The Meaning of Beatus: Disambiguating Latin with Contemporary AI Models

Eleonora Ghizzota[1,*,†], Pierpaolo Basile[1,*,†], Lucia Siciliani[1,*,†] and Giovanni Semeraro[1,†]

[1]Department of Computer Science, University of Bari Aldo Moro, via Edoardo Orabona 4, 70125, Bari, Italy

Abstract

The objective of this work is to assess the performance of Large Language Models (LLMs) on the task of Word Sense Disambiguation (WSD) for Latin.We evaluate state-of-the-art LLMs—including GPT-4o-mini and LLaMA variants—in both zero-shot and fine-tuned settings, using a dataset derived from the SemEval-2020 Latin Lexical Semantic Change task. Our study aims to determine whether instruction tuning and task-specific fine-tuning can significantly improve the models' ability to disambiguate Latin word senses. Results show that while LLMs demonstrate a non-trivial baseline ability in zero-shot settings, fine-tuning – particularly instruction-based – provides improvements in accuracy and F_1 scores. These findings highlight the potential of LLMs when applied to low-resourced historical languages.

Keywords

Lexical Semantics, Word Sense Disambiguation, Large Language Models, Latin, Low-resource languages

1. Introduction and Motivations

In terms of data availability and the impact of the study, some languages are more represented than others. Naturally, when developing a new Language Model or collecting data for a benchmark, most computational and research efforts focus on English. However, English is just one out of thousands of spoken languages, and many research teams continue working to fill this representation gap.

Latin is a suitable example of a former low-resource language to which many efforts were dedicated for creating *ad hoc* resources and datasets. Moreover, Latin is a perfect fit for several Natural Language Processing tasks thanks to a number of factors: *(i)* accessible digital data covering two thousand years of history, e.g., LiLa [1, 2, 3], LatinISE [4, 5, 6], Latin WordNet [7], *(ii)* available computational resources specially designed for Latin, e.g., Classical Language Toolkit [8], UDPipe [9, 10], *(iii)* ancient languages offer the opportunity to analyse long-term lexical semantic change and Latin itself is a prime example of a language that is not only ancient, but has also continued to be actively used long after the end of antiquity: the usage of Latin in written works covers a period of over 22 centuries, spanning from 200 BCE to modern-days. This temporal extension results in a wealth of textual data [11, 12] in which the language has undergone various diachronic evolution. Regardless of the few projects focusing on Latin, especially for semantic and syntactic annotations, very few evaluation campaigns and challenges are proposed, i.e., SemEval-2020 [13], EvaLatin [14, 15, 16], and when it comes to language modelling even fewer studies on Latin have been conducted, i.e., Latin BERT[1] [17, 18]. Nevertheless, the path to achieving equal representation of Latin is still far-reaching, especially when it comes to annotated datasets for automated learning, as well as language-specific generative models.

One of the historical [19] Natural Language Processing (NLP) tasks that suffers the most from the lack of resources is Word Sense Disambiguation (WSD), defined in [20] as "*the computational identification of meaning for words in context*". Having access to a language-specific model and extensive corpora is vital for the Word Sense Disambiguation task. As a matter of fact, [21] define the so-called *knowledge acquisition bottleneck* that characterizes WSD: it heavily relies on machine-readable knowledge resources that not only require extensive manual effort for their creation, but they also need to be updated or created from scratch anytime a variation occurs, e.g., a word has gained or lost a sense.

Over the years, techniques for tackling WSD have evolved significantly in tandem with advancements in Artificial Intelligence (AI) and Machine Learning (ML). Initially, the field was dominated by rule-based systems, which eventually transitioned to knowledge-based approaches as digital sense inventories became more accessible. The advent of digital corpora paved the way for supervised learning methodologies, utilising manually annotated datasets to improve WSD effectiveness.

The proliferation of web content has further revolu-

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

*Corresponding author.

†These authors contributed equally.

e.ghizzota@phd.uniba.it (E. Ghizzota); pierpaolo.basile@uniba.it (P. Basile); lucia.siciliani@uniba.it (L. Siciliani); giovanni.semeraro@uniba.it (G. Semeraro)

0000-0002-0751-3891 (E. Ghizzota); 0000-0002-0545-1105 (P. Basile); 0000-0001-7116-9338 (L. Siciliani); 0000-0002-9421-8566 (G. Semeraro)

[1]https://www.github.com/dbamman/latin-bert

tionised the landscape by providing vast corpora and extensive knowledge graphs extracted from online sources, thereby amplifying the capabilities of both supervised and knowledge-based methods. The introduction of transformer-based architectures [22] marked a significant turning point. These models use dense vector representations to capture semantic meaning in context, resulting in further advancements in disambiguation techniques. A significant development in this domain is the rise of Large Language Models (LLMs), which are built upon the Transformer architecture and trained on extensive text corpora. LLMs exhibit proficiency in a myriad of tasks in zero-shot or few-shot contexts, ruling out the necessity of task-specific training data. This implies an inherent capacity for semantic understanding within these models. Nonetheless, LLMs can also be fine-tuned on particular tasks by utilising tailored training data, enhancing their performance in specific applications.

Considering these premises, the intent of this work is to assess how state-of-the-art LLMs perform on under-represented languages like Latin through the lens of a long-standing task in NLP like WSD. In particular, our investigation has two objectives. First, we want to test models out-of-the-box ability to disambiguate Latin senses in a zero-shot setting. In this way, we aim to first establish how well the models inherent multilingual knowledge performs in accurate sense prediction. Next, we also perform task-specific fine-tuning, which enables us to adapt both standard and instruction versions of LLMs. The aim is to gauge the gain obtained with this additional training step.

The paper is structured as follows: Section 2 provides an overview of works related to solving the WSD task with LLMs; Section 3 introduces the corpus of choice for this study, while 4 illustrates the methodology. Section 5 describes the experimental setting and discusses the results and the limitations of the proposed strategy, while Section 6 summarises the takeaway messages of this paper and suggests some future works.

2. Related Work

2.1. Latin Word Sense Disambiguation

Currently, solving the WSD task for Latin using language models remains an unexplored strategy, with very few works investigating this line of research in recent years. The idea of using WSD for measuring the ability of language models to deal with Latin is supported by the work proposed by [17] in which Latin BERT is tested on the sense disambiguation task.

Latin BERT is a contextual language model tailored for Latin, trained on a corpus of 642.7 million words drawn from diverse sources ranging from the Classical period to the 21st century. It achieves state-of-the-art performance in Latin part-of-speech tagging across all Universal Dependency datasets. To capture the full range of linguistic variation, the model was trained on multiple corpora, including the Corpus Thomisticum, the Internet Archive, the Latin Library, Patrologia Latina, Perseus, and the Latin Wikipedia. Latin BERT uses Latin-specific sentence and word tokenizers from the Classical Language Toolkit, resulting in a vocabulary of 32,895 subword units. To assess Latin BERT performance in the WSD task, the authors reformulated it into a binary classification task and created an *ad hoc* dataset of Latin sense examples extracted from the *Lewis and Short Latin Dictionary* [23]. In order to be selected, headwords must have at least two distinct senses – typographically denoted by "I." and "II." – supported by at least 10 sentences each, and longer than five words. For the task, only the two major senses of a headword were retained; the final dataset consists of 8,354 examples for 201 dictionary headwords. For each headword, an instance of Latin BERT was fine-tuned on 80% of the examples. The number of training instances per headword ranges from 16 (8 per sense) to 192 (96 per sense); 59% of headwords have 24 or fewer training examples. Latin BERT achieves 75.4% accuracy, compared to the 67.3% of a bidirectional LSTM with static word embeddings. These results show that, even with few training examples, Latin BERT was able to disambiguate senses.

A few years later, [24] fine-tuned Latin BERT on a portion of sense representations in the *Thesaurus Linguae Latinae*[2] (TLL). The TLL is the first comprehensive dictionary of ancient Latin usage up to 600 AD, offering a comprehensive, documented overview of every Latin word's history, including meanings and constructions, etymology, inflexion peculiarities, spelling, and prosody, as well as comments from ancient sources on the word itself. The ongoing TLL project begun in 1894 and has been regularly updated since; currently, it contains lemmata from *a* to *resurgēsco*, and it is estimated to contain approximately 56,000 entries. Inspired by the WSD dataset created by Bamman and Burns for Latin BERT, the authors requested data for the same lemmata from TLL, obtaining 25,227 quotes for 40 lemmata. The new dataset leads to a performance gain, with the Mean Macro F_1 increasing from .695 to .794.

Although both [17] and [24] achieved promising results, Latin is still an under-represented language for which very few annotated resources are available, when compared to English. [25] proposes a language pivoting framework for Latin. Language pivoting, borrowed from Machine Translation [26], consists of propagating annotations from high-resource languages to lower-resource ones. Starting from the 40 lemmata manually annotated

[2] https://tll.degruyter.com/about

for SemEval-2020 [13], the authors extract an aligned Latin-English dataset in which these lemmata occur. To this day, the dataset of SemEval-2020 Task 1 is the only benchmark for Latin, manually annotated by Latin experts. These lemmata were then mapped to WordNet, Latin WordNet[3] and Princeton WordNet [27], allowing for annotation propagation from English to Latin. The final result is a dataset of 3,886 annotated sentences for training and experimentation.

2.2. LLMs and Word Sense Disambiguation

Over the years, LLMs have consistently demonstrated their ability to perform various tasks in a zero- or few-shot setting with minimal or no specific training data, suggesting an intrinsic capability of LLMs to grasp the semantics behind language [28, 29].

[30] demonstrates that BERT-like models are capable of effectively differentiating between various word senses, even when only a few examples are available for each. Their analysis further reveals that although language models can perform nearly perfectly on coarse-grained noun disambiguation in ideal settings where training data and resources are abundant, such conditions are rare in practical scenarios, presenting ongoing challenges. Along the lines of BERT-like approaches, [31] examines multiple WSD methods, including those that use language models to extract contextual embeddings as input features and as a foundation for training supervised models on sense-annotated data. [32] assesses language models' WSD capabilities through three behavioural experiments designed to evaluate children's ability to disambiguate word senses. The study offers a compelling comparison between how children understand semantics and how it is encoded in transformer-based models. The authors identify a bias in the models toward the most frequent sense and observe a negative correlation between the size of the training data and model performance.

[33] evaluated WSD accuracy of LLMs on eight datasets via a multiple-choice question format, and [34] extended the analysis by gauging LLM performance on single-choice questions and examining how different model sizes affect disambiguation accuracy. Similarly, [35] creates a benchmark specific for the Italian language with the aim of evaluating LLMs' abilities in selecting the correct meaning of a word and in generating the definition of a word in a sentence. Finally, [36] analyses WSD capabilities of only open LLMs experimenting with different parameter configurations on several languages: English, Spanish, French, Italian and German. The authors extend the existing XL-WSD benchmark [37] to include two additional subtasks: (*i*) given a word occurrence within a sentence, the LLM must generate the appropriate definition; and (*ii*) given a word occurrence and a list of predefined meanings, the LLM must identify the correct one. Moreover, they use the training data of XL-WSD to fine-tune an open LLM based on LLaMA3.1-8B. The results indicate that while LLMs perform well in zero-shot settings, they still fall short of surpassing current state-of-the-art methods. Larger models achieve the strongest results, whereas medium-sized models tend to underperform. Notably, however, a fine-tuned model with a medium parameter size outperforms all others, including existing state-of-the-art approaches.

3. Dataset

3.1. Resource

The dataset of choice is the Latin annotated dataset for the Unsupervised Lexical Semantic Change Detection (LSCD) shared task of SemEval-2020 [13].

This dataset is a fragment of LatinISE[4] [5], a 13 million words diachronic, annotated Latin corpus. The primary source of LatinISE is the Latin portion of the IntraText digital library[5]. To semi-automatically annotate this corpus, 2013 state-of-the-art NLP tools – PROIEL[6], Quick Latin[7], and TreeTagger[8]– were used. Hence, LatinISE provides morphological annotations like part-of-speech tags and lemma for each word.

Back in 2020, for the SemEval-2020 Unsupervised Lexical Semantic Change task, two time-specific sub-corpora C_1 and C_2 were extracted from LatinISE [13, 6]: C_1 covers the period from 2^{nd} century BC to 0 (1.7M tokens), C_2 from 0 to 21^{st} century AD (9.4M tokens).

As concerns target words, they are either (*i*) words that changed their meaning(s) between C_1 and C_2; or (*ii*) stable words that did not change their meaning during that time. The choice of the set of lexemes for the annotation was based on an initial process of lexical selection and pre-annotation, carried out by a team member [6]. A list of target words comprising those whose meaning has been attested to have changed between the pre-Christian and Christian era [38, 39, 40, 23] was selected. The pre-annotation trial verified whether the corpus showed evidence of both the late antiquity senses and the previous senses, and whether the late antiquity senses appeared in the later texts only and the classical senses in the earlier texts, although they may also have occurred in later texts. Conversely, stable words were chosen since they are not known for having undergone lexical semantic change

[3] http://latinwordnet.exeter.ac.uk/

[4] Available at https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2506

[5] http://www.intratext.com

[6] https://www.hf.uio.no/ifikk/english/research/projects/proiel/

[7] http://www.quicklatin.com/

[8] https://www.cis.uni-muenchen.de/âĹijschmid/tools/TreeTagger/

associated with the period of late antiquity. The final list comprises 40 target words, of which 23 are stable, while 17 have undergone changes in meaning in relation to Christianity.

For each target word, its primary sense definitions were taken from the Latin portion of the *Logeion Online Dictionary*[9], which includes Lewis and Short's *Latin English Lexicon* [23], Lewis's *Elementary Latin Dictionary* [41], and Du Fresne Du Cagne's *Glossarium mediae et infimae latinitatis* [42]. Depending on the cases, the sense inventory was simplified, or the definitions were shortened, while maintaining the principal distinction between senses. Finally, for each target word 60 passages sample sentences were extracted, 30 from C_1 and 30 C_2 respectively, for a total of 2,398 passages.

The lack of native Latin speakers adds a further layer of complexity to the sense annotation process. 10 annotators with a high-level knowledge of Latin were recruited, ranging from undergraduate students to senior researchers. Annotators – only one per target word – scored the relatedness between a usage and a sense definition according to the Diachronic Usage Relatedness (DUReL) framework [43], specially designed for lexical semantic change annotations. The DUReL framework consists of a 4-point scale for quantifying the relatedness of a word usage and a sense, or score 0 if the annotator cannot decide:

- 0 - Cannot decide
- 1 - Unrelated
- 2 - Distantly related
- 3 - Closely related
- 4 - Identical

Table 1 shows an example of the usage annotation for target word *beatus*. The senses presented to the annotators were: (a) "*blessed*", (b) "*rich*", (c) "*fortunate*", (d) "*happy*" and (e) "*rewarded*". Let's focus on the sense "*blessed*", which only emerged later with the advent of Christianity. Notice how it scores 1 for the first usage, dated 46 BC, while it scores 4 for the second usage, dated circa 1100 AD.

Target word *virtus* was chosen for calculating the inter-annotator agreement between four annotators: the average pairwise agreement computed as Spearman correlation coefficient was 0.69, comparable with inter-annotator agreement for modern languages, e.g., English 0.69, Swedish 0.57 and German 0.59 [43]. See [6] for the detailed process behind the creation and annotation of the dataset.

3.2. Data preparation

Pairs of sense and sentence were split in a stratified manner, based on the scores assigned to each sense. This stratification process, 70% training and 30% testing, outputs a training set of 6,299 sentences and a testing set of 2,690. Due to the absence of annotations, sentences of the lemma *oportet* were excluded from the dataset. DuREL annotation statistics are summarised in Table 2 below.

We take full advantage of the annotations in the dataset by creating a separate prompt for each of the judgments assigned to each of the proposed senses for a single sentence. For example, if the annotator marked *virtute* as "4 - Identical" for the "*manliness, courage, virtue, strength*" and "1 - Unrelated" for the sense "*virtue, personified as a deity*", two separate prompts are created, each structured as shown in Listings 1 and 2.

Listing 1: Prompt generated by each sense annotation for regression task.

```
Instruction: Given the target word ''virtute''
     and the sentence in input where the word
     is enclosed by the [TARGET] tag, and the
     following meaning ''virtue, personified as
      a deity'', assign a score between 0 and
     4. The score meaning is the following:
0: Cannot decide
1: Unrelated
2: Distantly Related
3: Closely Related
4: Identical
Answer just with the score.

Input: <left context> [TARGET] virtue [TARGET]
     <right context>
```

This process yields a total of 8,989 prompts for the regression task.

As for the binary classification task, the DuREL 1-to-4 scale was binary encoded as follows:

- Pairs of sense and sentence scores equal to or above 3 were labelled as YES;
- Pairs of sense and sentence scores equal to or below 2 were labelled as NO.

The prompt is the following:

Listing 2: Prompt generated by each sense annotation for binary classification task.

```
Instruction: Given the target word ''virtute''
     and the sentence in input where the word
     is enclosed by the [TARGET] tag, and the
     following meaning ''virtue, personified as
      a deity'', assign a label "yes" or "no".
The label meaning is the following:
"yes": The sense for the target word occurrence
      is correct
"no": The sense for the target word occurrence
     is not correct
Answer just with the label.

Input: <left context> [TARGET] virtue [TARGET]
     <right context>
```

[9] https://logeion.uchicago.edu/

Table 1
Two annotated usages of lemma *beatus* [6]; the first one is extracted from a classical text, Cicero's "*Tusculanae disputationes*" (46 BC), the second one from a mediaeval text, "*De libero arbitrio*" by Robertus Grossetest, 12^{th} - 13^{th} century AD. The English translations are in Appendix A.

Text	Senses				
	"*blessed*"	"*rich*"	"*fortunate*"	"*happy*"	"*rewarded*"
[...] Dico enim constanter grauiter sapienter fortiter. Haec etiam in eculeum coiciuntur, quo uita non adspirat beata. - Quid igitur? solane **beata** uita, quaeso, relinquitur extra ostium limenque carceris, cum constantia grauitas fortitudo sapientia reliquaeque uirtutes rapiantur ad tortorem nullumque recusent nec supplicium nec dolorem? [...]	1	1	3	3	2
[...] Ex quo fit, ut de nihilo creauerit omnia." Eadem itaque ratione solus facit ominia, nulla adiutus natura. Horum autem obiectorum solutio haberi potest ut uidetur ex uerbis **beati** Bernardi sic dicentis: "Ipsa gratia Liberum arbitrium excitat, cum seminat cogitatum. Sanat, cum mutat affectum; roborat, ut perducat ad actum; seruat, ne sentiat defectum." [...]	4	1	3	3	2

Table 2
DuREL annotation statistics in training and testing sets.

Label	Training	Testing	Total
0	44	15	59
1	3,536	1,514	5,050
2	495	205	700
3	771	329	1,100
4	1,453	627	2,080
	6,299	2,690	8,989

Pairs of sense and sentence with score 0 were not considered in this experiment; thus, with respect to the scores distribution in Table 2, the training set for binary classification task consists of 6,255 instances instead of 6,299, and the testing set has 2,675 examples instead of 2,690, yielding a total of and 8,930 prompts. This binary encoded dataset comprises 956 instances of class YES and 1,719 NO, resulting in a very imbalanced dataset in which class YES represents only 35.73% of the entire dataset.

The idea behind this work is to leverage this dataset for building a benchmark for the evaluation of LLMs in disambiguating Latin words as described in the following section.

4. Methodology

As stated in the introduction, one of the aims of this paper is to assess whether fine-tuning on LLMs can improve their performance on a lower-represented language, compared to a zero-shot setting. To do so, we exploit the prompt dataset created from LatinISE, described in Section 3. Tables 3 and 4 introduce the LLMs of choice and summarise their characteristics.

Table 3
Strategies applied to GPT-4o-mini and LLaMA-3 variants.

	Zero-shot	Fine-tuning
GPT-4o-mini	✓	
LLaMA-3.3-70B-instruct-turbo	✓	
LLaMA-3.1-8B-instruct	✓	
LLaMA-3.1-8B-instruct-ft		✓

4.1. Zero-shot

We assess the zero-shot capabilities of two categories of instruction-tuned LLMs:

- **LLaMA-3 instruction-tuned.** We use publicly available checkpoints of Meta's LLaMA 3.3-70B

Table 4
Technical details of analysed LLMs.

	Parameters	Training tokens	Multilinguality
GPT-4o-mini	–	–	✓
LLaMA-3.1	8B	∼15 trillion	✓
LLaMA-3.3	70B	∼15 trillion	✓

and 3.1-8B variants with instruction tuning, accessed via the TogetherAI API[10] and Unsloth API[11], respectively;

- **GPT-4o-mini.** accessed via Microsoft Azure API, this model is used without any task-specific training. Prompting is designed to simulate realistic WSD instructions.

For zero-shot WSD, we directly use the prompt test set, unseen during fine-tuning (see Section 4.2). After a preliminary prompt engineering step, we use the prompt in Listing 3, which is the same as the one used for fine-tuning.

4.2. Fine-tuning

Using the training split of the dataset, we fine-tune the open-weight LLaMA-3.1-8B model. Given the computational constraints associated with full fine-tuning of large models, we adopt a parameter-efficient fine-tuning (PEFT) approach based on Low-Rank Adaptation (LoRA).

LoRA [44] introduces trainable, low-rank matrices into each transformer layer to adapt the model to a downstream task. Instead of updating all model parameters, LoRA freezes the pre-trained weights and injects a low-rank decomposition into the linear projections of the self-attention and/or feed-forward layers. This strategy significantly reduces the number of trainable parameters and memory usage, allowing efficient fine-tuning even on consumer-grade GPUs. We use the implementation provided by the Unsloth library, which enables us to reduce the required memory and accelerate the training process. During the training, we format the instruction data using the prompt reported in Listing 3 by relying on the chat template specific to the LLaMA models.

Listing 3: Prompt used for the fine-tuning.

```
System: <Instruction>

User: <Input>

Assistant: <Output>
```

During training, we use the following parameters: $rank = 32$, $alpha = 64$, $learning\_rate = 2e-4$ and $batch\_size = 32$. We train all models for five epochs on the whole training dataset. The training was performed using a single GPU NVIDIA RTX A6000 with 48GB of memory.

5. Evaluation

As mentioned in Section 1, our study has two objectives. First, we want to test the models ability to disambiguate Latin senses in a zero-shot setting. In this way, we aim to first establish how well the model inherent multilingual knowledge performs in accurate sense prediction. Next, we perform task-specific fine-tuning, which enables us to adapt both standard and instruction versions of LLMs. The objective is to quantify the gain obtained through this additional training step.

It is worth noticing that the dataset of choice was initially devised for the Unsupervised LSCD task [13], not for WSD; therefore, comparing the results of the shared task with the results of this work is not feasible.

GPT-4o-mini and LLaMA-3.3-70B-instruct-turbo act as a zero-shot baseline for this experiment, to assess the capabilities of models not specially devised or fine-tuned for the Latin WSD task.

It is crucial to note that the dataset is highly imbalanced, as many instances are annotated with 1, since each word occurrence is generally assigned a single meaning; consequently, all other meanings receive the lowest score. Notice that all the metrics are computed with the dataset imbalance in mind. Balanced Accuracy[12] is defined as the average recall obtained inch class. Weighted Precision[13], Recall[14] and F_1[15] calculate metrics for each label, and find their average weighted by support. Finally, Macro F_1 and Micro F_1 scores are variants of F_1. The former is the only metric that does not take into account label imbalance, but computes metrics for each label and finds their unweighted mean; the latter calculates metrics globally by counting the total true positives, false negatives and false positives. Details about the DuREL annotation statistics are reported in Table 2.

We release the following resources, available on GitHub[16]: i) the source code; ii) instruction fine-tuning and testing data; iii) links to the fine-tuned models on HuggingFace and the outputs of all evaluated models.

[10] https://www.together.ai/
[11] https://unsloth.ai/
[12] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html
[13] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html
[14] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html
[15] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html
[16] https://github.com/swapUniba/latin-wsd

Table 5
Comparison of model performance on regression task across various evaluation metrics (MSE, RMSE, Precision, Recall, Accuracy, F_1, Macro F_1, Micro F_1) for GPT-4o-mini and different LLaMA variants.

	MSE	RMSE	Precision	Recall	Accuracy	F_1	Macro F_1	Micro F_1
GPT-4o-mini	1.0743	2.4595	.6371	.4190	.3095	.4170	.2504	.4190
LLaMA-3.3-70B-instruct-turbo	1.4063	3.1550	.6056	.2743	.2372	.2543	.1742	.2743
LLaMA-3.1-8B-instruct	1.6491	3.7405	.4748	.1717	.1993	.1037	.1037	.1717
LLaMA-3.1-8B-instruct-ft	**0.7699**	**1.8093**	**.6854**	**.7071**	**.4354**	**.6940**	**.4456**	**.7071**

5.1. Regression task

Table 5 illustrates the results of the WSD task. Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) show that the fine-tuned model is better at predicting the annotation score. To give a complete overview of the results, we also provide classification metrics. Although GPT-4o-mini shows a higher precision, LLaMA-3.1-8B-instruct-ft outperforms every other model. It is interesting to note the high difference in performance between LLaMA-3.3-70B-instruct-turbo and LLaMA-3.1-8B-instruct-ft. These results prove that the fine-tuning of a medium-sized LLM using a single GPU can overcome a model of the same family with about nine times the number of parameters.

To better understand the behaviour of each model, we report the confusion matrix of each model in B. The matrices of GPT-4o-mini (Figure 1) and LLaMA-3.3-70B (Figure 2) show that the models often confuse the label 1 with other labels. It is interesting to note that GPT-4o-mini confuses the label 1 with the label 4 508 times. This behaviour is more evident in LLaMA-3.1-8B-instruct (Figure 3) where 913 instances labelled as 1 are confused with label 3 and 579 with labels 4.

The fine-tuned model LLaMA-3.1-8B-instruct-ft (Figure 4) is the best at recognising label 1. This behaviour is evident since the model tends to overfit on the more frequent class.

5.2. Binary Classification task

Results of the WSD task framed as a binary classification task are in Table 6, as well as the confusion matrix of each model in Appendix B. Our proposed fine-tuned model LLaMA-3.1-8B-instruct-ft shows a strong performance boost with respect to LLaMA-3.1-8B-instruct and LLaMA-3.3-70B-instruct-turbo. On the other hand, GPT-4o-mini performance is in line with LLaMA-3.1-8B-instruct-ft, and even surpasses it in Precision and Accuracy. In general, our LLaMA-3.1-8B-instruct-ft outperforms the baseline models. Figure 8 shows that LLaMA-3.1-8B-instruct-ft performs the best on class no, while GPT-4o-mini predicts class yes better.

6. Conclusions and Future Works

This study explores the ability of Large Language Models (LLMs) to address Word Sense Disambiguation (WSD) in Latin, a historically rich yet computationally low-resourced language. The first contribution of our work is the release of a dataset for evaluating the WSD abilities of LLMs in Latin. This dataset is created by leveraging an existing manually annotated dataset. Then, using the new dataset and through both zero-shot and fine-tuned evaluations, we observed that while general-purpose LLMs exhibit a promising baseline ability to handle Latin WSD, significant improvements are achieved through task-specific fine-tuning. The fine-tuned LLaMA-3.1-8B-instruct model outperformed larger and more resource-intensive models in accuracy and F1 scores, underscoring the impact of targeted instruction tuning, even on medium-sized architectures. Nevertheless, challenges remain. The dataset's inherent class imbalance, with a predominance of "unrelated" sense labels, likely influenced the models' predictions and underscores the need for more balanced and semantically diverse training data.

Future work will focus on three main directions: i) Expanding the annotated dataset to include more lemmata and a broader variety of senses; ii) Evaluating model performance on additional semantic tasks, such as definition generation and contextual paraphrasing in Latin; iii) Exploring multilingual and cross-lingual transfer learning strategies, leveraging annotations from related Romance languages to further boost Latin model capabilities.

Acknowledgments

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU.

References

[1] M. C. Passarotti, F. M. Cecchini, G. Franzini, E. Litta, F. Mambrini, P. Ruffolo, The lila knowledge base of

Table 6
Comparison of model performance on binary classification task across various evaluation metrics (Precision, Recall, Accuracy, F_1, Macro F_1, Micro F_1) for GPT-4o-mini and different LLaMA variants.

	Precision	Recall	Accuracy	F_1	Macro F_1	Micro F_1
GPT-4o-mini	**.7974**	.7634	**.7817**	.7682	.7573	.7634
LLaMA-3.3-70B-instruct-turbo	.7030	.6301	.6654	.6355	.6284	.6301
LLaMA-3.1-8B-instruct	.5947	.5271	.5547	.5336	.5253	.5271
LLaMA-3.1-8B-instruct-ft	.7901	**.7933**	.7637	**.7906**	**.7694**	**.7933**

linguistic resources and nlp tools for latin., in: LDK (Posters), 2019, pp. 6–11.

[2] M. Passarotti, F. Mambrini, G. Franzini, F. M. Cecchini, E. Litta, G. Moretti, P. Ruffolo, R. Sprugnoli, Interlinking through lemmas. the lexical collection of the lila knowledge base of linguistic resources for latin, Studi e Saggi Linguistici 58 (2020) 177–212.

[3] M. Passarotti, E. Litta, F. M. Cecchini, M. Pellegrini, G. Moretti, P. Ruffolo, G. Pedonese, The lila knowledge base of interoperable linguistic resources for latin. architecture and current state (2022).

[4] B. McGillivray, P. Cassotti, P. Basile, D. Di Pierro, S. Ferilli, Using graph databases for historical language data: Challenges and opportunities (2023).

[5] B. McGillivray, A. Kilgarriff, Tools for historical corpus research, and a corpus of latin, New methods in historical corpus linguistics 1 (2013) 247–257.

[6] B. McGillivray, D. Kondakova, A. Burman, F. Dell'Oro, H. Bermúdez Sabel, P. Marongiu, M. Márquez Cruz, A new corpus annotation framework for latin diachronic lexical semantics, Journal of Latin Linguistics 21 (2022) 47–105.

[7] S. Minozzi, Latin wordnet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'information retrieval, Strumenti digitali e collaborativi per le Scienze dell'Antichita (2017) 123–134.

[8] K. P. Johnson, P. J. Burns, J. Stewart, T. Cook, C. Besnier, W. J. B. Mattingly, 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, Association for Computational Linguistics, Online, 2021, pp. 20–29. URL: https://aclanthology.org/2021.acl-demo.3. doi:10.18653/v1/2021.acl-demo.3.

[9] M. Straka, J. Hajic, J. Straková, Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 4290–4297.

[10] M. Straka, J. Straková, Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe, in: Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies, 2017, pp. 88–99.

[11] W. Stroh, Latein ist tot, es lebe Latein!: kleine Geschichte einer grossen Sprache, List Taschenbuch, 2008.

[12] J. Leonhardt, Latin: Story of a world language, Harvard University Press, 2013.

[13] D. Schlechtweg, B. McGillivray, S. Hengchen, H. Dubossarsky, N. Tahmasebi, Semeval-2020 task 1: Unsupervised lexical semantic change detection, 2020. arXiv:2007.11464.

[14] R. Sprugnoli, M. Passarotti, F. M. Cecchini, M. Pellegrini, Overview of the EvaLatin 2020 evaluation campaign, in: R. Sprugnoli, M. Passarotti (Eds.), Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 105–110. URL: https://aclanthology.org/2020.lt4hala-1.16/.

[15] R. Sprugnoli, M. Passarotti, F. M. Cecchini, M. Fantoli, G. Moretti, Overview of the EvaLatin 2022 evaluation campaign, in: R. Sprugnoli, M. Passarotti (Eds.), Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages, European Language Resources Association, Marseille, France, 2022, pp. 183–188. URL: https://aclanthology.org/2022.lt4hala-1.29/.

[16] R. Sprugnoli, F. Iurescia, M. Passarotti, 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, 2024, pp. 190–197.

[17] D. Bamman, P. J. Burns, Latin bert: A contextual language model for classical philology, arXiv preprint arXiv:2009.10053 (2020).

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[19] W. Weaver, Translation, in: Proceedings of the

conference on mechanical translation, 1952.
[20] R. Navigli, Word sense disambiguation: A survey, ACM computing surveys (CSUR) 41 (2009) 1–69.
[21] W. A. Gale, K. W. Church, D. Yarowsky, A method for disambiguating word senses in a large corpus, Computers and the Humanities 26 (1992) 415–439.
[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[23] C. T. Lewis, C. Short, A latin dictionary. clarendon, 1879.
[24] P. Lendvai, C. Wick, Finetuning latin bert for word sense disambiguation on the thesaurus linguae latinae, in: Proceedings of the Workshop on Cognitive Aspects of the Lexicon, 2022, pp. 37–41.
[25] I. Ghinassi, S. Tedeschi, P. Marongiu, R. Navigli, B. McGillivray, Language pivoting from parallel corpora for word sense disambiguation of historical languages: a case study on latin, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 10073–10084.
[26] H. Wu, H. Wang, Pivot language approach for phrase-based statistical machine translation, in: A. Zaenen, A. van den Bosch (Eds.), Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 856–863. URL: https://aclanthology.org/P07-1108/.
[27] C. Fellbaum, WordNet: An electronic lexical database, MIT press, 1998.
[28] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, A. Mian, A comprehensive overview of large language models, ACM Trans. Intell. Syst. Technol. (2025). URL: https://doi.org/10.1145/3744746. doi:10.1145/3744746, just Accepted.
[29] L. Qin, Q. Chen, X. Feng, Y. Wu, Y. Zhang, Y. Li, M. Li, W. Che, P. S. Yu, Large language models meet nlp: A survey, arXiv preprint arXiv:2405.12819 (2024).
[30] D. Loureiro, K. Rezaee, M. T. Pilehvar, J. Camacho-Collados, Analysis and evaluation of language models for word sense disambiguation, Computational Linguistics 47 (2021) 387–443.
[31] M. Bevilacqua, T. Pasini, A. Raganato, R. Navigli, Recent trends in word sense disambiguation: A survey, in: International joint conference on artificial intelligence, International Joint Conference on Artificial Intelligence, Inc, 2021, pp. 4330–4338.
[32] F. Cabiddu, M. Nikolaus, A. Fourtassi, Comparing children and large language models in word sense disambiguation: Insights and challenges, in: Proceedings of the Annual Meeting of the Cognitive Science Society, volume 45, 2023.
[33] R. Kibria, S. Dipta, M. Adnan, On functional competence of llms for linguistic disambiguation, in: Proceedings of the 28th Conference on Computational Natural Language Learning, 2024, pp. 143–160.
[34] J. H. Yae, N. C. Skelly, N. C. Ranly, P. M. LaCasse, Leveraging large language models for word sense disambiguation, Neural Computing and Applications 37 (2025) 4093–4110.
[35] P. Basile, E. Musacchio, L. Siciliani, Ita-sense-evaluate llms' ability for italian word sense disambiguation: A calamita challenge, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, 2024.
[36] P. Basile, L. Siciliani, E. Musacchio, G. Semeraro, Exploring the word sense disambiguation capabilities of large language models, arXiv preprint arXiv:2503.08662 (2025).
[37] T. Pasini, A. Raganato, R. Navigli, Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 13648–13656.
[38] J. Clackson, A companion to the Latin language, John Wiley & Sons, 2011.
[39] J. Clackson, G. Horrocks, The Blackwell history of the Latin language, John Wiley & Sons, 2011.
[40] P. Glare, Oxford Latin Dictionary, number Num. 1-4 in Oxford Latin Dictionary, Clarendon Press, 1982. URL: https://books.google.it/books?id=H7HhzAEACAAJ.
[41] T. Lewis Charlton, An elementary latin dictionary, New York, Cincinnati, and Chicago. American Book Company (1890).
[42] C. d. F. Du Cange, Glossarium mediae et infimae latinitatis: AZ, volume 7, L. Favre, 1886.
[43] D. Schlechtweg, S. Schulte im Walde, S. Eckmann, Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 169–174. URL: https://aclanthology.org/N18-2027/. doi:10.18653/v1/N18-2027.
[44] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models., ICLR 1 (2022) 3.

A. Translation

Cicero's *Tuscolanae Disputationes*

la: [...] Dico enim constanter grauiter sapienter fortiter. Haec etiam in eculeum coiciuntur, quo uita non adspirat beata. - Quid igitur? solane beata uita, quaeso, relinquitur extra ostium limenque carceris, cum constantia grauitas fortitudo sapientia reliquaeque uirtutes rapiantur ad tortorem nullumque recusent nec supplicium nec dolorem? [...]

en: For I say constantly, gravely, wisely, and strongly. These things are also cast into the rack, to which life does not aspire for happiness. - What then? Is a blessed life alone, I pray you, left outside the door and threshold of the prison, when constancy, gravity fortitude, wisdom and the other virtues are snatched away to the torturer and refuse neither punishment nor pain?

Robertus Grossetest's *De libero arbitrio*

la: [...] Ex quo fit, ut de nihilo creauerit omnia." Eadem itaque ratione solus facit ominia, nulla adiutus natura. Horum autem obiectorum solutio haberi potest ut uidetur ex uerbis beati Bernardi sic dicentis: "Ipsa gratia Liberum arbitrium excitat, cum seminat cogitatum. Sanat, cum mutat affectum; roborat, ut perducat ad actum; seruat, ne sentiat defectum." [...]

en: From which it comes about that He created all things out of nothing." Therefore, by the same reasoning, He alone creates all things, without any help from nature. But the solution to these objections can be found, as can be seen from the words of Blessed Bernard, who says thus: "Grace itself awakens Free will when it sows thought. It heals when it changes affection; it strengthens, so that it may lead to action; it preserves, so that it may not feel a deficiency."

B. Confusion Matrices

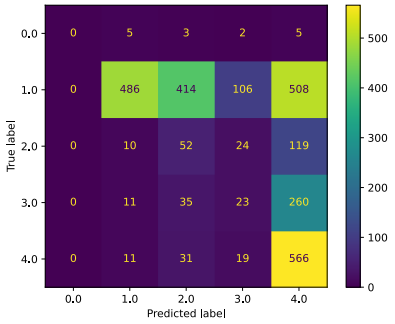

Figure 1: GPT-4o-mini confusion matrix (regression task).

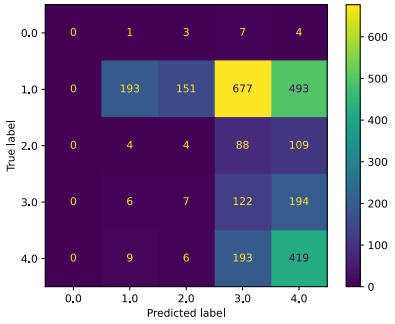

Figure 2: LLaMA-3.3-70B-instruct-turbo confusion matrix (regression task).

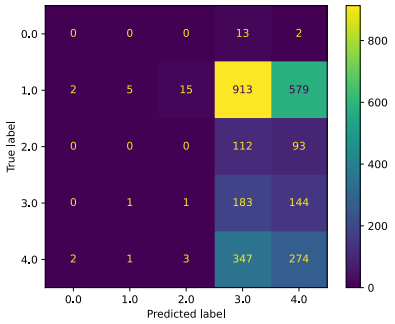

Figure 3: LLaMA-3.1-8B-instruct confusion matrix (regression task).

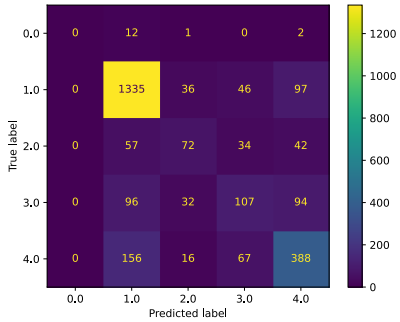

Figure 4: LLaMA-3.1-8B-INSTRUCT-FT confusion matrix (regression task).

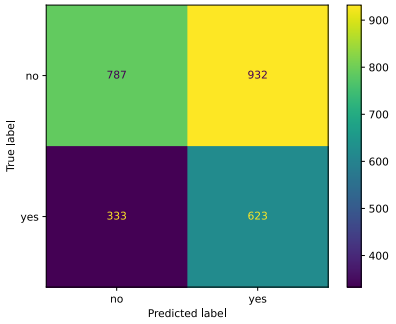

Figure 7: LLaMA-3.1-8B-INSTRUCT confusion matrix (binary task).

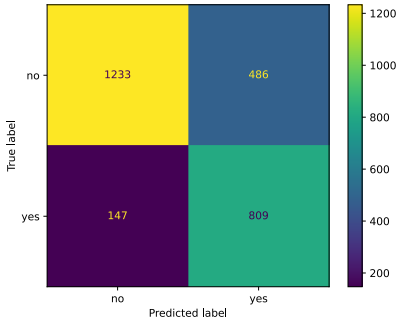

Figure 5: GPT-4O-MINI confusion matrix (binary task).

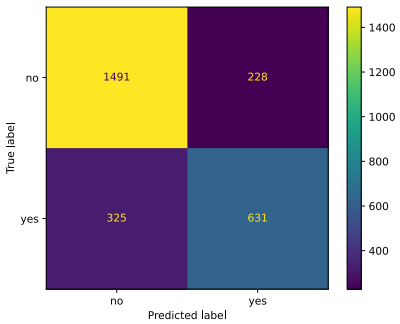

Figure 8: LLaMA-3.1-8B-INSTRUCT-FT confusion matrix (binary task).

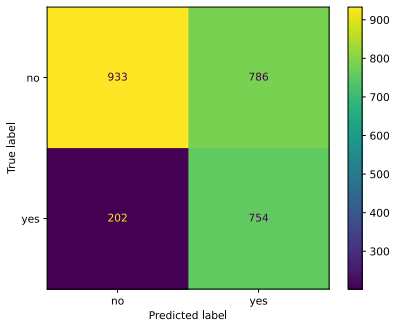

Figure 6: LLaMA-3.3-70B-INSTRUCT-TURBO confusion matrix (binary task).

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (OpenAI) and Grammarly in order to: Paraphrase and reword, Improve writing style, and Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.