Overview of the Named Entity Recognition Task at EvaLatin 2026

Valeria Irene Boano, Eleonora Litta, Matteo Romanello
KU Leuven, Università Cattolica del Sascro Cuore, Milano, University of Zurich
Oude Markt 13, 3000 Leuven, Largo Gemelli 1, 20123 Milan, Rämistrasse 71, 8006 Zürich
valeria.boano@kuleuven.be, eleonoramaria.litta@unicatt.it, matteo.romanello@uzh.ch

Abstract

This paper describes the organisation and results of the Named Entity Recognition and Classification (NERC) shared task, conducted as part of EvaLatin 2026. The fourth edition of this evaluation campaign for Natural Language Processing on Latin features two shared tasks, i.e. Dependency Parsing and NERC. After introducing the objective of the task and presenting the Ancient Named Entities Special Interest Group, which aims to address the specific challenges that this task presents, this overview details the annotation tagset, the data provided to the participants and their format. The evaluation metrics and the scorer are also described. Finally, the methodology used by each participating team and their results are presented and discussed.

Keywords: Latin, evaluation, named entities recognition

1. Introduction

EvaLatin 2026 marks the fourth iteration of the evaluation campaign dedicated to the benchmarking and advancement of Natural Language Processing (NLP) tools for the Latin language. Continuing the trajectory established by the inaugural edition in 2020 (Sprugnoli et al., 2020) and subsequent campaigns in 2022 (Sprugnoli et al., 2022) and 2024 (Sprugnoli et al., 2024), EvaLatin 2026 was organised as a core component of the Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA2026), co-located with the 2026 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2026).[1] The 2026 edition of EvaLatin was structured around two primary scientific challenges: Dependency Parsing (Iurescia et al., 2026) and Named Entity Recognition and Classification (NERC).[2] This report specifically details the organisation, dataset characteristics, and performance results of the NERC task, providing a critical analysis of the current state-of-the-art in identifying and categorising entities within Latin corpora. The shared task provided participants with standardised test data, which have been made publicly available to promote the iterative improvement of language technologies for Latin, together with comprehensive annotation guidelines, through a dedicated repository.[3] This NERC task also leverages the frameworks of the HIPE (Identifying Historical People, Places and other Entities) 2020 and 2022 campaigns,[4] reusing its file format and evaluation scorer as detailed in Section 3 and Section 4.

2. NERC

Named Entity Recognition and Classification (NERC) is a fundamental task in NLP that involves the automatic identification and classification of proper names in running text into predefined categories such as Persons, Places, Organisations. Although commercial NERC systems have achieved good performance in modern languages such as English, largely driven by massive datasets drawn from news corpora and the web, the application of these technologies to historical languages remains an active and critical frontier.

2.1. Ancient Named Entities

To address specific domain requirements, the Ancient Named Entities Special Interest Group has developed a Named Entity tagset tailored for Ancient Greek and Latin textual materials and their translations.[5] This schema offers labels that are specific to ancient literature while remaining mappable onto standard concepts employed in modern NLP datasets, supporting the creation of consistent annotated datasets of Named Entities in ancient texts by minimising annotator disagreement and ambiguity. The primary applications of this collaborative work is the development of robust machine learning methods and a basic level of interoperability and exchange across projects involving Named

[1] https://lrec2026.info/.

[2] EvaLatin is an initiative organised by the CIRCSE research centre at the Università Cattolica del Sacro Cuore in Milan, Italy.https://centridiricerca.unicatt.it/circse/en.html

[3] https://github.com/CIRCSE/LT4HALA/tree/master/2026/data_and_doc

[4] https://hipe-eval.github.io/HIPE-2022/about.

[5] Available from the SIG - Ancient NER github page.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 226–233.
11 May 2026.

Entities in Ancient Greek and Latin. The compilation of these guidelines aims at facilitating the integration of historical data into broader linguistic infrastructures. From a machine learning perspective, significant challenges arise from the high degree of inflexion of the target language, flexible word order, and complexities regarding tokenisation and lemmatisation. These difficulties are further compounded by the potential absence of capitalisation in some manuscript traditions, inconsistent spelling conventions, and the extreme heterogeneity of historical domains across different eras of Latinity.

The EvaLatin NERC shared task aims to improve a state of the art that is not optimal. With regard to ancient languages, NERC systems currently show different degrees of harmonisation, and Latin is not an exception in this respect. The diversity of the data currently available is an issue we are aware of and that needs to be addressed. This evaluation campaign aims to address this issue, and find strategies to deal with it successfully.

2.2. Tagset

The tagset developed by the Ancient Named Entities Special Interest Group provides a set of semantic labels to classify Ancient Named Entities. The tags provided include two hierarchical levels. Primary first-level tags include *person* (individuals/deities), *place* (geopolitical/geographical), *collective* (groups), *creature* (non-human), and *non-consecutive-entity* (split entity strings). The second-level tags capture nuanced semantic functions, such as *.ancestry*[6] (for familial relationships and lineage, e.g. "sons of Priamus"), *.ethnic* (identifying individuals or groups via membership in a geographically defined group), *.derivative* (used for adjectives derived from names/toponyms where the common noun is not a rigid designator, e.g. "Iberian" or "Platonic"), or *.epithet* (for capitalised nicknames or titles used as unambiguous identifiers, e.g. "Phoebus" for Apollo). First-level tags represent the type of object being designated by the named entity. The usage of first-level tags is intended to be strict, to ensure interoperability. Second-level tags are designed with more consideration for the semantic function of the name: they may point to the way in which a referent is designated (for example, an individual identified with an epithet rather than their own name) or the specifically identifiable subgroup to which an entity belongs (e.g. a political or ethnic collective). They are designed to be used, expanded, or selected for project-specific goals and contexts.

[6]The structural formula for a complete fine-grained annotation is defined by the concatenation of the primary category and the secondary sub-tag, separated by a full stop.

3. Data

This section describes the data provided for the NERC shared task. The EvaLatin data format, for both the sample and test subsets, is based on the one developed for the HIPE 2020 and 2022 NERC shared tasks.[7] Detailed counts by entity type can be found in Tables 1a and 1b.

3.1. HIPE format and tagging scheme

The HIPE format consists of a simple tab-separated column textual format, comparable to the CoNLL format. The annotation scheme follows the IOB (Inside-Outside-Beginning) format.

3.2. Sample Data

The Sample Data consisted of approximately 2,900 annotated tokens from Vergil (*Aeneid*) and Sallust (*Bellum Catilinae*). This set contained 88 mentions, primarily distributed across the *collective* (24), *person* (32), and *place* (27) categories, serving as a baseline for formatting and preliminary model alignment.

3.3. Test Data

The Test Data, totalling 26,308 tokens and 2,665 entity mentions, introduced significant complexity through three authors with distinct stylistic and thematic profiles: Tacitus, Pliny the Elder, and Ovid. These specific authors were selected to ensure the representation of both prose and poetic genres, each offering a rich and diverse array of named entities. A selection from Pliny the Elder's *Naturalis Historia* (Book IV, 1–85), an encyclopaedic treatise covering various aspects of the natural world, was included due to its high density of named entities pertaining to geography and demographic populations. Tacitus's *Germania*, an ethnographic treatise on the Germanic peoples and their territories, was incorporated to evaluate the model's capacity to identify not merely standard locations and persons, but also complex references to ethnicity and geographic affiliation. This serves as a critical benchmark for fine-grained named entity classification. Finally, Books IV and XIV of Ovid's *Metamorphoses* were specifically selected to challenge the system's ability to accurately distinguish between human entities and mythological or divine figures' epithets.

[7]https://hipe-eval.github.io/HIPE-2022/about.

(a) Coarse-grained types

Entity type	Sample	Test
collective	24	422
creature	0	23
event	1	0
language	0	4
misc	4	1
object	0	0
person	32	509
place	27	1704
time	0	1
work	0	1
Total	**88**	**2665**

(b) Fine-grained types

Entity type	Sample	Test
collective	3	7
collective.ancestry	0	13
collective.animal	0	0
collective.astronomy	0	0
collective.derivative	0	11
collective.epithet	0	0
collective.ethnic	20	391
collective.organization	1	0
creature	0	17
creature.animal	0	2
creature.astronomy	0	4
event	1	0
language	0	4
misc	4	1
object	0	0
person	27	364
person.ancestry	2	31
person.author	0	46
person.derivative	1	22
person.epithet	1	31
person.ethnic	0	15
place	25	1647
place.astronomy	0	1
place.derivative	2	56
place.epithet	0	0
time	0	1
work	0	1
Total	**87**	**2665**

Table 1: Entity counts by type in the released sample and test sets.

3.4. File structure

The test file encodes the annotations needed for the NER task, and contains the following lines:

- empty lines, which mark the boundaries between different documents;
- comment lines, starting with the character '#' followed by a space, which provide metadata about the document;
- annotated lines, which contain a token followed by its tab-separated annotations.

The annotation scheme originates from the CLEF-HIPE-2022 Shared Task and consists of 10 different columns, 8 of which contain detailed named entity annotation tags. However, for the evaluation phase of this specific shared task, only two columns are relevant, i.e. NE-COARSE-LIT and NE-FINE-LIT (as detailed in the list below). The tagset used corresponds to the one described in Section 2.2. In the HIPE format, each column contains specific information. The columns used for the task are indicated in bold:

1. **TOKEN**: the annotated token;
2. **NE-COARSE-LIT**: the coarse type (IOB-type) of the entity mention token, according to the literal sense;
3. NE-COARSE-METO: the coarse type (IOB-type) of the entity mention token, according to the metonymic sense;
4. **NE-FINE-LIT**: the fine-grained type (IOB-type.subtype) of the entity mention token, according to the literal sense;
5. NE-FINE-METO: the fine-grained type (IOB-type.subtype) of the entity mention token, according to the metonymic sense;
6. NE-FINE-COMP: the component type of the entity mention token;
7. NE-NESTED: the coarse type of the nested entity (if any);
8. NEL-LIT: the Wikidata Qid of the literal sense, or 'NIL' if an entity cannot be linked. Rows without link annotations have value '_';
9. NEL-METO: the Wikidata Qid of the metonymic sense, or 'NIL';
10. **MISC**: a flag which can take the following values:
    - NoSpaceAfter, indicating the absence of white space after the token; this information can be used to rebuild the original dataset;
    - EndOfLine, indicating the end of a layout line;
    - EndOfSentence, indicating the end of a sentence;

- Partial-START:STOP, marking the character range of a mention that covers only part of a token (most commonly the enclitic *-que*). The offsets are zero-based and follow Python slice semantics, where "abcd"[1:3] yields "bc";
- Non-specified values are marked by the underscore character "_".

For the evaluation phase, the test data have been provided in one single file adhering to the IOB format presented above, where the true NE values were "masked" and represented by underscores in every annotation column.

4. Tasks and Evaluation

4.1. Tasks

The EvaLatin NERC Shared Task focuses on two task types:

- Task 1 - NERC Coarse-grained: this task includes the recognition and classification of entity mentions according to coarse-grained types, where only first level categories should be recognised;
- Task 2 - NERC Fine-grained: this task includes the recognition and classification of entity mentions according to fine-grained types, where also second level categories should be recognised.

4.2. Metrics

NERC is evaluated in terms of macro and micro Precision, Recall, F1-measure. Two evaluation settings are considered: strict (exact boundary matching) and fuzzy (relaxed boundary matching, e.g. at least one token overlap). Each column is independently evaluated according to the following metrics:

- Micro average P, R, F1 at entity level (not at token level), i.e. consideration of all true positives, false positives, true negatives, and false negatives over all documents:
  - strict and fuzzy;
  - separately per type and cumulative for all types.
- Document-level Macro average P, R, F1 at entity level (not on token level), i.e. average of separate micro evaluation on each individual document:
  - strict and fuzzy;
  - separately per type and cumulative for all types.

Our definition of "macro" differs from the usual one, and macro measures are computed as aggregates on document-level instead of entity-type level. Specifically, macro measures average the corresponding micro scores across all the documents. This means that they reflect variance in document length but are insensitive to class imbalances. Note that in the strict scenario, predicting wrong boundaries leads to heavy penalty of one false negative (entity present in the gold standard but not predicted by the system) and one false positive (entity predicted by the system but not present in the gold standard). Although this may be severe, we keep this metric in line with CoNLL standards and refer to the fuzzy scenario if the boundaries of an entity are considered as less important.

4.3. Scorer

EvaLatin NERC evaluations were conducted using the CLEF-HIPE-2020-scorer,[8] a Python module for evaluating NERC and Entity Linking (EL) systems, developed and used in the context of the HIPE shared tasks on NE processing on historical documents. The scorer evaluates at the entity level, whereby entities (most often multi-words) are considered as the reference units, with a specific type as well as a token-based onset and offset. There are different evaluation regimes depending on how strictly entity type and boundary correctness is judged. The scorer provides strict and fuzzy evaluation regimes, where "strict" requires exact match of both entity type and entity boundaries, and "fuzzy" requires exact match of entity type and at least one token overlap. System performance has been computed, reported and published in terms of micro and macro P, R, and F1 for each Task. For each task, the systems have been ranked according to their F1 scores.

5. Results and Discussion

5.1. Team results

Three teams took part in the task: **argo-navis**, **KULeuven** and **uOttawa**. Argo-navis and uOttawa both submitted two runs for each task, while KULeuven submitted one run.

The **argo-navis** team approached the task with a zero-shot cross-lingual strategy (Ripoll-Alberola, 2026). The openly available GliNER2 model[9] (Zaratiana et al., 2025), an extension of the original GliNER model (Zaratiana et al., 2024), was used. GliNER concatenates task specifications with input text, jointly computes contextualised span and task

[8] https://github.com/hipe-eval/HIPE-scorer

[9] https://huggingface.co/collections/fastino/gliner2-family.

```
TOKEN	NE-COARSE-LIT	NE-COARSE-METO	NE-FINE-LIT	NE-FINE-METO	NE-FINE-COMP	NE-NESTED	NEL-LIT	NEL-METO	MISC
# evalatin2026:version = v0.1
# evalatin2026:title = Aeneid
# evalatin2026:original_source = data/preparation/corpus/la/curated/aeneid1-110_lat.xmi
# evalatin2026:license = CC-BY
# evalatin2026:document_id = aeneid1-110_lat
# evalatin2026:author = Vergil
# evalatin2026:applicable_columns = TOKEN NE-COARSE-LIT NE-FINE-LIT MISC
Arma	O	_	O	_	_	O	_	_	_
virumque	O	_	O	_	_	O	_	_	_
cano	O	_	O	_	_	O	_	_	NoSpaceAfter
,	O	_	O	_	_	O	_	_	_
Troiae	B-place	_	B-place	_	_	O	_	_	_
qui	O	_	O	_	_	O	_	_	_
primus	O	_	O	_	_	O	_	_	_
ab	O	_	O	_	_	O	_	_	_
oris	O	_	O	_	_	O	_	_	NoSpaceAfter
Italiam	B-place	_	B-place	_	_	O	_	_	NoSpaceAfter
```

Figure 1: Example of output file format.

embeddings via a shared encoder and feedforward networks, then scores each span-task pair with a dot-product sigmoid to predict entity probabilities. The team submitted two runs for each task, using two variants of the model: the multilingual[10] and the monolingual large.[11] Since GliNER accepts optional natural language descriptions for each task, the model was provided with the tag descriptions extracted from the annotation guidelines, after removing the examples to adapt them to the zero-shot approach. As GliNER is not trained on Latin texts, a cross-lingual approach was required: the input data was first translated into English using the deep-translator Python package[12] to query the Google Translate API. The results were then aligned back to the original Latin text using simalign (Jalili Sabet et al., 2020), using the embeddings of UGARIT/grc-alignment (Yousef et al., 2022). The team achieved third place in Task 1 (NERC Coarse-grained) and second place in Task 2 (NERC Fine-grained). The results showed different performance according to which variant of the model was used. The large model achieves higher precision, whereas the multilingual model delivers higher recall and a better performance on the fine-grained task, which in any case remains more critical.

The **KU_Leuven** team participated proposing a methodology that relies on a substitution-based probing strategy to identify entities within a text (Maria Mihaela Trusca and de Cruys, 2026). The core procedure involves taking a sentence and sequentially replacing individual words with predefined "probes", representative candidate entities of specific target types, such as "person" or "place". To quantify compatibility, the system calculates a Pseudo-Log-Likelihood (PLL) score, which measures how naturally a probe fits into the surrounding textual context. If the PLM predicts a natural fit, the original word is inferred to be functionally similar and likely an entity of that predefined type. These scores are then normalised across entity types. The primary language model utilised was LaBERTa, a RoBERTa-based encoder pretrained on Latin corpora, as preliminary tests showed it outperformed LatinBERT. To train the NER models, the team combined five datasets, one of which was generated by translating existing multilingual HIPE datasets into Latin using the Google Translate API. To align the NER labels, the researchers used a two-step procedure that balanced semantic similarity (via LaBERT embeddings) and orthographic similarity (via Levenshtein distance). The KULeuven team sent one run for each task. In Task 1 (NERC Coarse-grained), the team used a fine-tuned LaBERTa model integrated with the PLL-based auxiliary decoding signal, ranking second among competitors both in the strict and in the fuzzy evaluation. In Task 2 (NERC Fine-grained), the team applied their PLL-based scoring method in a purely zero-shot setting, utilising defined lists of probes for all category pairs, and achieved third place.

Team **uOttawa** addressed the task by leveraging commercial Large Language Models (LLMs), specifically Gemini-2.5-pro and Claude-sonnet-4-5, via prompt engineering (Chan, 2026). Their methodology involved creating prompt templates that contained a task overview, specific requirements, descriptions of the entity classes, and output guidelines mandating the Inside-Outside-Beginning (IOB) format. The team experimented with both zero-shot and few-shot learning strategies. They determined that few-shot prompting consistently outperformed zero-shot approaches because it provided essential contextual depth, grammatical cues, and domain knowledge. For their few-shot templates, they embedded three randomly sampled Latin sentences alongside their correct entity anno-

[10] https://huggingface.co/fastino/gliner2-multi-v1.

[11] https://huggingface.co/fastino/gliner2-large-v1.

[12] https://github.com/nidhaloff/deep-translator.

tations, drawn from a supplementary dataset. The target text was processed sentence by sentence, and light post-processing was applied to the LLM outputs to clean up formatting and automatically classify unlabelled tokens as non-entities. During preliminary testing, uOttawa discovered a distinct divergence in the strengths of their chosen models. They found that Claude-sonnet-4-5 performed best on the coarse-grained task, likely because it is adept at confidently assigning probabilities across a limited set of broad options. Conversely, Gemini-2.5-pro excelled at the fine-grained subtask, as it seemingly utilizes a more cautious predictive approach that is advantageous when navigating large label spaces with subtle semantic distinctions. Consequently, they submitted their official few-shot runs as 'nerc_1' (utilising Gemini) and 'nerc_2' (utilising Claude). A notable characteristic of uOttawa's results was that recall consistently exceeded precision across all models, prompting techniques, and subtasks. The team attributed this to an over-generation tendency, where the English-dominant pre-training data of the LLMs caused them to mistakenly flag Latin common nouns as proper entities if those words function as names in modern derivative languages (e.g., misidentifying the Latin common noun "gloria" as a name). Furthermore, while their few-shot method proved highly effective overall, the researchers acknowledged that the performance drop observed during the fine-grained task indicates that highly specialised, historically contextualised classical labels remain a significant hurdle for generalised commercial LLMs.

5.2. Comparative results

Limited availability of data shaped the approaches adopted by participants to this shared task. The absence of a sizeable training set meant that approaches relying on fine-tuning of transformer-based models were not applicable. As a result, the submitted systems use techniques like zero-shot, few-shot prompting, as well as cross-language transfer, to circumvent this challenge of limited data availability. With the exception of KULeuven, none of the other participating teams made substantial usage of data augmentation and synthetic data generation. The official results reveal that LLM-based few-shot prompting currently defines the upper bound for Latin NER, while probing-based and cross-lingual systems provide valuable, more computationally efficient alternatives.[13]

[13]Only Coarse and Fine Strict ranked results are reported here, for a comprehensive list of results for all teams check the results document in the LT4HALA repository.

System	P	R	F1
uOttawa_nerc-coarse_2	0.899	0.917	0.908
uOttawa_nerc-coarse_1	0.876	0.914	0.895
KULeuven_nerc-coarse_1	0.736	0.694	0.714
argo-navis_nerc-coarse_2	0.575	0.498	0.534
argo-navis_nerc-coarse_1	0.552	0.5	0.525

Table 2: NERC coarse strict

System	P	R	F1
uOttawa_nerc-fine_1	0.841	0.89	0.865
uOttawa_nerc-fine_2	0.841	0.871	0.856
argo-navis_nerc-fine_1	0.466	0.327	0.384
argo-navis_nerc-fine_2	0.491	0.218	0.302
KULeuven_nerc-fine_1	0.131	0.128	0.129

Table 3: NERC fine strict

The uOttawa system dominated both subtasks, with the Claude-sonnet-4-5 run (nerc_2) achieving the highest Coarse-Strict score. This performance confirms that generative LLMs, when provided with explicit formatting and semantic guidance, can effectively bridge the low-resource gap. The KULeuven results highlight a critical methodological distinction: while they were highly competitive in the Coarse task, their Fine Strict score (0.129) was a zero-shot result using only the PLL method without fine-tuning, as no guideline-compliant training data was available for the fine-grained subcategories. Similarly, argo-navis demonstrated that cross-lingual pipelines can achieve a viable baseline (F1 0.534), although they face precision drops in fine-grained tasks. Specifically, this model tends to favour the coarse categories, while rarely committing to the second-level distinctions. The results of the evaluation campaign show that the fine-grained task remains the most challenging. Persons and places are the best recognised categories across all three teams' attempts, reflecting their predominance in most NER multilingual pretraining data. The *creature* category was consistently poorly identified. Table 4 lists F1 performances per team per highest frequency first level entities.

Fine-grained categories such as *.epithet*, *.ancestry*, and *.author* posed a significant hurdle. On many occasions, models yielded zero scores here, illustrating that identifying that a token is a name is significantly easier than determining its semantic role (e.g., distinguishing a person mentioned as an author from a person mentioned as an ancestor). The guidelines defined the *.ethnic* tag for persons or collectives identified by their membership in a geographically or ethnically defined group, and the *.derivative* tag for adjectives derived from toponyms, personal names, or group names. The performance on these categories varied drastically

Entity Label	KULeuven	argo-navis_1	argo-navis_2	uOttawa_1	uOttawa_2
PERSON	0.663	0.43	0.568	0.839	0.855
PLACE	0.774	0.539	0.549	0.923	0.942
COLLECTIVE	0.576	0.684	0.507	0.899	0.88
CREATURE	0	0.475	0.333	0.333	0.318

Table 4: Strict F1 Scores by Team on First Level Entities

depending on the computational methodology employed by the participating teams. Team uOttawa achieved the most robust performance on derived adjectives by employing a few-shot prompting strategy using commercial Large Language Models (LLMs). Table 5 lists F1 performances per team per highest frequency second level entities.

A critical reflection on this campaign involves the tension between high performance and scientific reproducibility. The participating systems represent two distinct architectural philosophies: commercial LLMs, with uOttawa system achieving the highest scores by leveraging commercial models like Claude-sonnet-4-5 and Gemini-2.5-pro, versus open and specialised architectures with the argo-navis and KULeuven teams using openly available models such as GliNER2 and LaBERTa. While highly effective, the commercial LLM models operate as "black boxes" with proprietary training data and non-static versions, which still represents a challenge for the long-term reproducibility of results in a scientific context. Conversely, the open source approaches, while yielding lower absolute scores in this iteration, offer a higher degree of transparency and allow for consistent auditing and local deployment by the research community.

6. Conclusions

The EvaLatin 2026 NERC shared task demonstrates that the landscape of Natural Language Processing for historical languages is continuing to evolve. The results indicate that LLM-based few-shot prompting could potentially represent the state-of-the-art for this domain, effectively bridging the gap created by the absence of large-scale, domain-specific training sets. While coarse-grained entity recognition has reached a high level of maturity, the fine-grained classification of semantic roles remains a significant challenge. Specifically, distinguishing between the nuanced functions of entities continues to be a hurdle for both generalised and specialised models. Given the performance trends observed in the EvaLatin 2026 results, it is evident that the low score problem for fine-grained categories is the next major hurdle for Latin NERC. While models can reliably identify that a token is a name, they often fail to assign specific semantic roles like .epithet or .ancestry due to a lack of representative examples in existing corpora. To move beyond the limitations of existing manual annotations, more strategies involving Synthetic Data Generation can be used to "teach" models these subtle distinctions.

7. Acknowledgements

Our sincere thanks go to Margherita Fantoli and Laura Soffiantini for helping out with the compilation and gold standard annotation of the test data for this task. A special mention goes to all the other members of the Ancient Named Entities Special Interest Group, for their continuous feedback on how to better the tagset.

8. Bibliographical References

Callum Chan. 2026. Transfer learning for named entity recognition of classical latin through llm prompting. In *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026)*, Palma, Mallorca (Spain). ELRA.

Federica Iurescia, Marco Passarotti, and Rachele Sprugnoli. 2026. Overview of the dependency parsing task at evalatin 2026. In *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026)*, Palma, Mallorca (Spain). ELRA.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Violet Soen Ine de Daele Kevin Verbruggen Maria Mihaela Trusca, Mark Depauw and Tim Van de Cruys. 2026. Contextual probing for low-resource named entity recognition in latin. In *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient*

Fine-Grained Category	KULeuven_1	argo-navis_1	argo-navis_2	uOttawa_1	uOttawa_2
COLLECTIVE.ETHNIC	0.016	0.129	0.027	0.910	0.852
PERSON.AUTHOR	0.000	0.125	0.346	0.624	0.538
PLACE.DERIVATIVE	0.000	0.000	0.000	0.739	0.623
PERSON.ANCESTRY	0.056	0.000	0.000	0.638	0.358
PERSON.ETHNIC	0.000	0.000	0.000	0.529	0.200
PERSON.EPITHET	0.030	0.000	0.000	0.500	0.382
COLLECTIVE.ANCESTRY	0.000	0.000	0.000	0.296	0.071
CREATURE.ANIMAL	0.000	0.000	0.000	1.000	0.000

Table 5: F1 Scores by Team for Fine-Grained NERC Categories

Languages (LT4HALA 2026), Palma, Mallorca (Spain). ELRA.

Ripoll-Alberola. 2026. Classificatio sine iactu – that is, zero-shot nerc in latin. In *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026)*, Palma, Mallorca (Spain). ELRA.

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, Torino, Italia. ELRA and ICCL.

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.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 2020. Overview of the EvaLatin 2020 evaluation campaign. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 105–110, Marseille, France. European Language Resources Association (ELRA).

Tariq Yousef, Chiara Palladino, David J. Wright, and Monica Berti. 2022. Automatic translation alignment for Ancient Greek and Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107, Marseille, France. European Language Resources Association.

Urchade Zaratiana, Gil Pasternak, Oliver Boyd, George Hurn-Maloney, and Ash Lewis. 2025. Gliner2: An efficient multi-task information extraction system with schema-driven interface. *arXiv preprint arXiv:2507.18546*.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. Gliner: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376.