Detecting Latin in Historical Books with Varied Layout: A Multimodal Benchmark

Anonymous ACL submission

Abstract

This paper presents a novel task of extracting Latin fragments from mixed-language historical documents with varied layouts. We benchmark and evaluate the performance of large foundation models against a multimodal dataset of 753 annotated pages. The results demonstrate that reliable Latin detection with contemporary LLMs is achievable. Our study provides the first comprehensive analysis of these models' capabilities and limitations for this task[1].

1 Introduction

Accurate language identification at a granular level within historical documents is a key component to the study of the early modern period at scale. Latin, as the primary written language of Western Europe for more than a millennium, has a unique position, gradually ceding to vernaculars at varying paces across regions and genres (Marjanen et al., 2025). Throughout this transition, Latin fragments frequently appeared within predominantly vernacular texts, in quotations, specialist terminology, and instances of code-switching. Automated extraction of diverse Latin uses in context from historical corpora is crucial to studying language evolution, the interplay between classical and modern thought, and the dissemination of ideas (Sprugnoli et al., 2024; Gorovaia et al., 2024; Perrone et al., 2021; Burns et al., 2021). However, this task poses challenges due to wide variations in Latin usage, scripts, complex page layouts, and inconsistent print and scan quality in historical book databases.

This study focuses on detecting instances of Latin within the *Eighteenth Century Collection Online* (ECCO) (Tolonen et al., 2022) corpus using both book page images and the corresponding OCR text. The lack of an existing dataset specifically designed for multimodal and code-mixed Latin detection motivated us to create an annotated dataset for this purpose. Our dataset contains 753 pages sampled from historical documents, validated by specialists in 18th-century publishing culture to represent diverse use cases. While we focus on Latin due to its aforementioned importance for historical study, our fully benchmarked, manually annotated scenario provides a solid template for extending the method to other languages as well.

Given the complex nature of the task and factors ranging from OCR noise to varied print layouts, we explore the capabilities of modern Large Language Models (LLMs), including multimodal models (MLLMs) for this task. The way these models handle contextual information, recognize patterns in noisy data, and integrate textual cues with visual layout information has been found to help in disambiguating text in historical documents (Luo et al., 2024; Boros et al., 2024; Kanerva et al., 2025; Xie et al., 2025).

Our investigation exploits the new dataset to test a number of state-of-the-art models and approaches, and finds that reliable Latin detection in such challenging historical material is achievable. The benchmarking of different model architectures provides insight into their strengths and weaknesses when faced with the complexities inherent in the data. This work establishes a strong baseline for a novel NLP task and at the same time highlights the need for modality-aware approaches and robust evaluation frameworks in historical text analysis.

The main contributions of this article are:

- Defining the task of Latin detection in historical documents, an understudied multimodal case of language detection.
- Creating an expert-annotated dataset from 18th-century books, capturing diverse and challenging examples of Latin usage.

[1]The dataset is in the supplementary materials. Both the dataset and code will be published upon acceptance.

- Developing an evaluation framework for this task, considering challenges from both textual and visual modalities.
- Conducting a systematic comparison of contemporary LLMs to assess their efficacy for this task.
- Delivering a practical Latin detection pipeline, demonstrating its readiness for downstream applications in historical research.

2 Problem Definition

We define the task of *Latin language detection in historical documents* as a two-stage classification and extraction problem, where the input consists of a scanned page image and/or its OCR transcription. The task is to automatically detect whether any segments in the text are written in Latin, and if so, to extract text of those specific segments.

Formally, given a document page D, let I_D denote its image and T_D denote its OCR'd text. A system must perform the following two subtasks:

- **Task 1 (Page-level Latin Detection):** Predict a binary label $y_D \in \{0, 1\}$, where $y_D = 1$ indicates that the page contains at least one segment in Latin, and $y_D = 0$ otherwise.
- **Task 2 (Latin Segment Extraction):** If $y_D = 1$, extract a list of text spans $S_D = [s_1, s_2, \ldots, s_n]$, where each $s_i \in T_D$ is a contiguous Latin segment string.

The reason for specifying the task in two stages is mainly due to evaluation. As our final aim is to evaluate how well Latin is identified at a high level of granularity across different layouts, we measure task 2 in terms of per-page precision and recall. However, this type of measurement does not adequately cater to cases where Latin is not present on a page. Thus, task 1 has been designed to give us this more general information on e.g., whether the models are excessive in detecting Latin where there is none. Requiring segment-level output as strings rather than image regions aligns better with tasks most MLLMs are trained on, and enables simpler performance comparisons between the input modalities. More details on the metrics used will be given in section 5.

3 Related work

Latin in NLP Given its historical importance, Latin has attracted considerable attention within the NLP community (e.g., Sprugnoli et al., 2024; Schulz and Keller, 2016; Sprugnoli et al., 2022; Gorovaia et al., 2024; Perrone et al., 2021; Burns et al., 2021), though much of this research has centered on small, clean corpora of ancient literary texts. While some recent studies have ventured into Early Modern mixed-language documents (Stüssi and Ströbel, 2024; Volk et al., 2024), these also predominantly rely on manually curated and annotated data. In contrast, our work focuses on the foundational task of Latin *discovery*: detecting Latin within extensive, unedited, and noisy digitized collections like ECCO (Tolonen et al., 2022). This computational approach aims to detect Latin in a vast corpora, while the identified fragments can subsequently be analyzed using a range of established NLP tools developed for classical languages (Johnson et al., 2021; Burns, 2023; Straka and Straková, 2020; Kupari et al., 2024).

Code-mixed Language Detection From a methodology perspective, identifying Latin segments within British publications is a code-mixed language detection task (Aguilar et al., 2020). While extensive research in this area has focused on contemporary informal texts (Barman et al., 2014; Zhang et al., 2018), its application to historical documents, with challenges like archaic syntax, lexicon, and spelling, has been less explored (Schulz and Keller, 2016; Volk et al., 2022). Detecting classical languages in these complex historical contexts has traditionally involved rule-based systems and supervised machine learning approaches, notably Conditional Random Fields (CRFs) (Schulz and Keller, 2016; Sterner and Teufel, 2023; Volk et al., 2022). Alongside these, robust statistical tools like Lingua (Stahl, 2021) offer effective general language identification with support for mixed language. Given the recognized potential of modern LLMs to navigate linguistic nuances and noisy data, our work investigates their capacity to enhance detection performance.

LLMs for Historical Documents Recent LLMs, particularly Multimodal variants (MLLMs), have shown considerable potential in historical document analysis, demonstrating top performance in tasks like OCR, named entity recognition, and general document understanding from historical sources (Bai et al., 2025; Luo et al., 2024; Boros et al., 2024; Kanerva et al., 2025; Backer and Hyman, 2025; Xie et al., 2025), and in assessing general historical knowledge (Hauser et al., 2024). De-

spite these advancements, a significant gap persists for more specialized, complex applications. Specifically, there is a notable lack of dedicated benchmarks and systematic exploration for the fine-grained, page-level multimodal detection and extraction of embedded secondary languages (e.g., Latin) (Aguilar et al., 2020; Guzmán et al., 2017). This task is demanding due to noisy scans from historical archives, diachronic language context, and orthographic variation (Volk et al., 2022). Our work contributes to this underexplored area by introducing a systematic evaluation methodology designed to be scalable also to other languages.

4 Dataset

4.1 Sampling and Annotations

Our approach to dataset construction began with a targeted sampling strategy to identify pages with a high likelihood of containing Latin text. We queried the Reception Reader database (Rosson et al., 2023), which indexed text reuses across the ECCO corpus using noise-resistant detection methods. From this, we randomly selected 800 reuse instances where one book was cataloged as Latin and the other as non-Latin. To ensure broad representation and reflect the diversity of the ECCO collection (approximately 200,000 books), each sampled page was drawn from a different book, covering varied publication dates and topics. However, ECCO's language metadata is book-level, meaning that "Latin" books often contain significant non-Latin text like English introductions. Also, the reuse offsets only mark the textual overlap without specifying the language of the text segment.

These pages were then manually annotated. Annotators were tasked with drawing bounding boxes around all Latin fragments on the page images (see example in Figure 1). These visual annotations could later be reliably mapped to text offsets (locations within a string) using ECCO's OCR positional data for ground truth text extraction. Our annotation guidelines stated marking all instances of Latin text semantically used as Latin. This included single Latin words if presented with explanations in a dictionary, as well as Latin found in headlines, editorial annotations, or footnotes. Conversely, to ensure clarity and consistency, in-line Roman/Latin names (of people, places, plants) and jargon were not specifically annotated as Latin but were treated as part of the surrounding language, as well as abbreviations, such as 'etc'.

The annotation environment was Label Studio (Tkachenko et al., 2020-2025). The primary annotation was performed by three scholars familiar with Latin. Following this, an expert in historical texts meticulously reviewed and validated all annotations to ensure accuracy and consistency.

4.2 Dataset Characteristics

In total, 753 pages were annotated, with 623 identified as containing Latin. An expected finding during annotation was the frequent presence of other languages, such as French, German, and Greek, highlighting the dataset's challenging multilingual nature beyond simple Latin-English code-mixing.

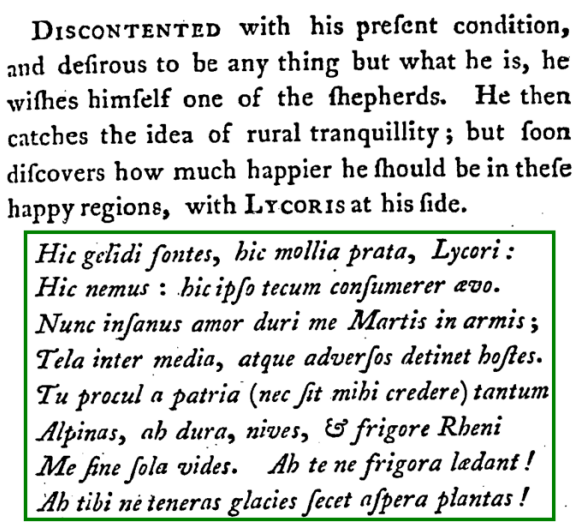
To contextualize model performance and to better understand the dataset's composition, we divided the sample pages into language integration categories. Each category represents a specific way in which Latin is used in 18th-century British books and how it relates to English-language text. Depending on their content, pages containing Latin text were assigned to one or multiple categories, with some frequently appearing together (e.g., bilingual editions and footnotes), while others are mostly exclusive (e.g., full Latin). By categorizing our evaluation dataset in this way, we can assess the performance of our Latin detection approach within different contexts of language integration. The following list defines the categories used:

1. Full Latin: Pages entirely written in Latin, without any other language present.
2. Bilingual Editions (Latin / English): Pages that contain both the original Latin text and its English translation.
3. Direct quotations in Latin: Pages where Latin phrases or sentences are quoted verbatim within an otherwise English text.
4. Independent Latin sections: Pages with original Latin text by the author, accompanied by English text on the same page.
5. Code Switching: Pages where the text alternates between Latin and English within the span, often for stylistic or rhetorical purposes.
6. Dictionaries: Pages with entries that define individual Latin words, often with translations or explanations in another language.
7. Footnotes: Pages with annotations or footnotes that provide explanations of Latin words or phrases used in the main text.

Table 1 shows the frequency of each annotation category within our dataset. Figures 3 and 4 in

	Category	Count
1.	Full Latin	134
2.	Bilingual Editions	65
3.	Direct quotes	258
4.	Independent Latin sections	169
5.	Code-switching	109
6.	Dictionaries	19
7.	Footnotes	69

Table 1: Frequencies of annotation categories.


DISCONTENTED with his preſent condition, and deſirous to be any thing but what he is, he wiſhes himſelf one of the ſhepherds. He then catches the idea of rural tranquillity; but ſoon diſcovers how much happier he ſhould be in theſe happy regions, with LYCORIS at his ſide.

Hic gelidi fontes, hic mollia prata, Lycori:
Hic nemus: hic ipſo tecum conſumerer ævo.
Nunc inſanus amor duri me Martis in armis;
Tela inter media, atque adverſos detinet hoſtes.
Tu procul a patria (nec ſit mihi credere) tantum
Alpinas, ah dura, nives, & frigore Rheni
Me ſine ſola vides. Ah te ne frigora lædant!
Ah tibi ne teneras glacies ſecet aſpera plantas!

Figure 1: An example of an annotated Latin fragment.

the Appendix A show category examples. This form of page-level annotation reduces the need for costly instance-level labeling, which is particularly challenging for Latin due to expertise requirements and high annotation volume. Moreover, it supports context-rich evaluation aligned with the page-level structural nature of our detection task inputs.

5 Evaluation Setting

5.1 OCR Post-Correction and Normalization

Evaluating models on the ECCO corpus is complicated by significant OCR quality discrepancies: modern models with vision capabilities often produce cleaner text than ECCO's original OCR, while text-based models may or may not replicate the noise in their input. Such differences make direct string-based comparisons problematic and distort evaluation. To ensure meaningful assessment across all model types, we post-correct both the ground-truth Latin segments and the full input page texts. This OCR post-correction is performed using the OpenAI o1 model (Jaech et al., 2024) with a specialized prompt from (Kanerva et al., 2025).

Even after the post-correction, residual noise and other variation still remain in the data. Thus, for token-based evaluation, we apply a more traditional rule-based pre-processing pipeline to both predicted and reference strings. This deterministic pipeline, informed by our extensive experience with OCR data and domain-specific knowledge, targets common superficial textual variations. The pipeline includes Unicode normalization, ligature replacement, lowercasing, digit removal, dehyphenation, and punctuation stripping. Subsequent to these cleaning operations, the strings are tokenized into word-level units. More details of the processing steps could be found in the Appendix B.1.

5.2 Metrics

The goal of Task 1 is to detect whether a page has Latin on it. We measure this by reporting precision, recall, and F1-score in percentage. To evaluate the Latin segment extraction performance in Task 2, we calculate precision, recall, and F1 score in percentage based on token-level matches between model predictions and ground truth. A fuzzy matching mechanism is applied to pair predicted and reference tokens one by one. A match is considered valid if the token-level edit-distance is not larger than a threshold proportion θ compared to the ground-truth token length, which serves as a tunable hyperparameter. This approach accounts for minor lexical variations and OCR-induced distortions in single tokens, offering a more flexible and robust evaluation compared with exact token matching The pseudocode of the fuzzy matching algorithm is shown in the Appendix B.2. Overall metrics are averaged across the full evaluation set.

6 Method

Our evaluation investigates the application of general instruction-following LLMs, particularly multimodal variants, for Latin segment extraction from historical documents. We propose a unified, prompt-based pipeline designed to be both practical for real-world deployment and robust for systematically and fairly evaluating the capabilities of diverse LLMs on this task.

Unified Prompting Strategy We employ a single, high-level instructive prompt designed to elicit responses that inherently address both sub-tasks within a unified and simply formatted output. This approach simplifies interaction with the models and

the subsequent processing of their outputs, thereby contributing to the overall ease of application.

This unified prompt asks the LLM to extract all Latin segments to a list, including single Latin words, without further instruction. The distinction in our experiments lies solely in the input provided to this consistent prompt, where the specific prompts are shown in the Appendix C:

- **Text-only input:** The LLM receives the OCR-extracted and post-corrected text, appended to the prompt.
- **Image-only input:** The LLM receives the page image, with the prompt guiding it to identify Latin script based on visual signal.
- **Multimodal input:** The LLM receives both the scanned page image and the post-corrected OCR-extracted text, allowing it to leverage both visual and textual information for the task.

Structured Output and Postprocessing The LLMs are instructed to output their predictions as a list of Latin segments, which directly corresponds to the output requirement for Task 2. The presence of a non-empty list implicitly indicates the presence of Latin script on the page (Task 1, $y_D = 1$), while an empty list indicates its absence ($y_D = 0$).

Model-Agnostic Compatibility Because the method does not rely on any model-specific architecture or training, it can be directly applied to a wide range of general-purpose foundation language models. This makes the approach particularly suitable for scalable deployment across large historical corpora with variable OCR quality and image-text alignment conditions.

7 Experiments

7.1 Experiment Setup

Model Selection To explore how modern instruction-tuned language models handle the new task of Latin segment detection and extraction in noisy multimodal historical documents, we benchmark a representative suite of LLMs across modalities, scales, and architectures. The model selection follows three guiding principles: (i) in the absence of dedicated multimodal benchmarks for historical language understanding in documents, we refer to leaderboard performance on **DocVQA** (Tito et al., 2021) and comprehensive open evaluations such as **OpenCompass** (Contributors, 2023) and **MMMU** (Yue et al., 2024); (ii) we prioritize lightweight to medium-scale models (7B-32B) to better reflect realistic research use cases in historical academic and low-resource scenarios. Specifically, the most notable selected models include (additionally, we evaluate further models detailed in the Appendix E.2):

- **Qwen2.5-VL series** (32B, 7B) (Bai et al., 2025): flagship open-source multimodal LLMs with strong document understanding and fine-grained visual grounding. We include both 32B and 7B variants, and ablate visual inputs via text-only configurations (Qwen2.5-32B/14B/7B).
- **InternVL3 series** (14B, 9B, 8B) (Zhu et al., 2025): notable academic multimodal LLMs with a two-stage visual encoder adding transformer design. We ablate the visual inputs in the 9B model via testing its pure language counterpart Internlm3-8B-Instruct.
- **Mistral-Small-3.1-24B-Instruct** (Mistral AI, 2025): an efficient model using Mixture-of-Experts (MoE). Notably, for its strong text performance and reasoning capabilities supports an extended 128k token context window.
- **DeepSeek-R1-Distill variants** (Qwen-32B, Qwen-14B, Llama-8B) (DeepSeek-AI, 2025): pure-text distilled models focused on reasoning, probing extraction capability with thinking capability but without visual cues.

Baseline We employ **Lingua** (Stahl, 2021), a statistical language identifier based on character n-gram modeling, and the only off-the-shelf tool we found that supports token-level Latin detection in mixed-language text. While not designed for noisy OCR, it offers a useful traditional baseline to contextualize the difficulty of our task and the potential advantages and drawbacks of LLM-based approaches. More configuration of the baseline is in Appendix D.

Implementation Details All LLMs are deployed on a unified inference backend using the `vLLM` framework (version 0.8.5) with server mode to simulate realistic client-side usage. Inference is performed on a 4×A100 (40GB) GPU node, with no data parallelism and a batch size of 1 for all cases, resulting in an average of 16 GPU hours per model. We use each model's default generation parameters without further tuning. For the Mistral model, we specifically load the Mistral-Small-3.1-24B-Instruct-2503 version checkpoint. The edit-

distance threshold θ is set to 0.2 based on empirical evidence, discussed further in the Appendix E.3.

7.2 Main Results and Comparisons

The main experimental results are shown in Table 2. The traditional Lingua language identifier proves to be a surprisingly strong baseline. In the page-level detection task (Task 1), Lingua achieves a strong F1 score of 92.89. In the more challenging segment extraction task (Task 2), it still attains 76.74 F1 using simple token-level identification.

Only the largest instruction-tuned foundation models such as Qwen2.5-VL-32B and DeepSeek-R1-Distill-32B consistently outperform Lingua across both tasks. Qwen in particular demonstrates strong document-oriented OCR capabilities, excelling even in vision-only settings. Interestingly, the DeepSeek-R1-Distilled version of Qwen does worse than plain Qwen on Task 1, but better on Task 2. Looking in more details at the numbers, the distilled version seems to be simply optimising for recall at the cost of precision on the page-level metric, while then being more accurate in the actual segment extraction. Smaller models, like InternVL3-14B, can still improve over Lingua when vision is used to guide text modeling, but fail in vision-only configurations, illustrating that OCR robustness is not easy to achieve.

Performance tends to improve with model scale, especially within the same model family. Larger models have been shown to more effectively memorize and generalize low-resource language phenomena, consistent with neural scaling laws (Gordon et al., 2021; Kaplan et al., 2020). However, size alone does not guarantee strong performance. Some smaller models illustrate that fusion architecture and fine-tuning strategy can outweigh raw scale. Both Qwen2.5-VL-7B and InternVL3-8B share the Qwen2.5-7B language backbone, yet Qwen2.5-VL-7B's distinct native dynamic image resolution support and instruction-tuning on multilingual document-centric corpora gain advantages. A direct comparison between Qwen2.5-7B and InternLM3-8B reveals that Qwen's textual pretraining data composition and cross-lingual instruction tuning confer superior robustness to OCR noise and low-resource Latin patterns, beyond what can be achieved by architecture scale alone.

Across all model families, multimodal variants (I+T) consistently outperform their unimodal counterparts under the same model size. This is especially evident in the InternVL3 series, where I-only models fall well below their I+T variants on both tasks. This highlights the value of aligning visual features with textual reasoning. However, text-only models can still perform competitively, particularly those with fine-grained OCR-style pretraining, such as Qwen2.5. Additionally, most models cannot extract Latin solely from images, where only Qwen2.5-VL-32B matches its text-only performance, further highlighting historical document OCR challenges and our dataset contributions on dealing with OCR texts and annotations.

7.3 Performance by Category

Figure 2 shows the results on the five top models considering the coverage of different input modalities, over different page categories, which were specified in Section 4.2. As can be seen from the left-hand side of the figure, for pages with only a single type, there is a large disparity between category difficulty: models yield almost perfect performance for full-Latin and bilingual pages while struggling with code-switching and to a lesser degree with dictionaries. The difference between modalities is also evident: image and combined inputs remain robust across categories, while text-only models demonstrate more variety. In particular, text-based Qwen fails in code-switch detection: it does not find any correct Latin text at all in 5 pages out of 9 in this category, showing its shortcomings in nuanced detail extraction. Since footnotes almost never exist in isolation, we are unable to directly report performance relating to them. In addition, considering multi-label cases, the performance metrics should be dominated by the performance on the longer text category. For example, the performance on independent Latin sections plus dictionaries is better than dictionaries alone. However, when looking at the performance of footnotes in conjunction with other short text types such as code-switching and quotations, we can still clearly find that the performance of footnotes is even worse than that of code-switching.

7.4 Behavior on Non-Latin pages

To gain further insight into model behavior, we analyze performance on pages devoid of Latin. For Task 1, we report the recall of the negative class. For Task 2, we report false positive token rates (Table 3). It is evident from these analyses that all models over-detect Latin as compared to our ground-truth. The best performance is demonstrated by Qwen2.5 with text-only input, which

Model Details				Page-level (Task 1)			Token-level (Task 2)		
Model	**Variant**	**Size**	**Mode**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
Lingua	-	-	T	92.89	86.85	**99.84**	76.74	76.64	80.01
Qwen2.5	VL	32B	I+T	94.09	88.97	**99.84**	**84.15**	**86.21**	**84.25**
	VL	32B	I	94.22	89.45	99.52	78.86	81.31	78.97
	-	32B	T	**97.68**	**97.29**	98.07	80.53	84.60	79.46
	-	14B	T	91.54	88.46	94.86	70.12	74.6	72.04
	VL	7B	I+T	89.93	83.87	96.95	73.23	77.1	76.05
	VL	7B	I	92.86	88.04	98.23	71.16	74.38	73.36
	-	7B	T	88.37	83.33	94.05	52.43	53.5	64.17
InternVL3	-	14B	I+T	91.5	84.68	99.52	80.09	81.28	82.3
	-	14B	I	87.37	82.95	92.28	44.48	46.45	46.89
	-	9B	I+T	84.69	82.37	87.14	59.88	59.33	68.09
	-	9B	I	70.82	79.28	63.99	25.28	25.91	29.61
	-	8B	I+T	89.5	83.36	96.62	65.76	61.35	81.79
	-	8B	I	86.67	83.68	89.87	48.58	45.91	59.2
Internlm3	-	8B	T	82.33	83.01	81.67	47.49	52.3	51.32
Mistral-Small-3.1	-	24B	I+T	92.73	88.47	97.43	79.89	79.4	82.93
	-	24B	I	89.82	85.16	95.02	66.14	65.39	69.81
	-	24B	T	88.47	83.26	94.37	72.97	72.61	78.79
DeepSeek-R1-Distill	Qwen2.5	32B	T	94.05	89.42	99.2	81.87	84.37	82.96
	Qwen2.5	14B	T	91.69	85.12	99.36	77.98	81.62	78.82
	Llama-3.1	8B	T	89.74	84.01	96.3	66.59	73.06	68.38

Table 2: Experimental results on selected LLMs, compared with Lingua baseline. "I" indicates image input, "T" indicates text input, and "I + T" refers to the combination of both modalities.

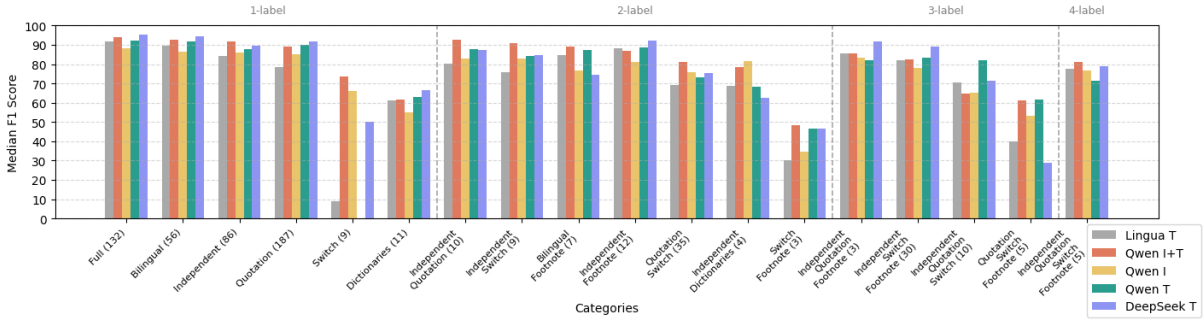


Figure 2: Result on different category labels for 5 top-performing models: median of the token-level F1-score for each page. The number of instances for each label is shown in parentheses. We filter out the subsets with fewer than 3 cases to ensure the statistical validity of the subset results.

correctly refrains from outputting Latin on 87% of the non-Latin pages, with a low proportion of Latin token output on misclassified pages. This could suggest that it benefits from a stronger sensitivity to linguistic context, enabling more cautious predictions in ambiguous cases.

The statistical baseline Lingua does remarkably worse in this evaluation, misidentifying Latin on over 70% of non-Latin pages. This is primarily because the ngram-based model struggles with deeper linguistic nuances and context, and outputs false positive tokens on most pages. Since our goal is to process a collection of 200K books, a large number of false positives would pose a significant problem.

In general, the number of Latin tokens identified on non-Latin pages remains small, pointing to the possibility of using further filtering or thresholding approaches in downstream tasks.

Model Details				Page	Token
Model	**Variant**	**Size**	**Mode**	**Neg. Recall**	**FP Rate**
Lingua	-	-	T	28.24	2.65
Qwen2.5	VL	32B	I+T	41.22	3.41
	VL	32B	I	44.27	1.94
	-	32B	T	**87.02**	**0.53**
DeepSeek	Qwen2.5	32B	T	44.27	4.84

Table 3: Analysis on non-Latin pages: Negative Class Recall (in pages) and False Positive Rate (percentage of tokens falsely identified as Latin on the whole page).

Prompt Config		Page			Token		
Non-Latin	**Abbrev**	**F**	**P**	**R**	**F**	**P**	**R**
		94.09	**88.97**	**99.84**	84.15	**86.21**	84.25
✓		93.37	87.82	99.68	**84.28**	85.34	**85.42**
✓	✓	93.01	87.31	99.52	84.26	85.74	84.89

Table 4: Prompt modification experiments.

7.5 Qualitative Evaluation and Error Analysis

Our qualitative evaluation and error analysis are informed by the ECCO dataset's characteristics and ambiguities in defining Latin. These factors complicated annotation, influenced model predictions, and set practical limits on achievable evaluation scores. For examples and example-oriented descriptions of the issues outlined here, see Appendix E.1.

First, poor image quality in ECCO, along with complex page layouts, such as multi-column text, footnotes, marginalia, and varied fonts or spacing (see Figures 3 and 4, Appendix A), often led to inaccurate OCR and fragmented transcriptions, hampering reliable annotation, adding noise to even the OCR-post-corrected page texts.

The error analysis of Latin-containing segments shows lower performance in categories like code-switching, dictionaries, and footnotes. This is caused by both the brevity of Latin snippets in these contexts and the fact that they are more likely to be affected by severe OCR errors, the latter often due to complex layouts and challenging font types or sizes on poor-quality scans. Figure 6 in Appendix E.1 shows an example of how OCR errors can lead to discrepancies between our ground truth and predictions.

Content-wise, an interesting definitional mismatch is observed, both decreasing precision as well as increasing false positive rates. In our ground-truth annotation guidelines, we specified that Roman named entities, common anglicized Latin phrases (e.g. "e.g.", "etc.", etc.) and jargon within otherwise English text were not to be considered Latin. However, these are frequently identified by the models as Latin (for an example, see Figure 7 in Appendix E.1). In terms of performance, these discrepancies often account for the full difference between ground truth and prediction, meaning that with a slight change in definitions, the model performance would be even stronger.

7.6 Prompt Modification Experiments

To assess the impact of prompt fine-tuning, we conduct a study on the best model (Qwen2.5-VL-32B) by (i) explicitly instructing it to output an empty list when no Latin is detected (Non-Latin column) and (ii) adding explicit instructions to not extract common Latin abbreviations found in other languages (Abbrev column). As shown in Table 4, these modifications yield only marginal changes, indicating small gains in some cases compensated by losses in others. This is particularly interesting concerning the abbreviations, as it means that the model has a strong innate definition of what it considers Latin, which cannot be easily overcome by prompting.

8 Conclusion

This paper presented a novel task and a dataset for Latin extraction from early-modern book pages. We systematically evaluated diverse foundational models and found that this task can be solved with excellent performance, without fine-tuning. However, such performance can be achieved only with bigger models (32B parameters).

Our results show that 94% F1 performance on page-level detection can be achieved with only image input. This has practical implications since a visual model can be used even in collections that have no OCR, or whose OCR is not of sufficient quality. In this case, OCR or post-OCR correction can be performed only on pages preselected by a visual model, which would save computation cost.

We also found that semantic analysis is crucial to distinguish actual Latin phrases from usages of Latin words in English text. Thus we argue for the usage of large foundation models, or a pipeline of visual and textual models, rather than pure statistical methods.

Further steps will include processing the whole ECCO collection and publishing a complete dataset of Latin fragments in the ECCO books. We also plan to expand our results to other Early-modern collections, such as the French BNF collection.

Limitations

All our work is performed on a single collection. Results from this single corpus (18th-century British ECCO texts) should be interpreted with caution, as potential corpus-specific biases may affect broader generalizability, highlighting the importance of future cross-corpus studies. Since Latin has been widely used across Europe, it would be interesting to validate our methods on other collections, e.g., the main texts written in Romance languages, such as French.

A drawback for per-category evaluation in the current dataset is the lack of per-segment type annotations, which leads to a shortage of instance-level performance analysis. These will be implemented in the next version of the dataset.

The current work does not utilize a separate validation set, a decision necessitated by the considerable difficulty and cost of annotating the historical data. Consequently, model selection was performed without an independent dataset for optimization before final testing, instead referring to the results from previous work. Incorporating a validation set is an important consideration for future extensions of this research.

All reported results are from single experimental runs in each model's default configuration. This approach was a necessary compromise due to the significant computational costs and GPU resource constraints associated with evaluating the diverse range of large-scale models investigated. Future work could explore the impact of running variability by conducting multiple trials and tuning seeds, temperatures, and other hyperparameters.

It would also be interesting to compare the off-the-shelf LLMs with some models trained or fine-tuned specifically for the Latin extraction task. However, given the surprising performance of general-purpose LLMs on this task, the practical usefulness of such experiments is questionable.

Ethics Statement

The underlying literary works from which our dataset is derived, sourced from 18th-century texts within the Eighteenth Century Collections Online (ECCO), are in the public domain. The compilation and sharing of our dataset, which comprises annotated excerpts and portions of page images from this collection, are conducted for research purposes under the permissions granted. We are committed to ensuring that the creation and dissemination of this dataset adhere to relevant copyright considerations and ethical guidelines.

We used ChatGPT and Gemini for grammar and spell-checking and stylistic polishing of the draft of this manuscript. All suggestions were critically reviewed and edited by the authors to ensure factual accuracy and originality.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Samuel Backer and Louis Hyman. 2025. Bootstrapping ai: Interdisciplinary approaches to assessing ocr quality in english-language historical documents. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 251–256.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23, Doha, Qatar. Association for Computational Linguistics.

Emanuela Boros, Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, and Frédéric Kaplan. 2024. Post-correction of historical text transcripts with large language models: An exploratory study. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 133–159, St. Julians, Malta. Association for Computational Linguistics.

Patrick J Burns. 2023. Latincy: Synthetic trained pipelines for latin nlp. *arXiv preprint arXiv:2305.04365*.

Patrick J Burns, James A Brofos, Kyle Li, Pramit Chaudhuri, and Joseph P Dexter. 2021. Profiling of intertextuality in latin literature using word embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4900–4907.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Mitchell A Gordon, Kevin Duh, and Jared Kaplan. 2021. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5915–5922.

Svetlana Gorovaia, Gleb Schmidt, and Ivan P. Yamshchikov. 2024. Sui generis: Large language models for authorship attribution and verification in Latin. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 398–412, Miami, USA. Association for Computational Linguistics.

Gualberto A Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for modeling code-switching across corpora. In *Interspeech*, pages 67–71.

Jakob Hauser, Daniel Kondor, Jenny Reddish, Majid Benam, Enrico Cioni, Federica Villa, James Bennett, Daniel Hoyer, Pieter Francois, Peter Turchin, et al. 2024. Large language models' expert-level global history knowledge benchmark (hist-llm). *Advances in Neural Information Processing Systems*, 37:32336–32369.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Kyle P. Johnson, Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. 2021. The Classical Language Toolkit: An NLP framework for pre-modern languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 20–29, Online. Association for Computational Linguistics.

Jenna Kanerva, Cassandra Ledins, Siiri Käpyaho, and Filip Ginter. 2025. Ocr error post-correction with llms in historical documents: No free lunches. *arXiv preprint arXiv:2502.01205*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Hanna-Mari Kristiina Kupari, Erik Henriksson, Veronika Laippala, and Jenna Kanerva. 2024. Improving Latin dependency parsing by combining treebanks and predictions. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 216–228, Miami, USA. Association for Computational Linguistics.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15630–15640.

Jani Marjanen, Tuuli Tahko, Leo Lahti, and Mikko Tolonen. 2025. Book printing in latin and vernacular languages in northern europe, 1500–1800. In *The Hermeneutics of Bibliographic Data and Cultural Metadata*, pages 27–66. National Library of Norway.

Mistral AI. 2025. Mistral-small-3.1-24b-instruct-2503. https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503. Instruction-tuned multimodal model with 24B parameters, 128k context length, and vision-text capabilities.

Valerio Perrone, Simon Hengchen, Marco Palma, Alessandro Vatri, Jim Q Smith, and Barbara McGillivray. 2021. Lexical semantic change for ancient greek and latin. *Computational approaches to semantic change*, 6.

David Rosson, Eetu Mäkelä, Ville Vaara, Ananth Mahadevan, Yann Ryan, and Mikko Tolonen. 2023. Reception reader: Exploring text reuse in early modern british publications. *Journal of Open Humanities Data*, 9(1).

Sarah Schulz and Mareike Keller. 2016. Code-switching ubique est - language identification and part-of-speech tagging for historical mixed text. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 43–51, Berlin, Germany. Association for Computational Linguistics.

Rachele Sprugnoli, Federica Iurescia, and Marco Passarotti. 2024. Overview of the evalatin 2024 evaluation campaign. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024*, pages 190–197.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, Margherita Fantoli, and Giovanni Moretti. 2022. Overview of the EvaLatin 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 183–188, Marseille, France. European Language Resources Association.

Peter M. Stahl. 2021. Lingua: The most accurate natural language detection library for python. https://github.com/pemistahl/lingua-py. Python bindings for the Lingua language detection library.

Igor Sterner and Simone Teufel. 2023. TongueSwitcher: Fine-grained identification of German-English code-switching. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–13, Singapore. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2020. UDPipe at EvaLatin 2020: Contextualized embeddings and treebank embeddings. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 124–129, Marseille, France. European Language Resources Association (ELRA).

Elina Stüssi and Phillip Ströbel. 2024. Part-of-speech tagging of 16th-century Latin with GPT. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 196–206, St. Julians, Malta. Association for Computational Linguistics.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2021. Document collection visual question answering. In *International Conference on Document Analysis and Recognition*, pages 778–792. Springer.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Mikko Tolonen, Eetu Mäkelä, and Leo Lahti. 2022. The anatomy of eighteenth century collections online (ecco). *Eighteenth-century studies*, 56(1):95–123.

Martin Volk, Dominic Philipp Fischer, Lukas Fischer, Patricia Scheurer, and Phillip Benjamin Ströbel. 2024. LLM-based machine translation and summarization for Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 122–128, Torino, Italia. ELRA and ICCL.

Martin Volk, Lukas Fischer, Patricia Scheurer, Bernard Silvan Schroffenegger, Raphael Schwitter, Phillip Ströbel, and Benjamin Suter. 2022. Nunc profana tractemus. detecting code-switching in a large corpus of 16th century letters. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2901–2908, Marseille, France. European Language Resources Association.

Yunting Xie, Matti La Mela, and Fredrik Tell. 2025. Multimodal llm-assisted information extraction from historical documents: The case of swedish patent cards (1945-1975) and chatgpt. In *The 9th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2025), March 5–7, 2025, Tartu, Estonia*, pages 1–15. University of Oslo Library.

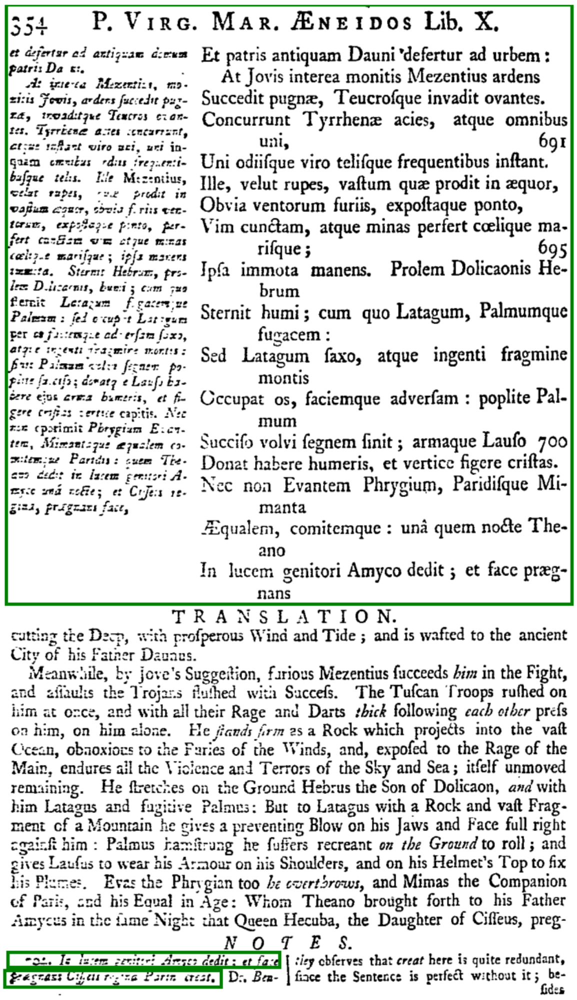
Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Yuan Zhang, Jason Riesa, Daniel Gillick, Anton Bakalov, Jason Baldridge, and David Weiss. 2018. A fast, compact, accurate model for language identification of codemixed text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 328–337, Brussels, Belgium. Association for Computational Linguistics.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Dataset Illustrations

Figures 3 and 4 show examples of Latin text categories. Both figures feature independent Latin text in the main text box at the top of the pages and footnotes at the bottom. Figure 3 is a bilingual edition of a Latin text, with an English translation directly below the Latin text at the top of the page. The footnotes in Figure 4 also include instances of code-switching and direct quotations of Latin text.


354 P. VIRG. MAR. ÆNEIDOS Lib. X.

Et patris antiquam Dauni defertur ad urbem:
At Jovis interea monitis Mezentius ardens
Succedit pugnæ, Teucrofque invadit ovantes.
Concurrunt Tyrrhenæ acies, atque omnibus uni,
Uni odiifque viro telifque frequentibus inftant.
Ille, velut rupes, vaftum quæ prodit in æquor,
Obvia ventorum furiis, expoftaque ponto,
Vim cunctam, atque minas perfert cœlique marifque;
Ipfa immota manens. Prolem Dolicaonis Hebrum
Sternit humi; cum quo Latagum, Palmumque fugacem:
Sed Latagum faxo, atque ingenti fragmine montis
Occupat os, faciemque adverfam: poplite Palmum
Succifo volvi fegnem finit; armaque Laufo
Donat habere humeris, et vertice figere criftas.
Nec non Evantem Phrygium, Paridifque Mimanta
Æqualem, comitemque: unâ quem nocte Theano
In lucem genitori Amyco dedit; et face prægnans

TRANSLATION.

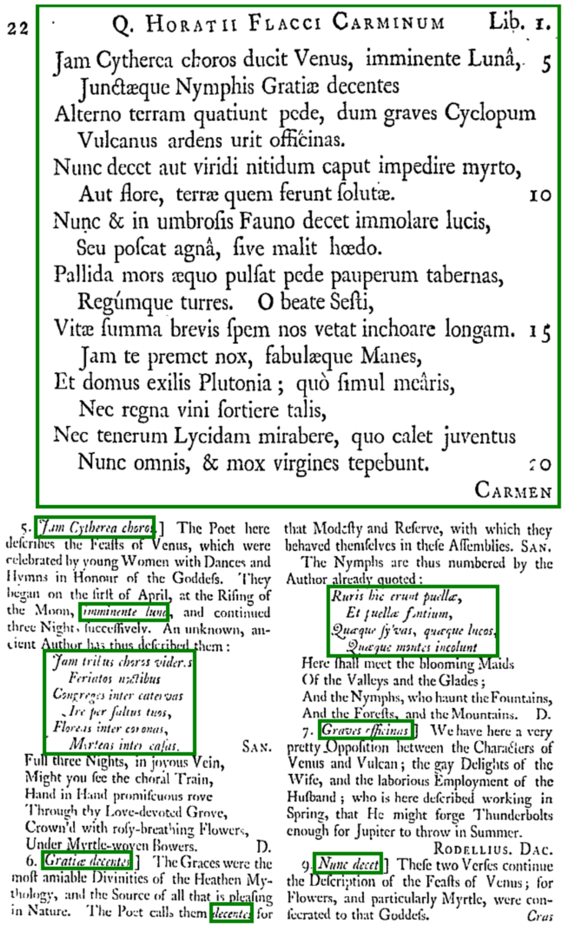
cutting the Deep, with profperous Wind and Tide; and is wafted to the ancient City of his Father Daunus.

Meanwhile, by Jove's Suggeftion, furious Mezentius fucceeds *him* in the Fight, and affaults the Trojans flufhed with Succefs. The Tufcan Troops rufhed on him at once, and with all their Rage and Darts *thick* following *each other* prefs on him, on him alone. He *ftands firm* as a Rock which projects into the vaft Ocean, obnoxious to the Furies of the Winds, and, expofed to the Rage of the Main, endures all the Violence and Terrors of the Sky and Sea; itfelf unmoved remaining. He ftretches on the Ground Hebrus the Son of Dolicaon, *and* with him Latagus and fugitive Palmus: But to Latagus with a Rock and vaft Fragment of a Mountain he gives a preventing Blow on his Jaws and Face full right againft him: Palmus hamftrung he fuffers recreant *on the Ground* to roll; and gives Laufus to wear his Armour on his Shoulders, and on his Helmet's Top to fix his Plumes. Evas the Phrygian too *he overthrows*, and Mimas the Companion of Paris, and his Equal in Age: Whom Theano brought forth to his Father Amycus in the fame Night that Queen Hecuba, the Daughter of Ciffeus, preg-

NOTES.

In lucem genitori Amyco dedit: et face prægnans Cifseis regina Parin creat. Dr. Ben- ... *ley* obferves that *creat* here is quite redundant, fince the Sentence is perfect without it; be-

Figure 3: An example page with Latin fragments.


22 Q. HORATII FLACCI CARMINUM Lib. I.

Jam Cytherea choros ducit Venus, imminente Lunâ,
Junctæque Nymphis Gratiæ decentes
Alterno terram quatiunt pede, dum graves Cyclopum
Vulcanus ardens urit officinas.
Nunc decet aut viridi nitidum caput impedire myrto,
Aut flore, terræ quem ferunt folutæ.
Nunc & in umbrofis Fauno decet immolare lucis,
Seu pofcat agnâ, five malit hœdo.
Pallida mors æquo pulfat pede pauperum tabernas,
Regúmque turres. O beate Sefti,
Vitæ fumma brevis fpem nos vetat inchoare longam.
Jam te premet nox, fabulæque Manes,
Et domus exilis Plutonia; quò fimul meâris,
Nec regna vini fortiere talis,
Nec tenerum Lycidam mirabere, quo calet juventus
Nunc omnis, & mox virgines tepebunt.

CARMEN

5. *Jam Cytherea choros*] The Poet here defcribes the Feafts of Venus, which were celebrated by young Women with Dances and Hymns in Honour of the Goddefs. They began on the firft of April, at the Rifing of the Moon, *imminente lunâ*, and continued three Nights fucceffively. An unknown, ancient Author has thus defcribed them:

Jam tribus choros videres
Feriatos noctibus
Congreges inter catervas
Ire per faltus tuos,
Floreas inter coronas,
Myrteas inter cafas. SAN.

Full three Nights, in joyous Vein,
Might you fee the choral Train,
Hand in Hand promifcuous rove
Through thy Love-devoted Grove,
Crown'd with rofy-breathing Flowers,
Under Myrtle-woven Bowers. D.

6. *Gratiæ decentes*] The Graces were the moft amiable Divinities of the Heathen Mythology, and the Source of all that is pleafing in Nature. The Poet calls them *decentes* for that Modefty and Referve, with which they behaved themfelves in thefe Affemblies. SAN. The Nymphs are thus numbered by the Author already quoted:

Ruris hic erunt puellæ,
Et puellæ fontium,
Quæque fylvas, quæque lucos,
Quæque montes incolunt

Here fhall meet the blooming Maids
Of the Valleys and the Glades;
And the Nymphs, who haunt the Fountains,
And the Forefts, and the Mountains. D.

7. *Graves officinas*] We have here a very pretty Oppofition between the Characters of Venus and Vulcan; the gay Delights of the Wife, and the laborious Employment of the Hufband; who is here defcribed working in Spring, that He might forge Thunderbolts enough for Jupiter to throw in Summer. RODELLIUS. DAC.

9. *Nunc decet*] Thefe two Verfes continue the Defcription of the Feafts of Venus; for Flowers, and particularly Myrtle, were confecrated to that Goddefs. *Cras*

Figure 4: An example page with Latin fragments.

that this applies to the evaluation step only, while Latin extracting models take an input text without these steps.

For each text string, the following sequential operations are performed:

1. **Unicode Normalization:** Each string undergoes Unicode normalization using the `normalize` with "NFKD" method from Python's built-in `unicodedata` module. This step decomposes characters into their canonical forms, for example, separating accents from base characters, which helps in standardizing character representation.

2. **Ligature Replacement:** A predefined set of common ligatures is replaced with their constituent characters. Examples of replacements include 'ﬀ' to `ff`, 'æ' to `ae`, and importantly for some historical contexts, '&' to `et`.

3. **Lowercasing:** All alphabetic characters in the string are converted to lowercase.

B Evaluation Setting Details

B.1 Preprocessing

This section details the text preprocessing pipeline for evaluation, implemented in Python, to normalize both ground truth and predicted text strings before unigram token extraction. The primary goal of this pipeline is to standardize textual representations, thereby mitigating the impact of superficial variations (e.g., from OCR noise or stylistic differences) on downstream metrics calculation. Note,

4. **Digit Removal:** All sequences of digits are removed from the string to avoid prediction ambiguity on digits, e.g., OCR digits in footnote notations.

5. **De-hyphenation (Word Merging):** This step addresses common OCR inconsistencies in handling end-of-line hyphens from historical documents. To ensure textual uniformity for subsequent analysis, word segments that were hyphenated, typically due to line breaks in the original source, are consistently merged into single tokens.

6. **Punctuation Stripping:** All standard punctuation marks, as defined by Python's `string.punctuation` constant, are removed from the string.

7. **Word Tokenization:** After the above cleaning steps, each processed sequence is tokenized into a list of individual words using the `word_tokenize` function from the NLTK library (`nltk.tokenize`, version 3.9.1).

B.2 Fuzzy Matching Algorithm in Token-level Metrics

To evaluate segment correspondence, we apply a fuzzy matching algorithm to compare lists of preprocessed ground truth tokens against predicted tokens for each sample. This approach calculates Precision, Recall, and F1-score while being robust to minor textual variations. The core matching logic is outlined in Algorithm 1.

The algorithm performs a greedy, one-to-one fuzzy match: each predicted token is compared against available ground truth tokens using a match indicator function (`IsFuzzyMatch`) based on edit distance and a predefined proportion threshold θ. A match only holds when the edit distance is less than or equal to θ proportion of the length of the ground truth token string. A ground truth token can only be matched once to ensure an accurate count of distinct true positive matches. This fuzzy approach is beneficial as it offers robustness to minor textual variations that may persist even after preprocessing, leading to a more meaningful evaluation of segment correspondence.

C LLM Prompt Details

This section details the exact prompt templates employed to instruct the LLMs for the task of Latin script detection and extraction. The prompts were adapted based on the input modality being used. In the templates below, the placeholder `{page_text}` indicates where the OCR output corresponding to the processed page image was dynamically inserted.

Image + Text

```
Identify and extract all segments written in Latin (e.g., Classical or Medieval Latin) from the provided image, using the accompanying OCR text as a reference. Include even single-word segments. Return the results as a list of strings in the JSON format: ["text1", "text2", ...].

OCR Text: {page_text}
```

Image-only

```
Identify and extract all segments written in Latin (e.g., Classical or Medieval Latin) from the provided image. Include even single-word segments. Return the results as a list of strings in the JSON format: ["text1", "text2", ...].
```

Text-only

```
Identify and extract all segments written in Latin (e.g., Classical or Medieval Latin) from the OCR text of an image. Include even single-word segments. Return the results as a list of strings in the JSON format: ["text1", "text2", ...].

OCR Text: {page_text}
```

D Configuration of Baseline

For comparative language identification, we employed Lingua (version 2.1.0) (Stahl, 2021) as a baseline. The `LanguageDetector` was specifically configured to operate with a predefined restricted set of eight languages: English, French, German, Greek, Italian, Spanish, Portuguese, and Latin. This selection aims to encompass Latin itself and a set of the most frequently occurring languages within our target corpus ECCO (English,

Algorithm 1 Fuzzy Matching and Token Metrics Output

procedure CALCULATEFUZZYMETRICS($GT\_Tokens, Pred\_Tokens, \theta$)
▷ Input: GT_Tokens, $Pred\_Tokens$ (lists of preprocessed tokens)
▷ θ (edit distance ratio threshold for a match)
▷ Output: Precision, Recall, F1-score
    $TP \leftarrow 0$
    $matched\_gt\_indices \leftarrow \emptyset$
    for each $pred\_token$ in $Pred\_Tokens$ **do**
        for each gt_token in GT_Tokens (with index gt_idx) **do**
            if $gt\_idx \in matched\_gt\_indices$ **then continue**
            end if
            if ISFUZZYMATCH($gt\_token, pred\_token, \theta$) **then**
                $TP \leftarrow TP + 1$
                Add gt_idx to $matched\_gt\_indices$
                break ▷ Current $pred\_token$ matched
            end if
        end for
    end for
    $FP \leftarrow$ length$(Pred_Tokens) - TP$
    $FN \leftarrow$ length$(GT_Tokens) - TP$
    $Precision \leftarrow TP/(TP + FP)$
    $Recall \leftarrow TP/(TP + FN)$
    $F1 \leftarrow 2 \times (Precision \times Recall)/(Precision + Recall)$
    return $Precision, Recall, F1$
end procedure

French, German, and Greek), while also including languages present in ECCO that share orthographic or lexical similarities with Latin (Italian, Spanish, and Portuguese). Including these similar languages was intended to create a more global and robust test scenario for accurate Latin identification in ECCO.

In our pipeline, Lingua's function to detect multiple languages within a given text (`detect_multiple_languages_of` method) was utilized on each page's OCR output. From the resulting language segments identified by Lingua, only those substrings classified as Latin were subsequently extracted for our analysis and evaluation.

E Additional Results

E.1 Qualitative Results

More qualitative results are shown as examples to illustrate the best model's performance and the error modes.

Figure 6 shows an example of a mismatch caused by significant OCR noise caused by poor original image quality. Here, the post-corrected OCR of our ground truth differs so much from the OCR visual or multimodal models produced during the prediction process that not even our edit-distance based fuzzy ground truth matching can recover what is essentially a full match. This kind of error especially affects pages in the footnotes, code switching and dictionary categories, since the Latin texts in these categories tend to be printed in harder to detect fonts and layouts, which are additionally more likely to be affected by bad scan quality.

Figure 7 shows an example of a definitional mismatch between our annotations and the predictions. Although there is no Latin text on the page, the prediction contains the Roman names appearing in the page text.

Figure 8 shows an example of a page where the prediction contains hallucinations. The model took part of the text and translated it into Latin in the prediction, without being prompted to do so.

E.2 Other LLMs

To further contextualize the performance of contemporary foundation models on our challenging Latin discovery task, we evaluated additional models beyond those in the main comparison (Table 2). This section presents results for variants from the Pixtral and Phi-4 families. Specifically, we exam-

ined **Pixtral-12B** (Agrawal et al., 2024), an efficient Mixture-of-Experts (MoE) model, and models from the **Phi-4** family (Abdin et al., 2024), including *Phi-4-Multimodal-Instruct* (5.6B parameters) noted for its smaller footprint and a larger text-only *Phi-4* (14B parameters) variant. The results are detailed in Table 5.

The results in Table 5 reveal varied performance. Notably, some smaller multimodal models, such as Phi-4-Multimodal-Instruct (5.6B) particularly in image-only (I) mode, struggled significantly with the fine-grained token-level extraction task, achieving an F1 score as low as 2.48. Similarly, Pixtral-12B showed a substantial performance drop in its image-only configuration for token-level results. This suggests that factors like smaller parameter counts or training data less attuned to the nuances of documents and noisy OCR may limit the out-of-the-box effectiveness of certain general-purpose models for this specialized task. In contrast, the text-only Phi-4 variant performed more competently, underscoring that model architecture with different input modalities and training focus are critical. These observations highlight the challenging nature of our proposed task.

E.3 Fuzzy Matching Threshold

The fuzzy matching threshold, θ (representing the maximum allowed normalized edit distance relative to ground truth token length), was empirically set to 0.2 for all main experiments. This choice aligns with a common heuristic of tolerating approximately "1 error in 5 characters," suitable for OCR-derived text, and is supported by our sensitivity analysis in Figures 5. It consistently shows that while F1 scores generally increase with θ, the most substantial and steepest F1 score improvements for the majority of evaluated models are concentrated in the range leading up to $\theta \approx 0.2$, effectively compensating for common, fine-grained textual variations attributable to OCR noise. Although metrics may continue to rise beyond this point for some configurations on our dataset, we maintain $\theta = 0.2$ as a principled trade-off. A higher universal threshold could risk over-tolerating more substantial prediction errors beyond typical OCR noise, potentially prioritizing the matching of token quantity or approximate form over precise content fidelity. This could also obscure true output quality differences, especially when comparing models with varying input noise levels (e.g., image-only versus OCR-input systems).

Model Details				Page-level (Task 1)			Token-level (Task 2)		
Model	**Variant**	**Size**	**Mode**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
Lingua	-	-	T	92.89	86.85	99.84	76.74	76.64	80.01
Pixtral	-	12B	I+T	89.46	84.65	94.86	69.70	71.48	74.13
			I	74.77	79.28	70.74	31.33	32.43	33.16
Phi-4	Multimodal	5.6B	I+T	76.46	85.04	69.45	39.53	42.43	46.02
	Multimodal	5.6B	I	40.84	86.91	26.69	2.48	4.13	2.78
	-	14B	T	91.64	86.24	97.75	70.14	74.17	75.12

Table 5: Experimental results for additional evaluated LLMs.

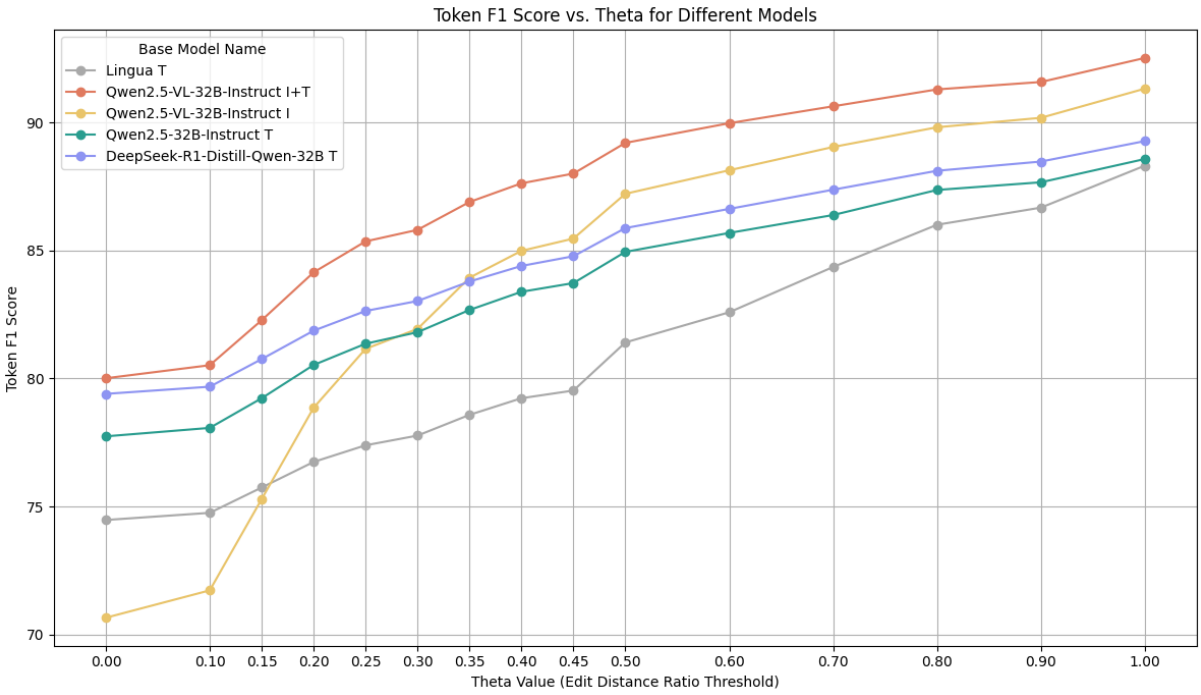


Figure 5: Token F1 scores on different θ value.

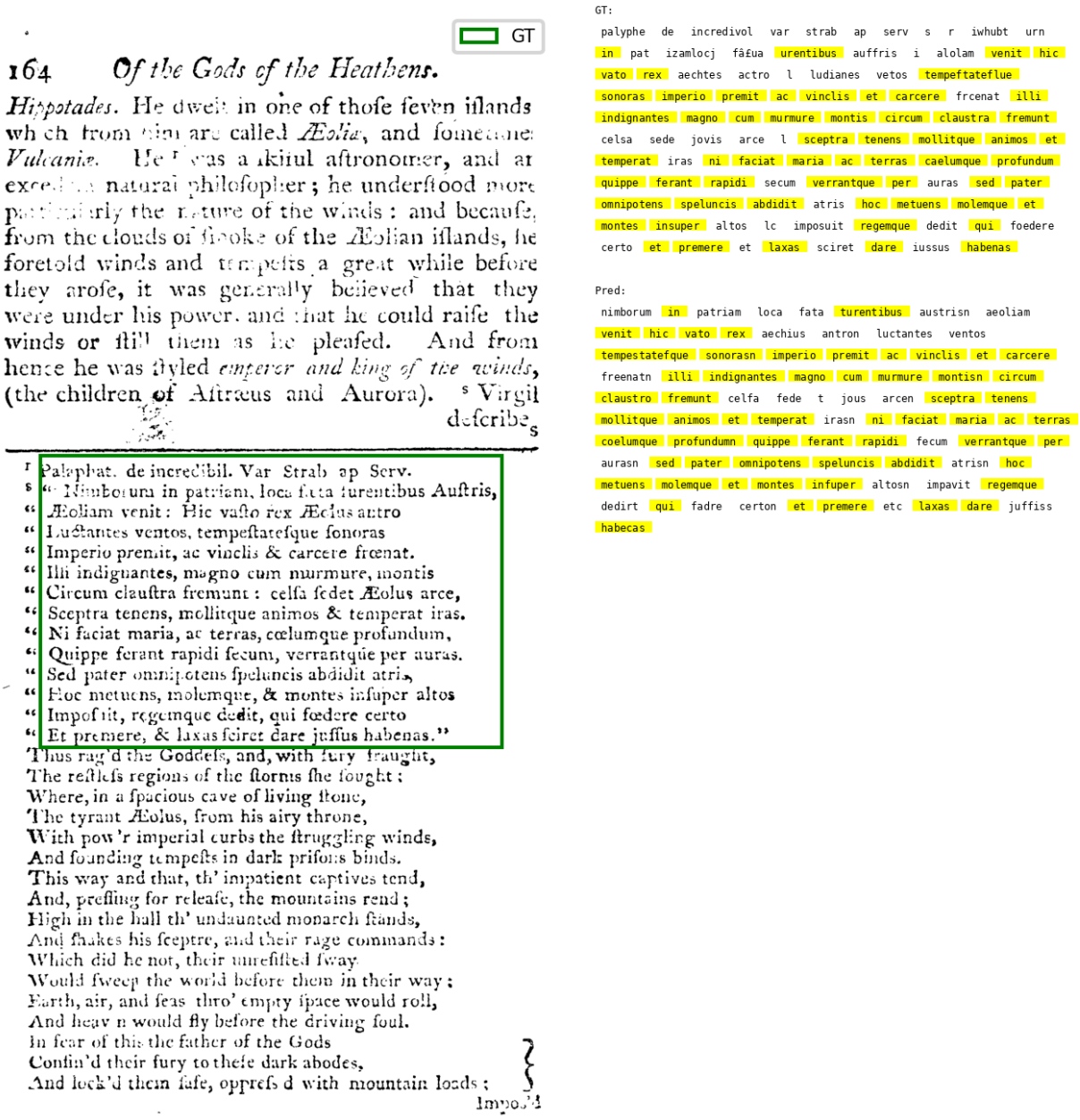
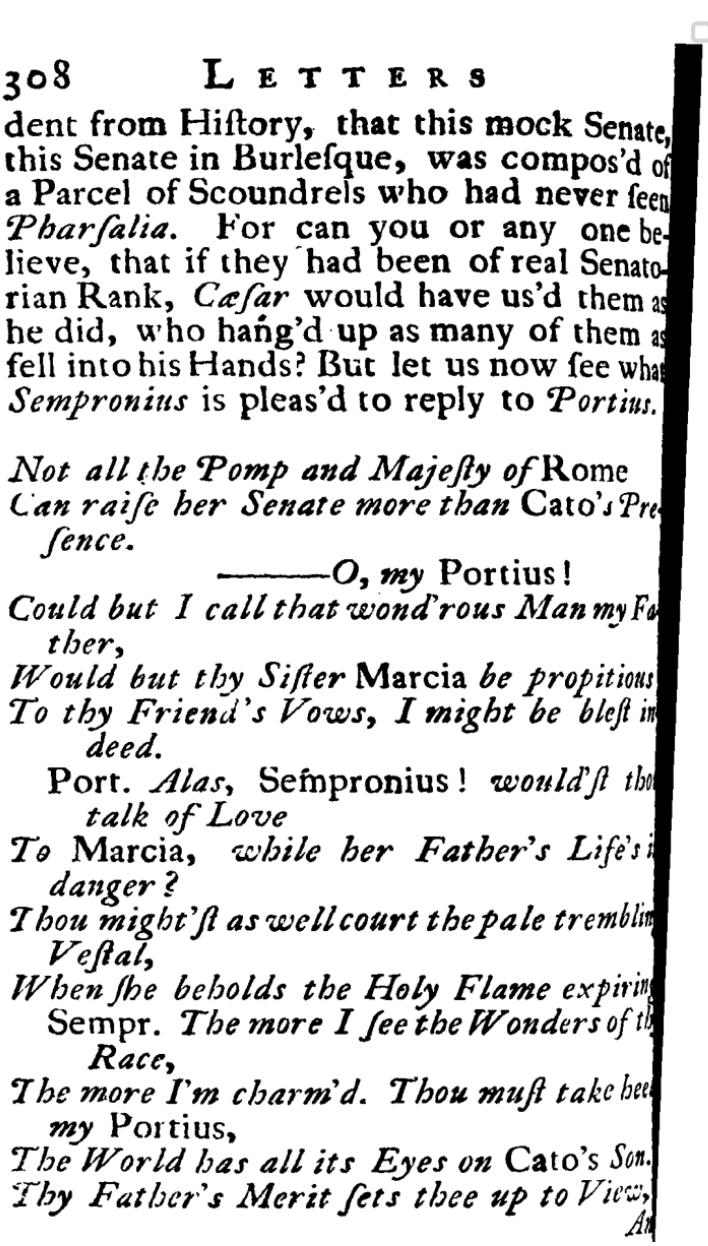


Figure 6: An example page with Latin fragments, together with our ground truth and prediction for that page.



308 LETTERS

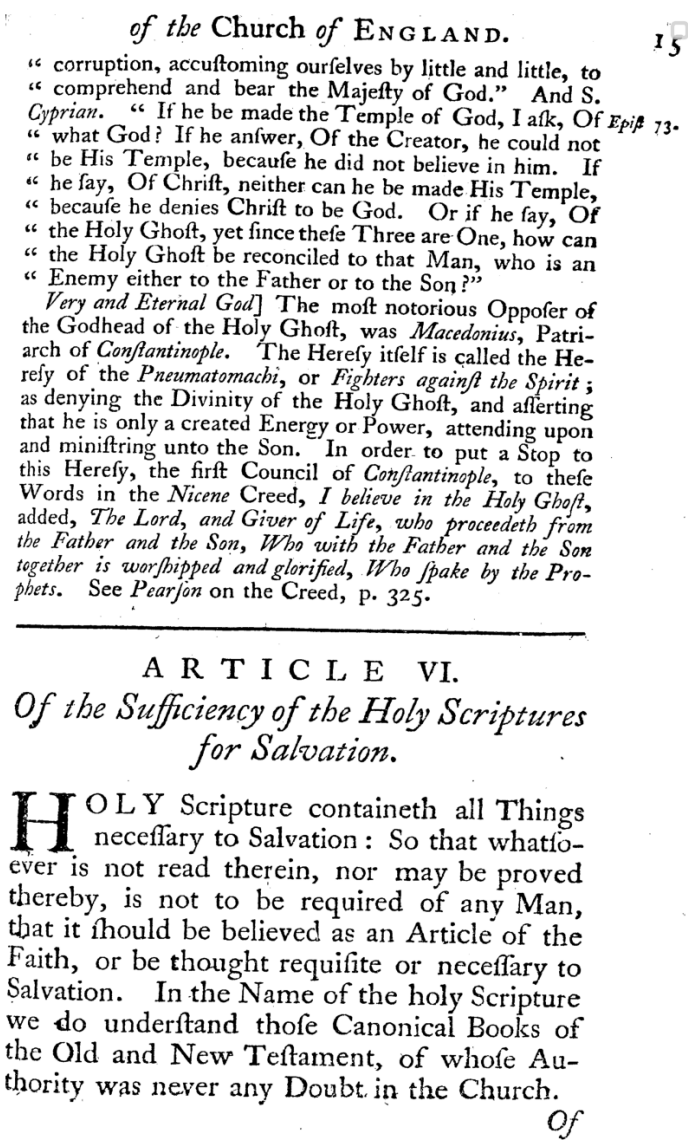
dent from Hiſtory, that this mock Senate,
this Senate in Burleſque, was compos'd of
a Parcel of Scoundrels who had never ſeen
Pharſalia. For can you or any one be-
lieve, that if they had been of real Senato-
rian Rank, *Cæſar* would have us'd them as
he did, who hang'd up as many of them as
fell into his Hands? But let us now ſee what
Sempronius is pleas'd to reply to *Portius.*

Not all the Pomp and Majeſty of Rome
Can raiſe her Senate more than Cato's *Pre-*
ſence.
———*O, my* Portius!
Could but I call that wond'rous Man my Fa-
ther,
Would but thy Siſter Marcia *be propitious*
To thy Friend's Vows, I might be bleſt in
deed.
Port. *Alas,* Sempronius! *would'ſt thou*
talk of Love
To Marcia, *while her Father's Life's in*
danger?
Thou might'ſt as well court the pale trembling
Veſtal,
When ſhe beholds the Holy Flame expiring
Sempr. *The more I ſee the Wonders of thy*
Race,
The more I'm charm'd. Thou muſt take heed
my Portius,
The World has all its Eyes on Cato's *Son.*
Thy Father's Merit ſets thee up to View,
An

GT:

Pred:
sempronius portius marcia pharsalia cato vesta

Figure 7: An example page without Latin fragments, together with our ground truth and prediction for that page.



of the Church *of* ENGLAND. 15

" corruption, accuſtoming ourſelves by little and little, to " comprehend and bear the Majeſty of God." And S. *Cyprian*. " If he be made the Temple of God, I aſk, Of *Epiſt* 73. " what God? If he anſwer, Of the Creator, he could not " be His Temple, becauſe he did not believe in him. If " he ſay, Of Chriſt, neither can he be made His Temple, " becauſe he denies Chriſt to be God. Or if he ſay, Of " the Holy Ghoſt, yet ſince theſe Three are One, how can " the Holy Ghoſt be reconciled to that Man, who is an " Enemy either to the Father or to the Son?"

Very and Eternal God] The moſt notorious Oppoſer of the Godhead of the Holy Ghoſt, was *Macedonius*, Patriarch of *Conſtantinople*. The Hereſy itſelf is called the Hereſy of the *Pneumatomachi*, or *Fighters againſt the Spirit*; as denying the Divinity of the Holy Ghoſt, and aſſerting that he is only a created Energy or Power, attending upon and miniſtring unto the Son. In order to put a Stop to this Hereſy, the firſt Council of *Conſtantinople*, to theſe Words in the *Nicene* Creed, *I believe in the Holy Ghoſt*, added, *The Lord, and Giver of Life, who proceedeth from the Father and the Son, Who with the Father and the Son together is worſhipped and glorified, Who ſpake by the Prophets*. See *Pearſon* on the Creed, p. 325.

ARTICLE VI.

Of the Sufficiency of the Holy Scriptures for Salvation.

HOLY Scripture containeth all Things neceſſary to Salvation: So that whatſoever is not read therein, nor may be proved thereby, is not to be required of any Man, that it ſhould be believed as an Article of the Faith, or be thought requiſite or neceſſary to Salvation. In the Name of the holy Scripture we do underſtand thoſe Canonical Books of the Old and New Teſtament, of whoſe Authority was never any Doubt in the Church.

Of

GT:

Pred:

credo in spiritum sanctum dominum et vivificantem qui ex patre filioque procedit qui cum patre et filio adoratur et conglorificatur qui locutus est per prophetas

Figure 8: An example page without Latin fragments, together with our ground truth and hallucinated prediction for that page.