkadian inscriptions, where they estimated the number of missing signs and trained BERT from scratch using each sign in the Akkadian vocabulary as a token. Assael et al. (Assael et al., 2025) trained a decoder-based model from scratch for the text restoration of Latin and Ancient Greek inscriptions, using an extra feed-forward head to predict the length of missing text given the context. This method treats each character as a token to ensure the alignment of the output and the predicted length of the missing text. These works resolved the length alignment issue at its root by adapting the tokenization scheme and training the model from scratch; however, this approach is more computationally expensive, requires more data, and effectively reduces the input context window. This may not be an issue for the restoration of inscriptions, as texts on hard materials tend to be shorter and have a lower requirement for a large context window. However, their method is deemed unsuitable for our task since notary documents tend to be longer.

Riemenschneider et al. (Riemenschneider and Frank, 2023) proposed several PLMs specifically for Latin. Their work introduces two groups of models: the first consists of encoder-based models utilizing the RoBERTa (Zhuang et al., 2021) architecture, the second comprises encoder-decoder models based on the T5 architecture (Raffel et al., 2020). All models were trained from scratch. Specifically, the first group includes LaBERTa (Latin only) and PhilBERTa (Latin, Ancient Greek, and English). The second group includes LaTa (Latin only) and PhilTa (Latin, Ancient Greek, and English). These multilingual variants were designed to investigate the impact of cross-linguistic data on historical language PLMs. Across a series of tasks that focused on semantic, syntactic, and morphological aspects of Latin, the monolingual models yielded similar results with no statistically significant differences in most tasks comparing to their multilingual counterpart. The authors suggested that these findings might be due to the relatively small size of their test set, leaving the impact of multilingual pre-training an open question.

In this study, we focus on the first group, as the encoder-only architecture allows us to formulate the restoration as a masked language modeling task rather than an open-ended text generation task. This approach provides greater control over the output length, enabling us to better align the predictions with the physical dimensions of the lacuna. We did not alter the original Byte-Pair Encoding tokenization strategy of the RoBERTa-based models, as doing so would require training from scratch. Instead, we proposed several decoding strategies for text restoration that do not require altering the original structure of the model. This method is presented in Section 3. We evaluate both LaBERTa and PhilBERTa to analyze the effects of multilinguality specifically on the text restoration.

3. Methodologies

Our proposed text restoration method operates under the assumption that the character length $L_{missing}$ of the missing text is known or estimated through physical analysis of the manuscript. We designed a four-stage decoding process to bridge the gap between language model's tokenization scheme and character-level length constraint. Figure 1 demonstrate the process of text restoration method:

- **Stage 1: Lacuna Expansion**
  To ensure compatibility between the physical lacuna and the model's Byte-Pair Encoding tokenization scheme, we first perform *boundary expansion*. If a lacuna begins or ends mid-word, we expand the lacuna to the nearest whitespace-delimited token boundaries. The partial strings covered by such expansion are treated as a known prefix and suffix.

  The target length is adjusted accordingly: $L_{adjusted} = L_{missing} + L_{prefix} + L_{suffix}$.

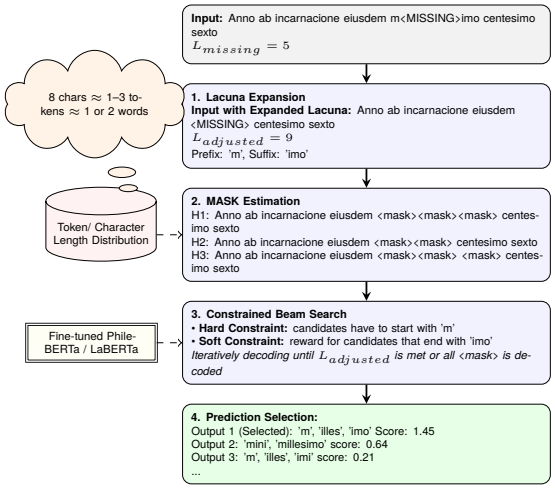

Figure 1: Text restoration workflow

- **Stage 2: Mask Estimation**
  Since encoder-based models maintain a strict one-to-one mapping between input $[MASK]$ tokens and output tokens, we must estimate the optimal number of masks N to represent $L_{adjusted}$ characters. Because BPE token lengths vary, we calculate the prior distribution of token counts for a given character length from the training set. We select the top k most frequent mask distributions. This results in k distinct input sequences, each representing a different hypothesis of the underlying token structure:

- **Stage 3: Prefix&Suffix Constrained Beam Search**
  We employ an iterative decoding strategy based on Beam Search, modified with two specific constraints:

  1. **Prefix Constraint:** During the first decoding step, the candidates are restricted to tokens that begin with the known prefix string.
  2. **Suffix Scoring Reward:** In the final decoding steps, we apply a bonus to the log-probability of sequences that terminate with the known suffix.

- **Stage 4: Prediction Selection** Decoding terminates once all $[MASK]$ positions are filled or the character length $L_{adjusted}$ is reached. Finally, the resulting sequences are aggregated from the k different input, and the candidate with the highest log-probability is selected as the optimal restoration.

4. Experiments

Our dataset comprises 1,184 notarial documents, primarily sourced from the Archivio di Stato di Bologna. These records were split into training, development, and testing sets according to an 8:1:1 ratio. On average, each document contains approximately 380 words. These deeds, written in Latin by 254 notaries, document various types of legal transactions such as sales, donations, exchanges, emphyteusis, wills, and other patrimonial instrumenta.

From a diplomatic point of view, it is also worth noting that the text of these documents is rigidly formalized in terms of discourse construction. In general, the content of a complete notary document can be partitioned into the protocol, the text body, and the eschatocol. There is also a special kind of document called rogationes, which are summary notes of the essential elements of the deed, drawn up by the notary in view of the subsequent drafting in extended form. An exemplary document is provided in Figure 2 along with its translation to provide a more concrete understanding of these concepts.

The LaBERTa and PhilBERTa models were fine-tuned using a standard MLM objective, incorporating an early-stopping strategy to prevent overfitting. To account for the non-deterministic nature of deep learning, we trained five independent checkpoints for both models and report the mean and confidence interval of our results.

For the evaluation of our model, we compare three variants of decoding strategies:

- **Proposed Iterative Decoding with Lacuna Expansion:** The full strategy including lacuna expansion, distribution-based mask estimation, and prefix&suffix constraints as described in Section 3.

- **Iterative Baseline (No Lacuna Expansion):** An iterative decoding strategy for comparison that skips the lacuna expansion step. In this variant, the number of mask tokens is estimated by dividing the missing text length by the average character length per token.

- **Non-Iterative Baseline:** A simple, single-pass decoding strategy. Similar to the iterative baseline, it estimates mask counts based on the average character length per token. This serves as a baseline for comparison, as it's a standard practice of text infilling with multiple masks.

To evaluate the model's restoration capabilities in a detailed way, a paleographer annotated the test set to delineate this structure. The 110 documents in the test set were categorized as follows:

- **Complete with rogationes(25):** documents that contain all three partitions and their corresponding rogationes.

Protocol
Latin: In nomine sancte et individue Trinitatis. Anno Domini millesimo centesimo trigesimo secundo, pridie kalendas aprelis, indicione decima.
English: In the name of the holy and indivisible Trinity. In the year of God one thousand one hundred and thirty-six, the day before the calends of April, tenth indiction.

Text Body
Latin: Ego quidem Rainerius filius Lamberti de Beio hoc donacionis instrumento presenti die dono in honore Dei et ecclesie Sancti Victoris et tibi donno Alberio priori eiusdem ecclesie tuisque fratribus et successoribus proprium in perpetuum conducticium unde pertinuerit, id est omne quod ego habeo et teneo et michi pertinet iure vel actione a radice montis Sancti Victoris usque ad crucem de Dilvino ab Aposa usque ad rivum ex illa parte Barbiani et peciam unam terre aratorie in loco ubi dicitur Castellioni prope cruce de Piro cum ingressu et egressu suo usque in via publica et cum omnibus super se et infra se habentem in integrum. Finis vero eius: ab uno latere a sero et uno capite a meridie possidet Albertus de Rigiza, alio latere a mane detinet Rusticus de Emma terra de socru sua, alio capite ab aquilone adest via publica et si qui alii affines sunt; omnium quod infra hos fines michi pertinet in integrum pro remedio anime mee meeque uxoris nec non et patris et matris mee, in presenti dono et trado atque concedo supradicte ecclesie et tibi donno Alberio priori tuisque fratribus ac successoribus ad habendum, tenendum ac possidendum et quicquid tibi tuisque fratribus ac successoribus deinceps placuerit ad utilitatem eiusdem ecclesie faciendum. Ut nullam litem nullamque controversiam deinceps a me vel a meis heredibus quolibet modo aliquo in tempore vos vel vestri successores de cetero sustineatis ab omni quoque homine prescriptas res legittime defendere et auctorizare tibi et tuis heredibus promitto. Et si ego vel mei heredes prelibatam meam donationem sinplicem in totum vel pro parte audaci nisu quandoque infringere temtavero et eam semper inviolatam custodire noluero, penam triginta denariorum Lucensium libras tibi vel tuis successoribus dare promitto et insuper hanc donationem semper intactam conservare promitto.
English: I, Rainerius, son of Lambertus of Beio, in the present day, give by this document of donation in honor of God and of the church of Saint Victor and to you, father Alberius, prior of the aforesaid church, and to your brothers and successors, full property in perpetuity, wherever it should have been pertained, which is everything I own and possess and belongs to me, by right and action, from the foot of the mountain of Saint Victor as far as the cross of Dilvino, and from the Aposa as far as the river on that side of Barbianum, and moreover one plot of arable land in the place called Castellione, near the cross of Pero, with its right of entry and exit as far as the public road, and with everything it has above and below it, in its entirety. Moreover, its boundaries are: on the western and southern side, the land owned by Albertus of Rigiza, on the eastern side, held by Rusticus of Emma, the land of his mother-in-law, on the northern side, the public road, and all other boundaries, if there are any other. Everything that I own within these boundaries, in the present day, I give and deliver and grant in his integrity, for the salvation of my soul, and of my wife, and of my father and mother, to the aforesaid church and to father Alberius, prior, and to his brothers and successors, so that they may have, hold and possess it, and use until it will be of any utility for the aforesaid church. So that no arguing nor dispute hereafter from me or from my heirs, in any way or at any time, you or your successors shall have to endure, I promise to rightfully defend and authorize the aforesaid property against everyone for you and your heirs. And if I or my heirs shall at any time, by bold attempt, try to violate this my simple donation in whole or in part, or if I shall not wish to keep it always inviolate, I promise to pay to you or your successors the penalty of thirty denarii of Lucca, and moreover I promise to preserve this donation always intact.

Eschatocol
Latin: Actum in canonica Sancti Iohannis in Monte, indicione predicta. Prenominatus Rainerius hoc donacionis instrumentum ut supra legitur scribere rogavit. Iohannes presbiter de ecclesia Sancta Tecla, Lambertus investitor filius Petri de Leo, Lambertus de Auria, Petrus filius Alberti de Vivelinda, Grimaldus filius Bonifantini de Sancto Rofillo, Eldus de Verona rogati sunt testes. Gerardus tabellio hoc donacionis instrumentum ut supra legitur scripsi et firmavi.
English: Done in the canonry of Saint John on the Mount, in the aforesaid indiction. The aforesaid Rainerius requested that this instrument of donation, as read above, may be written. Iohannes, presbyter of the church of Saint Tecla, Lambertus investor, son of Petrus of Leo, Lambertus od Auria, Petrus, son of Albertus of Vivelinda, Grimaldus, son of Bonifantinus of Saint Rufillo, Eldus of Verona are called as witnesses. I, Gerardus notary, wrote and signed this instrument of donation, as read before.

Rogationes
Latin: Pridie kalendas aprelis, indicione x. Testis Iohannes presbiter de Sancta Tecla et Lambertus investitor filius Petri Leo et Lambertus de Auria et Petrus filius Alberti de Vivelinda et Eldus de Verona, Grimaldus filius Bonifantini de Sancto Rofillo. Cartulam donationis fecit Rainerius filius Lamberti de Beio pro remedio anime sue et de uxore sua et patris et matris sue in honore Dei et ecclesie Sancti Victoris et donno Alberio priori eiusdem ecclesie suisque fratribus ac successoribus de omnibus iuris et actionibus quod sibi pertinet a pede montium Sancti Victoris usque ad crucem Dilvini a Aposa usque ad rivum ex illa parte Barbiani et insuper peciam unam terre aratorie prope crucem de Pero in loco qui dicitur Castellioni sub pena et defensione.
English: The day before the calends of April, tenth indiction. Witnesses Iohannes, presbyter of Saint Tecla, and Lambertus investor, son of Petrus Leo, and Lambertus of Auria and Petrus, son of Albertus of Vivelinda and Eldus of Verona, Grimaldus, son of Bonifantinus of Saint Rufillo. Rainerius, son of Lambert of Beio, made a charter of donation for the salvation of his own soul, and of his wife, and of his father and mother, in honor of God and of the church of Saint Victor and of father Alberius, prior of the aforesaid church, and his brothers and successors, regarding all the rights and actions that belong to him from the foot of the mountain of Saint Victor as far as the cross of Dilvino, and from the Aposa as far as the river on that side of Barbianum, and moreover one plot of arable land near the cross of Pero, in the place called Castellione, under penalty and warranty.

Figure 2: Example of a notary document, produced in 1132, archive shelfmark *Archivio di Stato di Bologna, Corporazioni religiose soppresse, S. Giovanni in Monte, 2/1342 n. 10a*

- **Complete with no rogationes (68):** documents that contain the three primary parts but lack rogationes.
- **Incomplete (17):** documents that consist of only one or two parts.

5. Results

We performed the text restoration task using two types of span length settings: short spans, ranging between 5-10 characters, and long spans, ranging between 10-20 characters. To facilitate evaluation of the model's restoration capabilities across different sections (protocol, text body, eschatocol, and rogationes) of the notarial text,no spans were allowed to cross section boundaries. It is worth noting that these evaluation examples were created artificially by removing text from the notarial documents, as we lack ground-truth data for real damages in the documents.

Tables 2 and 4 present the Character Error Rate (CER), Overlap Score, and Hit@1 rate for various decoding strategies. Tables 3 and 5 provide the stratified results of iterative decoding performance across different sections of the notary documents. The Overlap Score is defined as follows:

$$\text{Overlap Score} = \frac{|LMS|}{|LMS| + |\Delta_{gt}| + |\Delta_p|}$$

where $|LMS|$ is the length of the longest matched substring between the ground truth and the prediction, $|\Delta_{gt}|$ represents the length of the non-overlapping segment of the ground truth, $|\Delta_p|$ represents the length of the non-overlapping segment of the prediction.

To ensure a fair comparison, we stripped off the prefix and suffix from predictions generated by the iterative decoding with lacuna expansion method. Consequently, the ground truth remains consistent across all three evaluated decoding methods.

The results in Tables 2 and 4 demonstrate the significant advantage of expanding lacuna to align the lacuna boundary with the PLM's tokenization scheme across all evaluation metrics. Without this alignment method, the performance of the standard iterative approach remains marginal. Specifically, in the short-span setting, the results are nearly indistinguishable from simple parallel decoding. While the standard iterative approach shows a slight improvement over simple decoding in long-span settings, it still falls considerably short of the performance achieved when the lacuna expansion is applied. Notably, the multilingual PhilBERTa model did not demonstrate an advantage over the monolingual LaBERTa model in general. Its performance consistently lagged across all decoding modes and exhibited higher variance.

To illustrate the model's behavior qualitatively, Table 1 presents several examples of LaBERTa's predictions in different decoding modes. By expanding the damage boundary and aligning it with the model's tokenization scheme, the model tends to

produce coherent predictions and, in some cases, offers reasonable alternatives to the ground truth. In contrast, the other two decoding modes tend to generate output that is less convincing grammatically and semantically.

As for text type specific results of iterative decoding method with lacuna boundary expansion, as demonstrated in Tables 3 and 5, both models followed similar performance patterns across sections. Performance on the rogationes and eschatocol sections tended to be the lowest, whereas performance on the protocol and text body are better.

Interestingly, while falling behind in the protocol, text body, and eschatocol sections, PhilBERTa surpasses LaBERTa by a marginal advantage within the rogationes segments. Although this advantage is not statistically significant, as the confidence intervals of the two models' metrics overlap considerably, it suggests that the multilingual training of PhilBERTa may offer specific benefits in this context. A defining characteristic of the rogationes is that it serves as a short draft by the notary to summarize the notary document's content. It is more information-packed than the other sections by nature. To investigate this further, we calculated the lexical density of various document parts using the Type-Token Ratio (TTR) (Laurs, 2024) and Hapax Legomena. Here we define Hapax Legomena as words that occured only once in the specific text type. As shown in Table 6, the rogationes type exhibits both the highest TTR and the highest Hapax Legomena-to-token ratio across all document segments, confirming its high lexical richness. Given this distinct linguistic profile, one might speculate that the multilingual training of PhilBERTa could technically offer advantages in a high lexical richness and non-formulaic context, although the present results do not provide statistically significant evidence for this, and the observed difference remains inconclusive.

In general, the results of both models are far from perfection, with the CER ranging from 39.51% to 72.83% depending on the text type.These results suggest that while the model cannot yet automatically restore text with full autonomy, it holds significant potential as an assistive tool for paleographers.

6. Conclusions

In this paper, we addressed the challenging task of restoring historical notary documents in Latin language by leveraging encoder-based pre-trained language models. We investigated two primary research questions concerning the technical misalignment between physical lacuna and subword tokenization (RQ1) of pre-trained language models and the potential benefits of multilingual pre-training for this specific text genre (RQ2). We addressed the misalignment issue through an iterative decoding strategy with lacuna expansion, which significantly outperformed baselines. Our comparative analysis showed that while the monolingual LaBERTa model generally excelled in most sections, the multilingual PhilBERTa demonstrated marginal advantage in the lexically dense scenario, though this difference was not statistically significant. Although current error rates suggest that fully autonomous text restoration is not yet feasible, the model's performance highlights its value as an assistive tool for paleographers.

A limitation of this study is that our evaluation assumes perfect knowledge of the missing text length, as we used the exact number of characters removed during reference. In a real-world application, length would be estimated from the physical dimensions of the damage, introducing potential errors. Future work will focus on utilizing visual inputs of manuscripts to automatically estimate the character length of lacunae and investigate the impact of estimation precision on model output. Additionally, we plan to experiment with generative language models, such as those proposed by Riemenschneider et al. (Riemenschneider and Frank, 2023).

7. Acknowledgements

This study is funded by FutureData4EU project (Grant Agreement n. 101126733). FutureData4EU is Co-Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or REA. Neither the European Union nor the granting authority can be held responsible for them.

The authors acknowledge the CINECA award under the ISCRA initiative for the availability of high-performance computing resources.

8. 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*, 645(8079):141–147.

Giorgio Cencetti. 1977. *Notariato medievale bolognese. Tomo I: Scritti di Giorgio Cencetti*. Number 3, 1 in Studi storici sul notariato italiano. Consiglio nazionale del notariato, Roma.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek,

Input Text	Label	Decoding Mode		
		Iterative Boundary	Iterative	Simple
...ego quidem donno Paulus Dei gratia abbas Sancti Stefani pres[SPAN]ie iu re enphiteosin vendo tibi Dominico filio Alberti de Oliveti et...	enti d	**enti d**,enti ven,enti iure,enti enph,enti in d,enti do et,enti present,ente d,enti di e,enciarum	entiarum,en,enci,encia,encio, on,encium,enti,es,esi	entiarum
Peto a te donno Guido abbas monasterii Sancti Stephani qui v[SPAN]ti me Rodulfo filio Arnaldi seu filiis et heredibus meis aut cui ego...	ocatur Hierusalem u	**ocatur Hierusalem u**,ocatur in Claterna u,ocatur Gerusalem uti pro,ocatur in Strada u,ocatur in Strata u,ocatur in Musiliano u,ocatur Gerusalem uti a,ocatur in Musciliano u,ocatur Iehrusalem u,ocatur Ierusalem uti dign	itate Hierusalem uti,itate Ierusalem uti,itate vocatur Hierusalem,itate vocatur Ierusalem,. 19. Pe,. 32. Pe,. 29. Pe,itate mona a,itate dicitur Hierusalem,itate Hierusalem pro	itate Hierusalem preen
Et si ego vel mei su[SPAN]mnia firmiter semper ob servare noluerimus, pene nomine decem solidos Lucensium dare vobis et vestris heredibus pro mitto...	cessores predicta o	**cessores predicta o**,cessores prelibata o,cessores vel fratres vel,cessores hanc enphiteos,cessores vel mei heredes,cessores omnia predicta,cessores de cetero sustin,cessores vel ego vel mei,cessores vel mei successores,essores predicta omnia firmiter	essores predicta omnia,essores predicta pe,essores omnia predicta,essores predicta li,essores predicta instru,essores predicta ho,essores prescripta,cessores predicta omnia,essores predicta omni,essores predicta le	essores omnia omni

Table 1: Some examples of LaBERTa's output using different decoding modes. Predictions are separated by comma.

Decoding Mode	Model	CER	Overlap	HR@1
Iterative Boundary	LaBERTa	$\mathbf{55.40}_{\pm 1.09}$	$\mathbf{56.08}_{\pm 0.53}$	$\mathbf{35.47}_{\pm 1.04}$
	PhilBERTa	$63.25_{\pm 2.63}$	$48.70_{\pm 2.71}$	$28.56_{\pm 2.80}$
Iterative	LaBERTa	$\mathbf{72.30}_{\pm 0.88}$	$\mathbf{26.67}_{\pm 0.71}$	$\mathbf{1.87}_{\pm 0.31}$
	PhilBERTa	$75.34_{\pm 1.13}$	$23.12_{\pm 1.31}$	$1.32_{\pm 0.21}$
Simple	LaBERTa	$\mathbf{72.38}_{\pm 0.86}$	$\mathbf{26.31}_{\pm 0.65}$	$\mathbf{1.80}_{\pm 0.24}$
	PhilBERTa	$75.83_{\pm 1.15}$	$22.48_{\pm 1.24}$	$1.15_{\pm 0.20}$

Table 2: Overall Performance, span length 5-10 characters (Mean $\pm$ 95% CI, HR stands for hit-rate)

Text Type	Model	CER	Overlap	HR@1	HR@3	HR@5	HR@10
Protocol	LaBERTa	$\mathbf{46.93}_{\pm 2.71}$	$\mathbf{61.84}_{\pm 1.31}$	$\mathbf{41.03}_{\pm 2.10}$	$\mathbf{51.70}_{\pm 4.42}$	$\mathbf{56.39}_{\pm 4.82}$	$\mathbf{62.99}_{\pm 3.37}$
	PhilBERTa	$48.18_{\pm 4.91}$	$59.66_{\pm 4.17}$	$40.21_{\pm 6.14}$	$48.00_{\pm 4.43}$	$53.30_{\pm 5.39}$	$59.46_{\pm 5.18}$
Text Body	LaBERTa	$\mathbf{49.84}_{\pm 1.63}$	$\mathbf{60.91}_{\pm 1.29}$	$\mathbf{38.98}_{\pm 2.09}$	$\mathbf{51.21}_{\pm 1.37}$	$\mathbf{54.90}_{\pm 1.23}$	$\mathbf{58.81}_{\pm 0.84}$
	PhilBERTa	$63.36_{\pm 2.99}$	$49.23_{\pm 2.90}$	$28.49_{\pm 3.37}$	$41.48_{\pm 4.76}$	$44.80_{\pm 5.11}$	$48.53_{\pm 4.26}$
Eschatocol	LaBERTa	$\mathbf{62.38}_{\pm 3.04}$	$\mathbf{50.48}_{\pm 1.44}$	$\mathbf{32.81}_{\pm 1.92}$	$\mathbf{46.66}_{\pm 2.19}$	$\mathbf{50.05}_{\pm 0.84}$	$\mathbf{53.34}_{\pm 1.85}$
	PhilBERTa	$66.11_{\pm 5.36}$	$46.49_{\pm 4.15}$	$27.62_{\pm 3.81}$	$40.21_{\pm 3.24}$	$43.37_{\pm 3.28}$	$48.20_{\pm 4.51}$
Rogationes	LaBERTa	$76.77_{\pm 4.34}$	$38.09_{\pm 1.78}$	$16.27_{\pm 2.46}$	$\mathbf{29.57}_{\pm 3.82}$	$33.33_{\pm 3.72}$	$\mathbf{37.42}_{\pm 3.42}$
	PhilBERTa	$\mathbf{73.22}_{\pm 6.27}$	$\mathbf{38.22}_{\pm 2.52}$	$\mathbf{16.80}_{\pm 2.51}$	$28.72_{\pm 2.76}$	$\mathbf{33.79}_{\pm 4.00}$	$36.95_{\pm 4.25}$
Overall	LaBERTa	$\mathbf{55.40}_{\pm 1.09}$	$\mathbf{56.08}_{\pm 0.53}$	$\mathbf{35.47}_{\pm 1.04}$	$\mathbf{48.27}_{\pm 1.01}$	$\mathbf{52.01}_{\pm 0.93}$	$\mathbf{56.09}_{\pm 0.80}$
	PhilBERTa	$63.25_{\pm 2.63}$	$48.70_{\pm 2.71}$	$28.56_{\pm 2.80}$	$40.88_{\pm 3.62}$	$44.55_{\pm 3.76}$	$48.83_{\pm 3.54}$

Table 3: Stratified Iterative Boundary Results, span length 5-10 characters(Mean $\pm$ 95% CI, HR stands for hit-rate)

Decoding Mode	Model	CER	Overlap	HR@1
Iterative Boundary	LaBERTa	$\mathbf{55.22}_{\pm 0.95}$	$\mathbf{41.18}_{\pm 0.61}$	$\mathbf{18.75}_{\pm 1.32}$
	PhilBERTa	$62.40_{\pm 3.65}$	$33.52_{\pm 2.99}$	$12.97_{\pm 2.69}$
Iterative	LaBERTa	$\mathbf{63.98}_{\pm 0.60}$	$\mathbf{27.66}_{\pm 0.47}$	$\mathbf{1.15}_{\pm 0.17}$
	PhilBERTa	$67.53_{\pm 1.57}$	$23.50_{\pm 1.15}$	$1.10_{\pm 0.29}$
Simple	LaBERTa	$\mathbf{65.03}_{\pm 0.54}$	$\mathbf{25.39}_{\pm 0.58}$	$\mathbf{0.82}_{\pm 0.12}$
	PhilBERTa	$68.92_{\pm 1.64}$	$21.50_{\pm 1.02}$	$0.79_{\pm 0.08}$

Table 4: Overall Performance, span length 10-20 characters (Mean $\pm$ 95% CI, HR stands for hit-rate)

Text Type	Model	CER	Overlap	HR@1	HR@3	HR@5	HR@10
Protocol	LaBERTa	$\mathbf{39.51}_{\pm 3.63}$	$\mathbf{52.47}_{\pm 1.97}$	$\mathbf{27.84}_{\pm 3.73}$	$\mathbf{36.91}_{\pm 5.02}$	$\mathbf{43.90}_{\pm 5.82}$	$\mathbf{49.14}_{\pm 5.43}$
	PhilBERTa	$42.94_{\pm 4.66}$	$49.14_{\pm 3.52}$	$24.74_{\pm 3.73}$	$35.07_{\pm 4.05}$	$39.22_{\pm 4.78}$	$44.44_{\pm 5.08}$
Text Body	LaBERTa	$\mathbf{50.93}_{\pm 1.73}$	$\mathbf{45.56}_{\pm 0.93}$	$\mathbf{21.62}_{\pm 1.21}$	$\mathbf{31.00}_{\pm 1.85}$	$\mathbf{34.38}_{\pm 1.92}$	$\mathbf{37.51}_{\pm 2.27}$
	PhilBERTa	$62.20_{\pm 4.37}$	$34.07_{\pm 3.67}$	$13.13_{\pm 3.13}$	$21.14_{\pm 3.08}$	$23.58_{\pm 3.89}$	$26.45_{\pm 3.67}$
Eschatocol	LaBERTa	$\mathbf{63.70}_{\pm 0.92}$	$\mathbf{34.21}_{\pm 1.23}$	$\mathbf{14.55}_{\pm 1.55}$	$\mathbf{21.62}_{\pm 1.83}$	$\mathbf{23.95}_{\pm 2.70}$	$\mathbf{26.17}_{\pm 3.28}$
	PhilBERTa	$67.45_{\pm 4.76}$	$29.46_{\pm 3.52}$	$10.65_{\pm 3.52}$	$17.73_{\pm 3.96}$	$20.47_{\pm 5.14}$	$23.60_{\pm 3.94}$
Rogationes	LaBERTa	$74.11_{\pm 1.18}$	$\mathbf{22.83}_{\pm 1.43}$	$3.47_{\pm 1.48}$	$\mathbf{7.81}_{\pm 3.45}$	$\mathbf{11.70}_{\pm 6.16}$	$\mathbf{13.64}_{\pm 4.48}$
	PhilBERTa	$\mathbf{73.23}_{\pm 2.11}$	$22.63_{\pm 1.43}$	$\mathbf{4.00}_{\pm 2.62}$	$6.52_{\pm 3.46}$	$9.03_{\pm 3.60}$	$11.55_{\pm 3.13}$
Overall	LaBERTa	$\mathbf{55.22}_{\pm 0.95}$	$\mathbf{41.18}_{\pm 0.61}$	$\mathbf{18.75}_{\pm 1.32}$	$\mathbf{27.22}_{\pm 2.02}$	$\mathbf{30.78}_{\pm 2.05}$	$\mathbf{33.81}_{\pm 2.11}$
	PhilBERTa	$62.40_{\pm 3.65}$	$33.52_{\pm 2.99}$	$12.97_{\pm 2.69}$	$20.65_{\pm 2.94}$	$23.39_{\pm 3.94}$	$26.58_{\pm 3.08}$

Table 5: Stratified Iterative Boundary Results, span length 10-20 characters(Mean $\pm$ 95% CI, HR stands for hit-rate))

Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Giovanni Feo, editor. 2001. *Le carte bolognesi del secolo XI*. Number 53 in Fonti per la storia d'Italia medievale. Istituto storico Italiano per il Medio Evo, Roma.

Thomas Laurs. 2024. Towards a readability formula for latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024*, pages 170–175.

Koren Lazar, Benny Saret, Asaf Yehudai, Wayne Horowitz, Nathan Wasserman, and Gabriel Stanovsky. 2021. Filling the gaps in ancient akkadian texts: a masked language modelling approach. *arXiv preprint arXiv:2109.04513*.

Maddalena Modesti. 2012. *Studi per l'edizione delle carte bolognesi del secolo XII: prosopografia dei notai ed edizione critica di due cartulari notarili*. Pàtron Editore, Bologna.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Frederick Riemenschneider and Anette Frank. 2023. Exploring large language models for classical philology. *arXiv preprint arXiv:2305.13698*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Tag	Type-Token Ratio	Hapax-Token Ratio
Protocol	0.137	6.8%
Eschatocol	0.223	14.8%
Text Body	0.129	6.7%
Rogationes	0.310	19.7%

Table 6: Lexical richness and Hapax Legomena-to-token ratio across notary document sections.

9. Language Resource References