Across Generations: A Comparative Analysis of NER for Latin Inscriptions from Classical Machine Learning to LLMs

Wenhui Cui, Phillip Benjamin Ströbel
University of Zurich
wenhui.cui@uzh.ch, phillip.stroebel@uzh.ch

Abstract

Latin epigraphic texts are a challenging type of historical data for natural language processing (NLP). They are often fragmentary, contain inconsistent spelling, and follow complex Roman naming conventions. This paper investigates Named Entity Recognition (NER) for this domain by comparing several approaches, including feature-based Support Vector Machines, neural models such as BiLSTM and TreeLSTM, pre-trained language models like LatinBERT, fine-tuned Transformer models based on BERT, and large language models used with prompting and supervised fine-tuning. We introduce a manually annotated dataset of 1,000 inscriptions from the *Epigraphik-Datenbank Clauss-Slaby*, labelled with a fine-grained BIO scheme that captures the internal structure of Roman personal names. Results show that the fine-tuned BERT model achieves the highest performance, with a weighted F1 score of 91.1% and a macro F1 of 68.7%, and clearly outperforms other methods. Additional linguistic features, such as part-of-speech tags and dependency information, yield only limited improvements, likely due to the irregular nature of inscriptional texts. This work provides a new benchmark for NER on Latin inscriptions and offers practical insights into applying modern NLP techniques to historical, non-standardised language.

Keywords: Latin NLP, Named Entity Recognition, Epigraphy, Low-resource Languages, Machine Learning, LLMs

1. Introduction

Named entity recognition (NER) is a key natural language processing (NLP) task that identifies references to persons, locations, and other entities (Munnangi, 2024). Due to the high density of named entities in historical inscriptions, these texts constitute an important linguistic resource for the study of people, society, and geography in the ancient world (Bodel, 2012).

In recent years, there has been growing interest within the Digital Humanities and NLP communities in applying advanced language models to historical and low-resource languages (Manjavacas and Fonteyn, 2022). While NER has achieved strong performance on modern high-resource languages using neural and Transformer-based models, increasing attention is now being directed toward extending these approaches to historical languages (Brandsen et al., 2021; Schweter and März, 2020; Konle and Jannidis, 2020). As epigraphic collections become increasingly digitised, there is growing interest in applying NLP to these materials. However, unlike well-edited literary texts, inscriptions are often fragmentary, contain inconsistent spelling, and follow conventionalised formulaic patterns, which makes automatic processing difficult (Heřmánková et al., 2021).

Latin inscriptions are particularly challenging because Roman personal names consist of multiple components, such as *praenomen* (EN *first name*), *nomen gentilicium* (EN *family/clan name*), and *cognomen* (EN *nickname*),[1] which constitute the so-called *tria nomina* and may appear in abbreviated or incomplete forms (Salway, 1994). E.g., in the name *M. Tullius Cicero*, we find the abbreviation for the *praenomen* ***Marcus***, followed by the *nomen gentilicium* ***Tullius*** and the *cognomen* ***Cicero***. Inscriptions also include place names and institutional titles embedded in fixed expressions. These domain-specific features require a more detailed annotation scheme than standard person and location labels.

In this paper, we address three research questions: (RQ1) How do traditional feature-based models, neural sequence labellers, Transformer encoders, and LLMs compare when applied to NER on Latin inscriptions? (RQ2) To what extent do additional linguistic features, such as POS tags and dependency parses, improve performance on highly irregular inscriptional texts? (RQ3) How competitive are large language models, used via prompting or supervised fine-tuning, compared to task-specific fine-tuned Transformers in this low-resource, historical setting?

The contributions of this paper are the following:

- We develop and evaluate a NER system for Latin inscriptions using a manually annotated dataset of 1,000 inscriptions from the

[1] The translations of the different parts of the name are approximate in the sense that these names were not necessarily used how we would use them nowadays. Moreover, the usage of the different names has changed over the centuries (see, e.g., Ayer (2014)).

Proceedings of LT4HALA 2026 @ LREC 2026, pages 112–124.

Epigraphik-Datenbank Clauss-Slaby (EDCS, Clauss et al. (1985)) as a gold standard.[2]

- We apply existing Latin NER models, especially LatinCy[3] (Burns, 2023), and test large language models (LLMs) in a zero-shot setting to establish baselines to compete against.
- We trained the following models: (1) A linear SVM baseline using handcrafted features, including prefixes, suffixes, POS tags, and dependency relations, and (2) neural models, including a TreeLSTM for hierarchical structure and a BiLSTM CNN model for morphological and contextual features. We fine-tuned (3) Transformer models, i.e, multilingual BERT and a Latin-specific RoBERTa, both of which use token classification for BIO tagging. We (4) applied LLMs using few-shot prompting and supervised fine-tuning, combined with CRF and rule-based post-processing.
- We test all models on a fixed stratified dataset using token-level precision, recall, and micro/macro F1, enabling a systematic comparison across traditional, neural, Transformer, and LLM approaches.[4]

2. Related Work

The development of NER for Latin has been relatively slow, particularly outside literary texts (Erdmann et al., 2016). Early research was mostly part of broader projects like the Classical Language Toolkit (CLTK) (Johnson et al., 2021) and LatinCy (Burns, 2023). While these initiatives focused on establishing core linguistic processing capabilities, most notably part-of-speech (POS) tagging, they created the essential computational infrastructure that later enabled more advanced tasks such as NER. As noted by Erdmann et al. (2016), who achieve a 90% F-score on literary data, performance differences are expected when applying NER tools to texts from different Latin genres and linguistic styles.

Building upon this infrastructure, early dedicated NER efforts for Latin typically used Conditional Random Field (CRF) models with handcrafted linguistic features such as word shapes and POS tags, as exemplified by the recognition of named entities in Medieval Latin charters (Chastang et al., 2021). While these methods achieved high performance on literary texts, with F1 scores up to 0.95 for exact matches, they have not been applied to inscriptional material. Given the fragmented and formulaic nature of inscriptions, the frequent use of abbreviations, and variations in spelling, feature-based models are expected to face substantial challenges when applied to epigraphic data.

Later, the field shifted towards machine learning models tailored for sequence labelling tasks. Neural architectures such as the BiLSTM-CRF model enable models to learn complex patterns, including contextual cues and morphological variation, directly from data, reducing reliance on handcrafted linguistic features (Lample et al., 2016). This class of models has become a standard approach in NER, achieving strong performance across multiple languages and related domains. Research has shown that explicitly incorporating linguistic structures, such as syntactic constituents, into neural models can improve NER by enabling the model to leverage hierarchical phrase structure information and semantic features from syntactic trees (Li et al., 2017). Building on these ideas, researchers have also adapted sequence labelling approaches to Latin, as in the *Herodotos Project*, which developed NER systems for Classical Latin and ancient Greek texts that accurately identify groups, persons, and places, illustrating the feasibility of machine learning–based entity recognition in historical languages (Erdmann et al., 2023). Their model achieved micro-F1 scores of 0.99 on in-domain data and only slightly lower scores on out-of-domain data (i.e., data the model was not trained on).

Following the introduction of Transformer architectures (Devlin et al., 2019), Latin-specific models such as LatinBERT (Bamman and Burns, 2020) demonstrated that contextualised representations can effectively capture linguistic patterns in Latin. Subsequent work explored other Transformer variants, including ELECTRA-based models for lemmatisation and POS tagging on Classical Latin corpora (Mercelis and Keersmaekers, 2022). Building on these foundations, researchers have adapted Transformer approaches to historical language scenarios, including NER (Beersmans et al., 2023). Also, multilingual models such as mBERT and XLM-RoBERTa have been fine-tuned for NER on medieval charter corpora in Latin, French, and Spanish, achieving strong cross-lingual performance without significant degradation compared to monolingual baselines (Torres Aguilar, 2022). Another approach uses cross-lingual annotation projection to transfer NER labels from modern languages to low-resource classical languages, such as Latin and ancient Greek, leveraging parallel corpora and neural word alignment to expand training data (Yousef et al., 2023).

[2] See https://edcs.hist.uzh.ch.

[3] Available within *spaCy*, see https://spacy.io and for LatinCy especially https://spacy.io/universe/project/latincy.

[4] The code and data of this work are available at https://github.com/Wenhui620/latin-inscriptions-ner.

In recent years, LLMs have made substantial progress in NLP due to their scale and contextual learning ability, and have increasingly been applied to NER (Brown et al., 2020). Because NER is a sequence labelling task while LLMs are generative models, two main adaptation strategies have been proposed: few-shot prompting and fine-tuning (Ji et al., 2025). Few-shot prompting incorporates task descriptions and a small number of annotated examples into the input, allowing models to infer labelling patterns through in-context learning without parameter updates (Federiakin et al., 2024). Approaches such as GPT-NER reformulate sequence labelling as text generation and show strong performance in low-resource settings (Wang et al., 2023). In contrast, fine-tuning updates model parameters on task-specific data to improve domain accuracy and consistency, with parameter-efficient methods reducing computational cost (Zhang et al., 2025).

For Latin and other historical languages, NER remains challenging due to limited annotated data and evolving naming conventions (Ehrmann et al., 2021). Recent work shows that LLMs can outperform traditional NLP frameworks such as spaCy and flair on historical NER, achieving gains of 7 to 22 % F1 through context-aware prompting, although increasing the number of examples beyond a certain point brings little improvement (Hiltmann et al., 2025). However, zero-shot experiments on historical sources reveal persistent difficulties, including inconsistent annotation, entity complexity, multilingual code-switching, and sensitivity to prompt design (González-Gallardo et al., 2023). Because naming rules change over time and annotated data remain scarce, historical NER remains difficult (Ehrmann et al., 2021). Consequently, both few-shot prompting and fine-tuning should be viewed as complementary to traditional supervised methods.

This overview of related work shows several desiderata for NER in Latin. First, none of the approaches mentioned addressed NER in Latin inscriptions. Secondly, there is only a distinction between names and other categories, where names *per se* are not annotated with more fine-grained categories. Thirdly, the problem of Latin being a low-resource language is aggravated in inscriptions, which constitute a specific, highly formulaic subgenre. Fourthly, there is no comprehensive comparison of machine learning methods applied to inscriptions, especially one that includes traditional methods, which, due to the hype generated by LLMs, are often overlooked. Our paper addresses all of these issues.

3. Method

Figure 2 shows the experimental framework used in this study. The workflow illustrates the systematic comparison of five distinct modelling paradigms applied to Latin inscription NER.

3.1. Dataset

The dataset used in this research consists of 1,000 Latin inscriptions manually collected from the EDCS. The selection process focused on gathering a representative sample of Roman names spanning from the early Republican to Imperial periods and distributed across multiple provinces throughout the Roman world. Hence, we compiled the dataset as a stratified sample, accounting for metadata such as date, classification, province, and material. This reduced the number of inscriptions from which we sampled from over 500,000 to approximately 200,000.[5] Each inscription was cleaned to remove modern symbols[6] used to mark missing, erased or supplemented text while keeping the original structure of the text. We applied a fine-grained BIO tagging schema to these texts to identify 10 specific entity types.

The annotation is inspired by existing guidelines for named entity annotation in Classics (e.g., Romanello and Najem-Meyer (2022)), particularly in its treatment of entity categories and boundary definitions, while introducing additional fine-grained labels tailored to the structure of Latin personal names in inscriptions. It includes several tags for fine-grained personal name elements: *PERS:PRAE* (e.g., *Publius*) for the first name, *PERS:NOMEN* (e.g., *Iulius*) for the clan name, *PERS:COG* (e.g., *Cicero*) for the family name, and *PERS:AG* (e.g., *Africanus*) for an additional nickname. It also identifies *PERS:FILI* to mark family relations like a father's name (e.g., *Marci filii*, EN *son of Marcus*) and *PERS:TRIBE* (e.g., *Palatina*) for the voting tribe. Beyond personal names, the dataset marks *PERS:TITLE* (e.g., *Augustus*) when a title is part of a name, and *TITLE* (e.g., *pontifex*) for independent positions such as priest or veteran. For coarse-grained annotation, the general *PERS* tag is used, which spans all name parts of a single person. Geographic information is captured by the

[5] We are aware that we limited the selection to certain factors, which might not result in a fully representative sample of the overall data, but we favoured this approach since it is 1) reproducible and 2) still represents a subset of the corpus. The selected inscriptions span the early Republican to the late Imperial period (roughly 3rd century BCE to 4th century CE) and cover provinces across the Italian peninsula, the western provinces (Gallia, Hispania, Africa), and the eastern Mediterranean, reflecting the geographic spread of the EDCS corpus.

[6] Added during the editing process.

LOC tag for cities or provinces. To ensure the quality of the manual annotation, the data was checked for consistency across different inscription types.

We divided the dataset into three parts for the experiments: 700 inscriptions for training, 150 for validation, and 150 for testing. Again, stratified sampling was used to ensure that each set had a similar mix of entity types. This data organisation enables a reliable evaluation of how different machine learning and transformer-based models handle the specific patterns found in Latin inscriptions.

3.2. Inter-Annotator Agreement

Category	A1 vs A2	A1 vs A3	A2 vs A3
AG	0.691	0.000	0.087
COG	0.789	0.685	0.724
FILI	0.556	0.528	0.645
LOC	0.333	0.000	0.000
NOMEN	0.742	0.835	0.755
PERS	0.303	0.214	0.391
PRAE	0.738	0.802	0.855
TITLE	0.510	0.382	0.419
TRIB	0.000	0.000	0.875
Micro-average	0.618	0.540	0.609
Macro-average	0.518	0.383	0.528

Table 1: Pairwise span-level F1-scores per entity category and global averages. In each pair (X vs Y), annotator X serves as the reference (proxy gold standard) against which Y is evaluated.

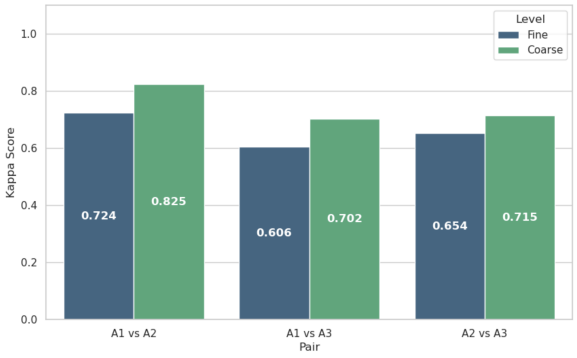

Figure 1: Pairwise comparison between annotators based on Cohen's κ.

Three annotators independently labelled 129 inscriptions from the test set to evaluate inter-annotator agreement. The annotation team comprised a Latin epigraphy expert, a historian in training, and a student responsible for data preparation and annotation coordination. Table 1 shows pairwise F1 scores, while Figure 1 shows Cohen's κ. The overall agreement, according to Fleiss' κ, is 0.658, indicating moderate agreement. However, the results generally show high overlap in annotations of the *tria nomina*, whereas the other parts of the names are more difficult to identify. This is largely due to the limited contextual information available in inscriptions. In many cases, especially in fragmentary or formulaic texts, additional name elements, such as filiations, tribal affiliations, and titles, lack sufficient context to be interpreted reliably. As a result, annotators may differ in both boundary detection and category assignment, leading to lower agreement.

3.3. Models

This study compares several modelling paradigms to evaluate their effectiveness in recognising entities in Latin inscriptions, ranging from traditional machine learning to modern neural and generative approaches.

Existing Latin NER Models We first evaluate an existing Latin NLP system as an external reference point. Specifically, we test the LatinCy pipeline in a zero-shot setting to establish a modern baseline. This model was trained on a mixture of Latin treebanks, web corpora, and historical text collections, but not on epigraphic material. To ensure alignment with our gold dataset, we construct spaCy `Doc` objects directly from the gold tokenisation, preventing the model from re-tokenising the text. The span-level entity predictions produced by LatinCy are converted into token-level BIO tags. Since LatinCy uses a coarse label set, its entity types are mapped to three evaluation categories: PERS, TITLE, and LOC, whereas all other predictions are mapped to O. On the gold-standard side, our fine-grained person name subtypes are collapsed into their coarse classes while preserving the BIO prefixes. Performance is measured using token-level Precision, Recall, and F1 score.

Support Vector Machine As a standard baseline, we use a linear SVM classifier implemented in `scikit-learn`.[7] The task is formulated as token-level BIO classification. We compare two feature settings. The baseline feature set includes the lowercase token form, word shape, prefixes, suffixes, and the previous and next tokens. The linguistic feature set is extended by adding POS tags, dependency relations, and the lowercase form of the syntactic head, which are obtained from the LatinCy parser. All features are converted into sparse vectors using `DictVectorizer`. To address class imbalance, class weights are set to `balanced`. The classifier is further wrapped with

[7] See https://scikit-learn.org/stable.

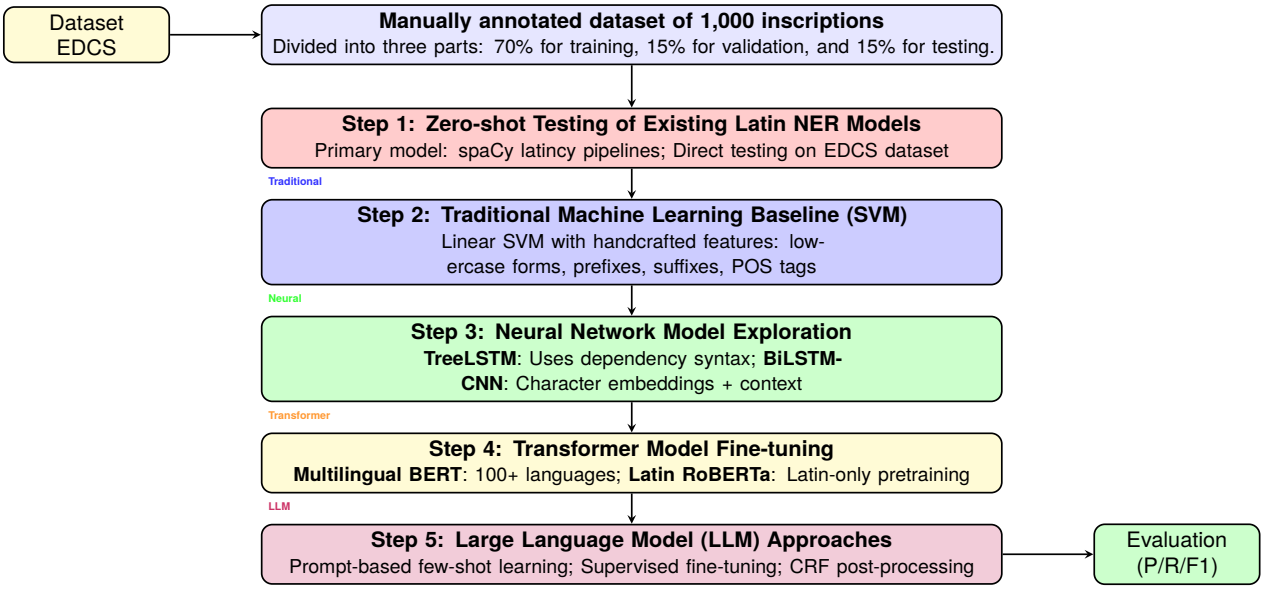

Figure 2: Overview of the experimental framework.

`CalibratedClassifierCV` using sigmoid calibration on the validation set to produce more reliable probability estimates.

Neural Models We then explore two neural network architectures implemented in PyTorch. The first is a TreeLSTM that incorporates dependency structure by aggregating information from child to parent nodes; its input representation concatenates word embeddings, POS and dependency embeddings, and character-level CNN embeddings, which are jointly learned during training, followed by a hidden size of 256 and a linear layer for BIO tag prediction. The second is a BiLSTM-CNN hybrid that processes tokens sequentially; each token representation combines word, character CNN, prefix, suffix, and surrounding token embeddings, which are fed into a bidirectional LSTM with hidden size 256 and a linear classification layer. During training, dropout with a rate of 0.5 is applied to reduce overfitting, and both models are trained using the Adam optimiser with a learning rate of 0.001 for 20 epochs and a batch size of 32; mini-batch gradient descent is employed, and padding tokens are ignored in the loss computation.

BERT-based Models For Transformer-based methods, we fine-tune two pretrained encoders for token classification. The first is `bert-base-multilingual-cased`,[8] which was pretrained on more than 100 languages, including Latin. The second is a Latin-specific RoBERTa-based model pretrained only on Latin corpora.[9] Both encoders are further fine-tuned on our inscription dataset as token-classification models. For multilingual BERT, we use the Hugging Face Transformers library with a standard token classification head and fine-tune for three epochs with a batch size of 16 and a learning rate of 5e-5. For the Latin RoBERTa model, we use the spaCy training framework with a Transformer-plus-NER pipeline, converting the data to spaCy's binary format prior to training.

LLMs Finally, we evaluate LLMs under both prompt-based and fine-tuned settings. In the prompt-based experiments, GPT-4.1-mini[10], Gemini-2.5-flash[11], and Claude-Sonnet-4[12] are tested with zero-shot and few-shot prompting (using 3-, 5-, and 10-shot prompts). The prompts define the BIO tagging task and require the models to output JSON objects containing aligned token and tag sequences.[13] For long inscriptions, the input is split into shorter segments, and the predictions are concatenated. After inference, a CRF model trained on the gold data is applied as a post-processing step, and rule-based constraints enforce common patterns in Roman names.

[8] See https://huggingface.co/google-bert/bert-base-multilingual-cased.

[9] See https://huggingface.co/pstroe/roberta-base-latin-cased.

[10] See https://developers.openai.com/api/docs/models/gpt-4.1-mini.

[11] See https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash.

[12] https://www.anthropic.com/news/claude-sonnet-4-5.

[13] The prompts are available in the GitHub repository accompanying this paper (see Footnote 4).

In the fine-tuning setting, Gemini and GPT models are further trained on the annotated corpus using supervised learning. The training data are formatted as JSONL files containing token and BIO tag sequences. Gemini fine-tuning is conducted via the Vertex AI interface using the default hyperparameters, such as learning rate, batch size, and optimiser type, which are platform-defined and do not expose fixed numeric values. GPT fine-tuning is performed for 3 epochs with a batch size of 1 and a learning rate multiplier of 2. All models are evaluated on the same fixed test set.

3.4. Evaluation

To evaluate the models' performance on Latin inscriptions, we use standard metrics, including Precision, Recall, and weighted & macro F1 scores. Because the dataset employs a fine-grained BIO tagging schema and exhibits an uneven distribution of entity types, we primarily focus on the weighted F1 score. This ensures that the performance on rare tags is reflected fairly in the final results.[14] For LLM outputs, we apply a format check to convert the generated text into standard BIO format so that it can be compared directly with the other models under the same criteria.

4. Results and Discussion

This section describes the results of our experiments and the subsequent error analysis. Table 2 summarises the performance, specifically weighted F1, macro F1, and accuracy, for all models evaluated in this study.

4.1. Overall Performance Comparison

Clear performance differences emerge across model families. Overall, the fine-tuned multilingual BERT reaches 91.1% weighted F1 and 68.7% macro F1, indicating strong performance even when averaging over all entity types rather than weighting by frequency. Fine-tuned transformer models perform best overall. The fine-tuned multilingual BERT exhibits strong performance on frequent name categories such as *praenomen*, *nomen*, and *filiation* markers, indicating effective adaptation to abbreviation patterns and orthographic variation.

By contrast, zero-shot models such as LatinCy perform substantially worse, often failing to detect complete entity spans. The Latin RoBERTa model, although pretrained exclusively on Latin corpora, achieves a weighted F1 of 88.9% and a macro F1 of 59.9% after fine-tuning on our dataset, indicating that Latin-specific pretraining helps but still falls short of the multilingual BERT baseline.

LLMs show higher variability. Performance improves with more examples, and 10-shot prompting enables Claude Sonnet to reach 90.75% accuracy, close to that of fine-tuned BERT. Fine-tuned LLMs such as Gemini and GPT achieve stable accuracy above 90%, but their macro F1 remains limited due to errors on rare labels and boundary detection.

Traditional SVM models remain strong baselines, with weighted F1 scores of 88.40% and 87.96%, indicating that surface features capture much of the signal. The BiLSTM also performs competitively, with 88.60%, whereas the TreeLSTM performs substantially worse, at 73.45%, likely due to unreliable dependency parses.[15]

Overall, fine-tuned transformers perform best, followed by fine-tuned LLMs, pretrained models, SVM and BiLSTM, and finally TreeLSTM. Models that depend on accurate syntax or zero-shot reasoning are less reliable than those trained on annotated inscription data. Our results show that relying on the latest developments, i.e., LLMs, does not yield the best performance, as LLMs are outperformed by a pre-trained BERT. These representations appear to contain information valuable for the NER task. This is confirmed by the almost equally strong results of LatinBERT.

Recall values follow a broadly similar pattern to weighted F1, though with notable divergences for specific entity classes. Fine-tuned BERT achieves the highest overall weighted recall, closely matched by the fine-tuned LLMs. Among few-shot LLMs, Claude with 10-shot prompting is competitive at the weighted level, but per-class recall reveals a striking weakness on the coarse PERS category, where it almost entirely fails to recover underspecified name spans, preferring to assign a fine-grained subtype or nothing at all. LOC recall is low across all models and architectures, underscoring that location recognition remains the most challenging category. These recall figures are particularly relevant for historical NER, where missed entities may represent irretrievable losses of prosopographic or geographic information. Per-class recall values for all models are available in the project repository (see Footnote 4).

[14] It should be noted that the weighted F1 score is computed over all token labels, including O, which represents the majority class. This inflates the absolute weighted F1 values relative to an entity-only metric. To provide a more entity-focused picture, we also report macro F1, which averages scores across entity categories without weighting by frequency and excludes the dominance of O-class tokens. Readers primarily interested in entity detection performance should therefore weight the macro F1 figures more heavily when comparing models.

[15] For future work, we plan on an evaluation of the parsing results by LatinCy on the inscriptions.

Model Category	Model Name	Weighted F1	Macro F1	Accuracy
Transformer (fine-tuned)	**BERT***	**0.911**	0.687	**0.912**
LLM (few-shot)	Claude-10shot*	0.904	**0.759**	0.908
LLM (fine-tuned)	GPT-fine-tuned	0.896	0.623	0.906
LLM (fine-tuned)	Gemini-fine-tuned*	0.901	0.610	0.905
Transformer (pretrained)	Latin-RoBERTa	0.889	0.599	0.891
Neural Network (sequential)	BiLSTM*	0.886	0.613	0.892
LLM (few-shot)	Gemini-10shot	0.883	0.622	0.890
Traditional (feature-based)	SVM-Linguistic*	0.884	0.625	0.890
Traditional (feature-based)	SVM-nonLinguistic	0.880	0.628	0.888
LLM (few-shot)	Claude-5shot	0.881	0.639	0.880
Neural Network (hierarchical)	TreeLSTM	0.735	0.459	0.746
LLM (baseline)	Claude-0shot*	0.799	0.507	0.814
LLM (baseline)	Gemini-0shot	0.794	0.457	0.803
LLM (baseline)	GPT-0shot	0.712	0.310	0.700
External Pipeline	LatinCy	0.580	0.240	0.580

Table 2: Overall performance of all models. The best performances per measure and the best overall model are in **bold**. The best models per category (feature-based, neural network, Transformer-based, LLM) are marked with an asterisk (*).

4.2. Fine-Grained Entity Performance

Label	BERT	SVM	BiLSTM	Cl-10s	Gem-FT
B-P:PRAE	*0.928*	**0.883**	**0.885**	**0.922**	**0.896**
B-P:NOMEN	0.881	0.789	0.814	*0.895*	0.828
B-P:COG	0.795	0.734	0.771	0.762	***0.813***
B-P:FILI	***0.965***	0.816	0.835	0.882	0.816
B-TITLE	0.711	*0.750*	0.624	0.723	0.714
B-LOC	*0.579*	0.333	0.240	0.480	0.419

Table 3: Fine-grained F1 scores across representative models. Best-performing models are in **bold**. Best performances per category are in *italics*.

As shown in Table 3, model performance differs considerably across entity types. For *praenomen* and *nomen*, supervised models achieve high scores. BERT reaches 92.82% F1 on *praenomen*, and Claude, with 10-shot prompting, performs similarly. For *nomen*, Claude 10-shot gives the best result, followed closely by BERT and the BiLSTM. SVM models perform slightly worse, especially on *nomen*, which shows greater lexical diversity.

Cognomen are more challenging because of their wider range of forms. Fine-tuned Gemini performs best in this category, followed by BERT. Feature-based and few-shot LLM models perform worse, indicating that this category benefits from stronger contextual modelling.

Filiation markers are easier for most systems. Their repetitive structure and fixed abbreviation patterns help all models. BERT achieves the highest score at 96.52%, and several other models also perform strongly.

Titles and location names are the most difficult categories. Titles appear in many syntactic positions and with varied forms. In this category, the SVM slightly outperforms other systems. Location names are sparse and highly variable in spelling. BERT performs best, but overall scores remain much lower than for personal name categories.

These results show that high-frequency and formulaic entity types, i.e., the *tria nomina*, are easier for all models, while rare and lexically diverse categories reveal clearer differences in model capacity and domain adaptation.

4.3. Error Analysis

This section analyses systematic discrepancies between model predictions and gold annotations. Using Claude-Sonnet-4 with 10-shot prompting as an example, Figure 3 provides a confusion matrix summarising category confusions.

Errors in Personal Names Labels such as *PERS:PRAE*, *PERS:NOMEN*, *PERS:COG*, *PERS:FILI*, and *PERS:AG* show high accuracy along the diagonal, indicating that the model broadly captures the semantic domain of Roman personal names. Nevertheless, fine-grained errors are substantial and reveal systematic difficulties in handling the internal structure of the *tria nomina* system.

The most prominent pattern is that *PERS:COG* functions as a default category, absorbing tokens

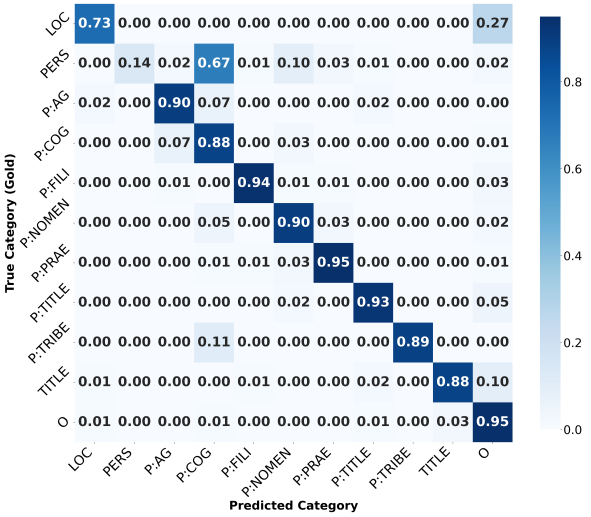

Figure 3: Semantic confusion matrix illustrating the misclassification patterns between entity labels.

from multiple related subtypes. Error rates for *PERS:COG* and the general *PERS* category are particularly high, reaching up to 67%, indicating frequent misassignment between the default *cognomen* label and the coarse-grained person category. Detailed analysis of the confusion matrix reveals specific bidirectional confusions. First, *PERS:COG* and *PERS:AG* exhibit mutual misclassifications (7% in both directions). This confusion is likely due to two factors: *agnomina* appear infrequently in the training data, and both categories occupy a similar structural position within the name. They follow the *nomen* and *cognomen* and serve as additional descriptive or honorific components. Second, *PERS:COG* and *PERS:TRIBE* show substantial bidirectional confusion (11%). This pattern reflects historical developments in Roman naming practices: over time, many tribal designations gradually became absorbed into the nomenclature, sometimes evolving into *nomina* or *cognomina*. As a result, some lexical forms can function as either tribal markers or name components, creating inherent ambiguity that challenges automatic classification.

Two further issues emerge with specific types of name components. First, filiation markers such as *filia*, *filius*, and *filiae* (forms of EN *daughter* or *son*), particularly in formulaic expressions (e.g., *Marci* ***filii***, EN son of Marcus), are annotated as *PERS:FILI* in the gold data because they encode family relationships. The model, however, often labels these tokens as *O*, treating them as ordinary nouns rather than structural components of the name. Second, with fragmented or abbreviated names, although the annotation scheme permits a general *PERS* label when the internal structure is unclear, the model often predicts more specific subcategories such as *PERS:NOMEN* or *PERS:COG* even for short or underspecified tokens. This behaviour reflects an implicit assumption of a complete *tria nomina* structure and overgeneralisation from canonical patterns rather than adaptation to underspecified contexts.

Overall, these patterns demonstrate that while the model captures the semantic notion of personhood, it struggles with the internal hierarchical and relational structure of Roman personal names. The concentration of errors in *PERS:COG* and the high confusion across both general and specific categories highlight the limitations of token-level classification for representing complex onomastic conventions, in which different components encode distinct types of social and familial information.

Errors in Titles and Locations The *TITLE* label is reserved for historically attested Roman administrative or military positions. Terms such as *consularis*, meaning consul, are annotated as *TITLE* when they appear in formal or official contexts, such as in inscriptions recording careers, offices, or honours. The model does not always follow this principle: while most *TITLE* tokens are correctly predicted, with 88% of instances on the diagonal, around 10% are misclassified as *O*. Such errors typically occur when titles resemble common nouns or appear in less explicit contexts, suggesting that the model sometimes treats these items as ordinary nouns rather than recognising their function as name components. Multiword title expressions that function as a single unit are occasionally fragmented, reflecting difficulties in modelling longer, formulaic sequences rather than misunderstandings of individual words.

Location recognition and segmentation exhibit similar challenges. Administrative units and province names are labelled as *LOC* when referring to concrete geopolitical entities. While a majority of *LOC* tokens are correctly identified, accounting for 73% of instances, a substantial proportion, approximately 27%, is misclassified as *O*. This pattern reflects a conservative prediction behaviour when the geographical meaning is ambiguous. Boundary inconsistencies arise when parts of multiword locations are labelled differently, resulting in split or shifted spans. Ethnic adjectives, such as *Gallorum*, meaning *Gallic*, further complicate recognition, as these forms are sometimes interpreted as locations due to their morphological similarity to place names, even when they function as descriptors of people. Overall, these patterns indicate that the model relies heavily on surface-form cues and struggles to distinguish between geographic reference and ethnic attribution in contexts without explicit signals.

4.4. Overall Discussion

Across models and entity types, consistent patterns emerge. Fine-tuned transformers achieve the best performance, particularly on frequent and regular personal name categories such as *PRAE* and *NOMEN*, underscoring the advantage of supervised learning for formulaic inscriptional data. Traditional feature-based models and pretrained transformers handle frequent patterns reasonably well, but struggle with rare or diverse entities. LLMs benefit from few-shot prompting but remain constrained by boundary detection and the scarcity of rare-label data.

Error analysis shows that most errors are systematic. Personal names are generally detected, but fine-grained subtypes are sometimes confused. Titles and locations are often missed or fragmented, especially in long formulaic expressions. Multiword expressions, abbreviations, and ambiguous morphological forms further increase difficulty. For example, ethnic adjectives are sometimes misclassified as locations due to surface similarity.

Overall, model performance depends on the extent to which inscriptional patterns match the learned categories. Frequent and regular entities are reliably recognised, while rare and complex cases remain challenging. These findings suggest that combining semantic modelling with targeted linguistic features and expanding annotated data can further improve Latin NER.

In summary, our experiments show that, in this noisy, low-resource, and highly formulaic historical setting, task-specific fine-tuned Transformer encoders still set the performance ceiling: none of the prompted or fine-tuned LLMs surpasses the fine-tuned BERT baseline, although they come close and exhibit complementary error profiles. This suggests that combining encoder-based models with LLMs, for example, via simple ensembling or post-editing, is a promising direction for future work rather than replacing supervised models outright.

5. Conclusion and Future Work

This study evaluated NER approaches for Latin inscriptions, including feature-based models, neural networks, and large language models. Fine-tuned transformers performed best, with BERT achieving a weighted F1 score of 91.1% and effectively capturing formulaic naming patterns, while few-shot Claude generalised well to rare entities (weighted F1 of 90.4%). Traditional SVMs that incorporated handcrafted linguistic features such as morphology, part-of-speech tags, and dependency relations provided a reasonable baseline, but their contribution was limited, as evidenced by the modest performance of syntax-dependent models such as TreeLSTM. Error analysis revealed systematic challenges: fine-grained subtypes of personal names were sometimes misassigned, titles and locations were often underrecognised or fragmented, and multiword or abbreviated expressions remained difficult to recognise. These patterns highlight the value of domain-specific fine-tuning and suggest that combining LLM semantics with selective linguistic cues may be more effective than relying solely on structural features. Limited overlap in errors between fine-tuned models and few-shot LLMs also indicates potential for ensemble methods. Future work could expand the annotated corpus to cover more regions, periods, and inscription types, thereby improving model generalisation and robustness. Moreover, a logical next step is to link names and locations to external knowledge sources. As such, this work opens the way to re-drawing the social network of the ancient Roman Empire as documented in inscriptions.

6. Limitations

Our study has several limitations. First, the inscription sample, although stratified by date, province, and material, is not fully representative of the entire EDCS corpus and underrepresents some regions and genres. Second, inter-annotator agreement is low for certain labels such as TRIB, TITLE, and LOC, which constrains the achievable upper bound and complicates the evaluation of these categories. Third, syntax-based models depend on LatinCy parses that are not optimised for fragmentary epigraphic Latin, which likely contributes to the poor performance of the TreeLSTM. Finally, we restrict ourselves to surface-level BIO tagging of flat entities; nested or discontinuous entities and document-level co-reference are left for future work.

7. Acknowledgements

We would like to thank Dr. Jens Bartels for his invaluable help in the annotation of the gold standard.

8. Bibliographical References

Meagan Ayer. 2014. *Allen and Greenough's New Latin Grammar for Schools and Colleges*. Dickinson College Commentaries, Carlisle, Pennsylvania.

David Bamman and Patrick J. Burns. 2020. Latin BERT: A Contextual Language Model for Classical Philology. ArXiv:2009.10053 [cs].

Marijke Beersmans, Evelien de Graaf, Tim Van de Cruys, and Margherita Fantoli. 2023. Training

and Evaluation of Named Entity Recognition Models for Classical Latin. In *Proceedings of the Ancient Language Processing Workshop*, pages 1–12, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

John Bodel. 2012. Epigraphic Evidence: Ancient History from Inscriptions.

Alex Brandsen, Suzan Verberne, Karsten Lambers, and Milco Wansleeben. 2021. Can BERT Dig It? – Named Entity Recognition for Information Retrieval in the Archaeology Domain.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners.

Patrick J. Burns. 2023. LatinCy: Synthetic Trained Pipelines for Latin NLP.

Pierre Chastang, Sergio Octavio Torres Aguilar, and Xavier Tannier. 2021. A Named Entity Recognition Model for Medieval Latin Charters. *Digital Humanities Quarterly*, 15(4).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. Named Entity Recognition and Classification in Historical Documents: A Survey.

Alexander Erdmann, Christopher Brown, Brian Joseph, Mark Janse, Petra Ajaka, Micha Elsner, and Marie-Catherine de Marneffe. 2016. Challenges and Solutions for Latin Named Entity Recognition. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 85–93, Osaka, Japan. The COLING 2016 Organizing Committee.

Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeux-Prunel, and Marie-Catherine de Marneffe. 2023. Herodotos Project Latin NER Tagger Annotation.

Denis Federiakin, Dimitri Molerov, Olga Zlatkin-Troitschanskaia, and Andreas Maur. 2024. Prompt Engineering as a New 21st Century Skill. *Frontiers in Education*.

Carlos-Emiliano González-Gallardo, Emanuela Boros, Nancy Girdhar, Ahmed Hamdi, Jose G. Moreno, and Antoine Doucet. 2023. Yes but.. Can ChatGPT Identify Entities in Historical Documents?

Petra Heřmánková, Vojtěch Kaše, and Adela Sobotkova. 2021. Inscriptions as Data: Digital Epigraphy in Macro-historical Perspective. *Journal of Digital History*, 1(1).

Torsten Hiltmann, Martin Dröge, Nicole Dresselhaus, Till Grallert, Melanie Althage, Paul Bayer, Sophie Eckenstaler, Koray Mendi, Jascha Marijn Schmitz, Philipp Schneider, Wiebke Sczeponik, and Anica Skibba. 2025. NER4all or Context is All You Need: Using LLMs for low-effort, high-performance NER on historical texts. A humanities informed approach.

Lixia Ji, Yiping Dang, Yunlong Du, Wenzhao Gao, and Han Zhang. 2025. Nested Named Entity Recognition: A Survey of Latest Research. *Expert Systems*, 42(7):e70052.

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.

Leonard Konle and Fotis Jannidis. 2020. Domain and Task Adaptive Pretraining for Language Models. In *Workshop on Computational Humanities Research*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition.

Peng-Hsuan Li, Ruo-Ping Dong, Yu-Siang Wang, Ju-Chieh Chou, and Wei-Yun Ma. 2017. Leveraging Linguistic Structures for Named Entity Recognition with Bidirectional Recursive Neural Networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2664–2669, Copenhagen, Denmark. Association for Computational Linguistics.

Enrique Manjavacas and Lauren Fonteyn. 2022. Adapting vs. Pre-training Language Models for

Historical Languages. *Journal of Data Mining & Digital Humanities*, NLP4DH(Digital humanities in languages). Publisher: Episciences.org.

Wouter Mercelis and Alek Keersmaekers. 2022. An ELECTRA Model for Latin Token Tagging Tasks. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 189–192, Marseille, France. European Language Resources Association.

Monica Munnangi. 2024. A Brief History of Named Entity Recognition.

Matteo Romanello and Sven Najem-Meyer. 2022. Guidelines for the annotation of named entities in the domain of classics. Developed in the context of the Ajax Multi-Commentary project and used for the CLEF HIPE 2022 shared task.

Benet Salway. 1994. What's in a Name? A Survey of Roman Onomastic Practice from c. 700 B.C. to A.D. 700. *Journal of Roman Studies*, 84:124–145.

Stefan Schweter and Luisa März. 2020. Triple E - Effective Ensembling of Embeddings and Language Models for NER of Historical German. In *Conference and Labs of the Evaluation Forum*.

Sergio Torres Aguilar. 2022. Multilingual Named Entity Recognition for Medieval Charters Using Stacked Embeddings and Bert-based Models. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 119–128, Marseille, France. European Language Resources Association.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. GPT-NER: Named Entity Recognition via Large Language Models.

Tariq Yousef, Chiara Palladino, Gerhard Heyer, and Stefan Jänicke. 2023. Named Entity Annotation Projection Applied to Classical Languages. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 175–182, Dubrovnik, Croatia. Association for Computational Linguistics.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2025. Instruction Tuning for Large Language Models: A Survey.

9. Language Resource References

Clauss et al. 1985. *Epigraphic Database Clauss Slaby (EDCS)*. Epigraphik-Datenbank Clauss/Slaby Project, University of Zurich; Catholic University of Eichstätt-Ingolstadt. Online database of Latin inscriptions.

A. Prompt for zero-shot experiments

B. Prompt for few-shot experiments

```
You are a Latin epigraphy Named Entity Recognition (NER) expert.
Your task is to assign BIO tags to each token of Latin funerary
inscriptions.

Use only the following entity tags:
- B-PERS:PRAE - Praenomen (personal name)
- B-PERS:NOMEN - Nomen (clan/family name)
- B-PERS:COG - Cognomen (family branch or nickname)
- B-PERS:FILI - Filiational terms (e.g., filius, libertus)
- B-PERS:AG - Agnomen (honorific)
- B-PERS:TITLE - Personal title like consul, pontifex
- B-TITLE - Official state/military/religious title
- B-LOC - Geographical names only (e.g., Roma, Tiberis)
- I-PERS, I-PERS:AG, I-PERS:TITLE - Inside-tag variants
- B-PERS - Only use this if:
    - The token is a standalone name, and
    - It cannot be confidently classified as PRAE, NOMEN, or COG.
    - Do NOT use B-PERS for names that belong to known Roman name
    structures.

Use the BIO format:
- "B-" means beginning of an entity
- "I-" means continuation
- "O" means not an entity

Output Format:
Return your output as exactly one valid JSON object, on a single line,
structured as:
{"tokens": [...], "tags": [...]}

Constraints:
- The number of tags MUST match exactly the number of tokens.
- Do not include explanations or comments.
- Do not output Markdown, ellipses, or formatting.
- Only return the final JSON.

You will be given a list of tokens:
<tokens>
{{TOKENS}}
</tokens>

Label each token with its BIO tag. Be strict with format and tag rules.
```

Figure 4: Zero-shot prompt used for LLM-based Latin NER.

```
You are a Latin epigraphy Named Entity Recognition (NER) expert.
Your task is to assign BIO tags to each token of Latin funerary
inscriptions.

Use only the following entity tags:
- B-PERS:PRAE - Praenomen (personal name)
- B-PERS:NOMEN - Nomen (clan/family name)
- B-PERS:COG - Cognomen (family branch or nickname)
- B-PERS:FILI - Filiational terms (e.g., filius, libertus)
- B-PERS:AG - Agnomen (honorific)
- B-PERS:TITLE - Personal title like consul, pontifex
- B-TITLE - Official state/military/religious title
- B-LOC - Geographical names only (e.g., Roma, Tiberis)
- I-PERS, I-PERS:AG, I-PERS:TITLE - Inside-tag variants
- B-PERS - Only use this if:
  - The token is a standalone name, and
  - It cannot be confidently classified as PRAE, NOMEN, or COG.
  - Do NOT use B-PERS for names that belong to known Roman name
  structures.

Use the BIO format:
- "B-" means beginning of an entity
- "I-" means continuation
- "O" means not an entity

Output Format:
Return your output as exactly one valid JSON object, on a single line,
structured as:
{"tokens": [...], "tags": [...]}

Constraints:
- The number of tags **must match exactly** the number of tokens.
- Do not include explanations or comments.
- Do not output Markdown, ellipses, or formatting.
- Only return the final JSON.

Here are some examples of correctly labeled Latin inscriptions in the
JSON line format:
{"tokens": ["qui", "et", "gni", "in", "pace", "positus"], "tags": ["O",
"O", "O", "O", "O", "O"]}
{"tokens": ["Dis", "Manibus", "Anthusae", "HiYppolytus", "coniugi",
"bene", "merenti", "fecit"], "tags": ["O", "O", "B-PERS", "O", "O",
"O", "O", "O"]}
{"tokens": ["Caius", "Vibius", "Polycarpus", "Caius", "Vibius",
"Dorus", "Halus", "Tiberi", "Claudi", "Caesaris", "aedituus", "de",
"aede", "Iovis", "porticus", "Octaviae"], "tags": ["B-PERS:PRAE",
"B-PERS:NOMEN", "B-PERS:COG", "B-PERS:PRAE", "B-PERS:NOMEN",
"B-PERS:COG", "B-PERS:AG", "B-PERS:PRAE", "B-PERS:NOMEN", "B-PERS:COG",
"O", "O", "O", "O", "O", "O"]}

Label each token with its BIO tag. Be strict with format and tag rules.
```

Figure 5: Three-shot prompt used for LLM-based Latin NER.