The Latin Language Evolved Over Time, Masked Models Disregard That

Miriam Cuscito[1], Alfio Ferrara[2], and Martin Ruskov[3]

[1] Department of Languages, Literatures, Cultures and Mediations, University of Milan, Milan, Italy
[2] Department of Informatics, University of Milan, Milan, Italy
[3] Department of Languages and Modern Cultures, University of Genoa, Genoa, Italy

Abstract

Training of Latin language models is rarely done with consideration of important historical watersheds. Here we demonstrate how this leads to a poor performance when specific socio-temporal contextualisation is sought, something common to humanities research. We perform an evaluation that compares the historical adequacy of Latin language models, i.e. their ability to generate tokens, representative for a historical period. We adopt a previously established method and refine it to overcome limitations due to Latin being an under-resourced language and one with intense tradition of intertextuality. To do this we extract word lists and concordances from the LatinISE corpus and use them to compare seven masked language models trained for Latin. We further perform statistical analysis of the results in order to identify the best and worst performing models in each of the historical contexts of interest. We show that BERT medieval multilingual best captures the Classical linguistic context. Four models are indistinguishably good in our evaluation of the the Neo-Latin linguistic context. These findings have broad implications for wider historical language research and beyond. Among these, we emphasise the need to train historical language models with due attention on consistent historical periods and we discuss the possible usefulness of noisy predictions. Historical research of language models provides a neat demonstration of how model biases could impact their performance in specific domains.

Keywords: historical adequacy, algorithmic bias, masked language models, Latin language, model evaluation

1 Introduction

The relevance of computational linguistics and language models to the digital humanities is continuously growing. BERT-like masked language models (MLMs) are a particularly important class of interest. On one hand, these are the large language models where most relevant training on historical texts has been done [5; 11; 15; 19], on the other these models are widely used as a starting point for a number of other important tasks in computational linguistics, such as text reuse detection [22] or semantic shift detection [6]. Historical analysis using computational linguistics is challenged by the fact that these MLMs are trained with disregard to historical changes in language and culture. Such disregard is commonly due to the need to collect a critical mass of text in a corpus for training data [9; 11]. This is particularly evident with the Latin language, which has inevitably evolved over its long timespan, but research in language models hugely ignores this. The work of Riemenschneider and Frank [23] is a notable exception, also attentive to the historical importance and intertextuality that also underlines the relevance of Latin to research in other languages from the classical and medieval periods.

Miriam Cuscito, Alfio Ferrara, and Martin Ruskov. "The Latin Language Evolved Over Time, Masked Models Disregard That." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 1275–1286. https://doi.org/10.63744/sLAHYnQdA8fu.

In this contribution we study the *historical adequacy* of different MLMs for Latin. We do this by applying a technique that was previously used to measure such *adequacy* of models to Early Modern English, by capturing the *historical bias* of models and interpreting it as a representation of the socio-temporal context of the model training corpus [9]. Key here is an important assumption in computational humanities research, that often remains implicit: this socio-temporal context should be close to the one being studied using these models. Due to the nature of MLMs and the abundance of lexical resources in computational linguistics, this approach is inherently lexical, but evidence shows that it is also able to capture particularities in orthography, grammar, semantics and culture [8; 9].

2 Background

To contextualise our work, this section provides an overview of available and relevant corpora and models within our knowledge.

2.1 Corpora

We present a number of corpora that are relevant to the experiments conducted here due to their impact in model training and broader computational linguistics. On a general note, a huge challenge in the context of historical research is the scarcity of corpora, due to a number of factors, including loss of manuscripts and limited digitisation. This, together with the inherent high intertextuality, leads to a noteworthy probability that the same text emerges across multiple corpora.

Name	Timespan	Words	Share
Nova	XV-XXI cent.	1.7M	15.1%
Mediaevalis	V II-XIV cent.	2.4M	21.4%
Romana Postclassica	I-VI cent.	4.6M	41.6%
Romana Classica	I cent. BC	1.8M	16.1%
Romana Antiqua	VII-II cent. BC	0.4M	3.2%

Table 1: Subcorpora in the evaluation corpus *LatinISE*.

As the name suggests, **LatinISE** is a Latin text corpus collected from three historical sources: LacusCurtius, Intratext and Musisque Deoque [16]. This corpus was developed to be used with SketchEngine[1] and thus was collected with semantic annotations in mind. As further explained in Section 4, this is our corpus of choice for the current evaluation and a breakdown of its historical subcorpora is featured in Table 1.

Corpus corporum is currently the largest curated corpus of Latin [24]. It is very likely to contain any other historical curated and publicly available corpora, including the treebanks discussed at the end of this section. This corpus is being actively expanded, and this is our explanation why models trained on it earlier, report smaller training set sizes than ones trained later.

The Internet Archive is a digitisation effort that also includes OCR and with this it has been used as a training corpus. It also includes an important chunk of Latin texts, but this digitsation is an unsupervised process and as a consequence the produced texts are often of poor quality [3].

CommonCrawl is the largest web scraping effort online and thus an important corpus for languages, including Latin which counts as one of its top 100 languages. A cleaned corpus with this top languages subset has been derived and named **CC100**. However, remotely similarly to the case of the Internet Archive, these corpora have been only automatically curated which has been

[1] `https://sketchengine.eu`

a source of concerns [1]. By construction it is a contemporary corpus, containing sources such as Wikipedia in Latin.

There is a number of other established smaller corpora, such as **The Latin Library**, **Wikipedia in Latin** and the syntactically annotated treebanks **Perseus Digital Library** and **PROIEL** project. These are typically included in other aggregation efforts and is part of the reason why the previously mentioned corpora have consequential overlaps in content.

2.2 Models

Model	Corpus	Tokens	Timespan
LaBerta [23]	Corpus Corporum	168M	700BC-2004[2]
RoBERTa Latin cased3	Corpus Corporum	232M	700BC-2004[2]
Cicero Similis [7]	4 small corpora	1.2M	N/A[3]
BERT adapter pfeiffer [21]	Wikipedia in Latin	N/A	2002-2025[4]
RoBERTa medieval	N/A	N/A	N/A
BERT medieval multilingual	CC100,...[5]	650M	500BC - 1600[6]
LatinBERT [2]	Internet Archive,...[7]	643M	200BC-1922[8]

Table 2: Models and their corresponding training corpora.

We present the models we included in our evaluation. Most of these were selected due to the documented research that accompanied to their training, but some less documented were also included due to their performance.

LaBerta is one of the models produced in a systematic effort to train language models on classical languages [23]. It was pre-trained from the RoBERTa [14] architecture on the version of Corpus Corporum that was available at its time, and we believe this explains the different reported number of tokens for this corpus in the context of different models in Table 2. Here we select the monolingual model *LaBerta*, even if the multilingual *PhilBerta* could also be very relevant to this evaluation.

Cicero Similis is another model that is presumedly BERT-based and trained on 4 relatively small corpora: Phi5, Tesserae, Thomas Aquinas, and Patrologes Latina [7].

BERT adapter pfeiffer is the Latin implementation of a generic adapter approach that uses Wikipedia in different languages as a training corpus [21]. Latin is not in the focus of the analysis included in the original publication, nor is mask-filling as a task.

LatinBERT was trained with the BERT architecture [10] on a bespoke corpus that combines the digitisation of materials from the Internet Archive [2; 3] and several curated corpora: Perseus Digital Library, the Latin Library, the Patrologia Latina, Corpus Thomisticum and Wikipedia in Latin. The Internet Archive represents more than 85% of the used corpus size, so the collection of this particular data is of central interest. It relies on the OCR digitisation performed by the Internet Archive, which is of varying quality. To mitigate this problem, the authors retain only the materials

[2] taken from current version of corpus, which is probably broader than the version at time of retrieval
[3] from the model card: "model training data excludes modern and 19th century texts"
[4] based on information from `https://en.wikipedia.org/wiki/List_of_Wikipedias#Wikipedia_editions`
[5] includes CC100, Corpus Corporum and other corpora that cumulatively represent less than 10%
[6] The model card contains controversial information about the end date of CC100, its biggest subcorpus - XVIII cent. is indicated in one instance.
[7] also Patrologia Latina, the Latin Library, Wikipedia in Latin, Corpus Thomisticum, Perseus
[8] Since the Internet Archive is the source of >85% of the data, we consider it to determine the overall period. This is obviously an approximation, e.g. due to the presence of Wikipedia in Latin

that have at least 40% valid tokens in Latin (which we note is a very low threshold) and upsample the other included corpora to reduce the share of the corpus from the Internet Archive to about 50% [2; 3].

As shown in Table 2, we found several other models, fine-tuned for Latin. From these we included the ones that indicated interesting results during preliminary evaluations. The models we additionally selected are **RoBERTa Latin cased v3**, **RoBERTa medieval** and **BERT medieval multilingual**. Several other models that we have identified and not included in the evaluation are listed in Appendix A.

We would also like to explicate the fact that when foundation models such as BERT or RoBERTa are fine-tuned, even if this fine-tuning is attentive to the socio-temporal context, contamination from the pre-training is probable.

3 Method

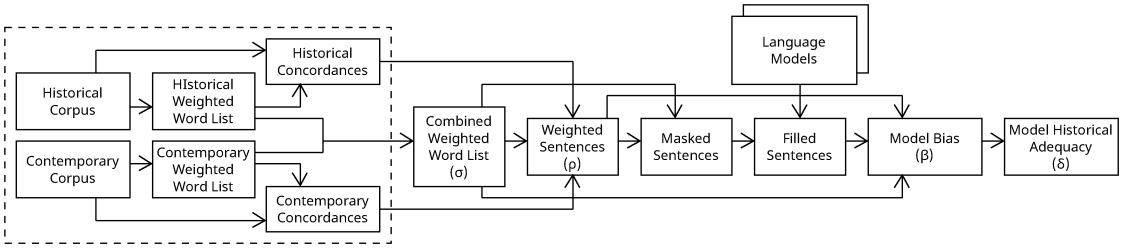

Figure 1: The process introduced by Cuscito et al. [9] and employed here. Note that "historical" and "contemporary" refer to *Classica* and *Nova* respectively in the context here. The dotted box indicates the perimeter of steps performed with SketchEngine.

Following the evaluation process of Cuscito et al. [9] illustrated in Fig. 1, we perform a masked-word prediction task to assess the historical adequacy of models trained on Latin to the variant of the language used in the Classical period it on one hand, and on the other – Neo-Latin (see respectively *Romana Classica* and *Nova* in Table 1). We deviate from the procedure originally defined in Cuscito et al. [9] only for the calculation of bias β, as shown in equation (4). The entire process is explained in detail below.

For the weighted word lists in the start of the process in Fig. 1, we define word frequency functions f_{nova} and $f_{classica}$, which given a word return its normalised relative frequency in the corresponding corpus. Thus, we can employ the equations (1) and (2) defined below, which are just a reformulation of equation (1) in Cuscito et al. [9]. This reformulation splits the calculation of *historical specificity* (σ) in two steps to make it more explicit. More specifically, we define an intermediate *distance function* (d) in equation (1), which is later used in equation (2). The constant four in the fourth root functions in equation (1) is adopted from Cuscito et al. [9]. It controls how words present in only one corpus mix with ones that are present in both. Values smaller than 4 leave a gap between σ scores for words that are present in both corpora, and those that are present in one only (i.e. one of the f functions is null). On the other hand, values greater than 4 lead to words that are exclusive to one of the corpora not getting a greater weight.

$$d(f_a, f_b, w) = \sqrt[4]{\sqrt[4]{f_a(w)} - \sqrt[4]{f_b(w)}} \quad \text{given that } f_a(w) \geq f_b(w) \tag{1}$$

$$\sigma(w) = \begin{cases} \frac{d(f_{nova}, f_{classica}, w)}{\max_w d(f_{nova}, f_{classica}, w)} & \text{if } f_{nova}(w) \geq f_{classica}(w) \\ \frac{-d(f_{classica}, f_{nova}, w)}{max_w d(f_{classica}, f_{nova}, w)} & \text{otherwise.} \end{cases} \tag{2}$$

With the *historical specificity* function, we can measure *sentence valence* (ρ) as the trimmed sum of the *specificity* of words contained in it. Notice that for this *valence* to arrive at an extreme for a sentence, at least two words need to have non-zero *specificity*. In this way, even when one of these words is masked, the others would still carry some non-zeros valence, rendering the masked sentence indicative of its socio-temporal context. Only such sentences are selected for evaluation, as reported in the next section.

$$\rho(s) = min(1, max(-1, \sum_{w_i \in s \cap \mathcal{C}} \sigma(w_i))) \tag{3}$$

Weighting the *historic specificity* of word tokens predicted by models (w_m), with the corresponding prediction probabilities ($\mathbf{p}_m$), allows us to subsequently define model *historical bias* (β). As a way to reduce the bias introduced by the very high probability of texts being present both in model training corpora and in our evaluation corpus (i.e. data leakage), during the computation of *historical bias* we exclude the originally encountered word from the assessment by considering *specificity* $\sigma(w) = 0$ for it. As a consequence, this also reduces the overall *historical bias* scores of models.

$$\beta(m, s) = \sum_{w_m \text{ for } s} \begin{cases} 0 & \text{if } w_m \text{ is the original token in the sentence} \\ \sigma(w_m)\mathbf{p}_m & \text{otherwise.} \end{cases} \tag{4}$$

From the above, also *historical adequacy* (δ) is derived. It provides a normalised measure of *historical bias* regardless of direction:

$$\delta(m, s) = 1 - \frac{\mid \rho(s) - \beta(m, s) \mid}{2} \tag{5}$$

To elucidate the results of this method, at the end we perform statistical significance testing on *historical adequacy* to see if any models could be claimed better than others in any of the two historical contexts represented by our word lists. First, we compare all models using repeated measures ANOVA. In the case of this indicating possible significance, we proceed to perform a combination of one-tailed paired t-tests to compare models against each other and calculate Cohen's d to measure the corresponding effect sizes. We do this separately for results on *Classica* and *Nova* sentences, as we expect that different models behave differently in the two contexts.

4 Evaluation

	senatus	uterque	fortuna	lex	populus	et	terrae	rex	vitae	christo	dei
$f_{classica}$	0.027	0.005	0.016	0.005	0.009	1.000	0.007	0.012	0.009	0.000	0.000
f_{nova}	0.000	0.000	0.007	0.004	0.008	1.000	0.008	0.018	0.022	0.011	0.041
σ	0.981	0.877	0.623	0.439	0.386	0.000	-0.361	-0.505	-0.603	-0.873	-0.947

Table 3: An illustrative sample with some values of relative word frequencies and *specificity* (σ) between the two corpora.

We use the LatinISE corpus as a diachronic reference, because we consider it both representative for the evolution of Latin, and – in combination with SketchEngine – it affords easily accessible word frequency lists and their corresponding concordances. To ensure representativeness, we consider the 1000 most frequent words. We discard the *Romana Antiqua* subcorpus from our considerations due to its small size, as indicated in Table 1. As a way to emphasise the linguistic evolution over time from the remaining subcorpora we choose the two which are most distant in time, i.e. *Romana Classica* and *Nova*. Taking the 1000 most frequent words from the two resulted

in a combined list of 1329 words scored with *historical specificity* (σ). A sample of this scoring is shown in Table 3.

T1/Tinos(0)/m/n' undefined For our analysis, we collect 5000 sentences (concordances from SketchEngine) from each of the *Romana Classica* and the *Nova* subcorpora of LatinISE with *sentence valence* of $\rho = 1$ and $\rho = -1$ respectively (which we call *threshold valence*), as samples from these corpora have highest probability to bear distinct socio-temporal connotations. Then we perform the comparison detailed in the previous section. A visual overview of the results could be seen in Fig 2.

For illustrative purposes, in Table **??** we explore in detail the results emerging from two sentences in our evaluation. Our selected example from the *Classica* subcorpus is the phrase "Nonne in terris multa, ut oppidum in Graecia Hippion Argos?" from *De re rustica* by Varro which translates to "Are there not many in the world, such as the Greek town of Argos Hippion?". Our example from the *Nova* subcorpus, "Sed statim Solymanus ipso patre acrior Pannoniam invasit." from *Icon Animorum* by John Barclay, could be translated as "But immediately Suleyman, more aggressive than his father himself, invaded Pannonia." Notably for the second, models propose the originally present word "in", but it doesn't contribute to the corresponding scores. This occurs also in other test sentences and, as already mentioned, draws the absolute value of overall scores of the models down.

The tokenisers of some models (as seen in Table **??** for *BERT medieval multilingual* on the *Nova* example) produce tokens that are not words and which do not get attributed *valence* even if they might be meaningful. Similarly, even though this did not make it into the illustrative example shown here, the *Latin BERT* model commonly predicts word tokens with orthographic defects, probably due to OCR, which also do not get *valence* assigned. As a consequence *Latin BERT* obtains a very low score and was excluded from the quantitative comparison.

For statistical significance testing, we fist perform repeated measures ANOVA on the results of the different models, with *historical adequacy* (δ) as a dependent variable and the evaluation corpus groupings as an independent variable (also referred to as within-subject factor) and this gave us a reference probability of $p = 0.0001$ and F-statistic $F = 121.2929$. This probability is below the 1‰ confidence threshold and thus very strong indication of statistical evidence that the models do not perform equally.

An overview of the pairwise t-tests on *historical adequacy* (δ) scores could be seen in Fig **??**. In it only pairwise significant differences are shown, with reported direction of significance (color), t-test probability and effect size (Cohen's d). The heatmap for *Classica* sentences shows that *BERT medieval multilingual* consistently outperforms other models, followed by *RoBERTa medieval* and *BERT adapter pfeiffer*. Then with less consistent significance, yet in this order, follow *Cicero Similis* and *LaBerta*, even if the former does not perform significantly better than the model that ranks last, i.e. *Roberta Latin cased3*. The result for *Nova* sentences shows that *BERT adapter pfeiffer* is worse than the other models, followed by *BERT medieval multilingual*. This pair is the only one that has statistical significance that is above the 1‰ confidence threshold, yet it is still below the 5% threshold which is commonly accepted in the social sciences and humanities. As for the other models, they are statistically indistinguishable in the *Nova* context. However, as seen in Fig **??**, the measured effect sizes turn out to be rather small. This is a particularly important sign, because of the large number of sampled sentences (n=5000 for each of the periods).

5 Discussion

In the adopted approach we consider the most frequent word tokens from the evaluation corpora, regardless of their content or grammatical function. As illustrated by the example sentences provided in the previous section, this has the effect of capturing not barely lexical differences, but more complex phenomena like geographic relationships or historical circumstances. These two are just

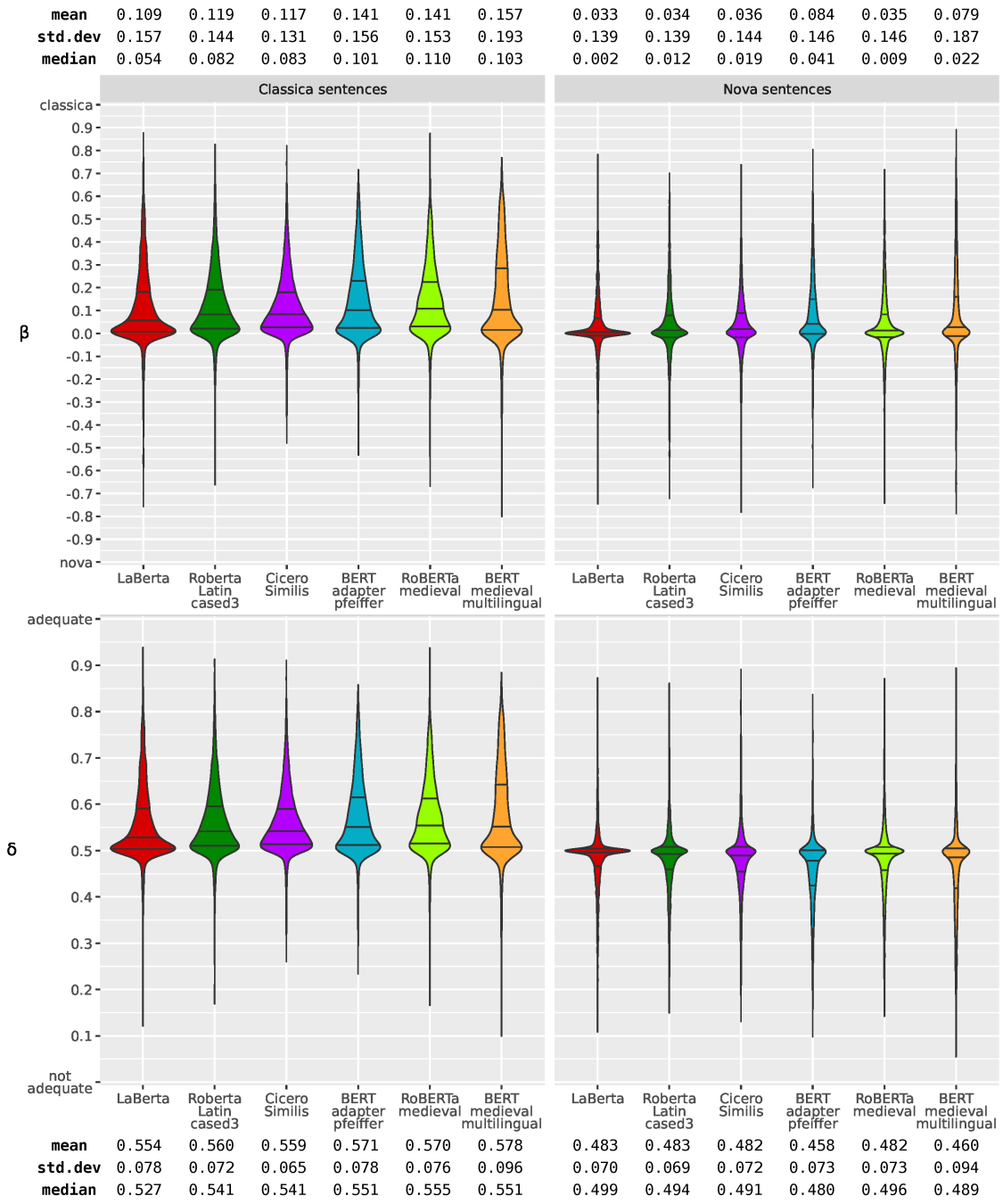

Figure 2: Comparison overview of results for bias β (above) and historical adequacy δ (below). It could be seen that for *Classica* sentences δ is just rescaling of δ, whereas for *Nova* the values are also mirrored vertically. Thus δ brings together the two scales of β.

examples of the context contributing to the evaluation of the choice of words. While admittedly just considering the most frequently used vocabulary does not capture all of this context, it does capture orthographic and morphological variation, grammatical evolution, and socio-cultural phenomena. However, due to Latin being a highly inflected and syntactically complex language, the integrity of this generic approach might be questioned and alternative ways of calculating specificity could be considered.

To justify the exclusion of the *Romana Antiqua* subcorpus due to the sentence identification

process, consider that sentences need to both contain the words to be masked, and have a *threshold valence* ($\rho = \pm 1$) as a sign that the sentences are contextually charged for the corresponding period. As an illustration of the importance of the latter, consider a contrast to a sentence we would not like to have included, i.e. one with generic references that does not provide clues about its socio-historical context. A naive illustrative example could be "Caelum stellatum aspicio". It translates to "I look at the starry sky" and the context it carries is generic and does not relate to a particular culture.

Overall, the evaluation tends to produce results for *bias* which are very close to neutral (i.e. $\beta = 0$). This is noticeably more so when compared to the results for historical models in the study of English reported by Cuscito et al. [9]. While undoubtedly our modification of excluding the original word token from the evaluation had an impact on this, we speculate about other contributing factors. Our main hypthesised reason is plainly that models were trained on corpora indiscriminately and thus have a mixed representation across periods. If the *LatinISE* corpus could be an indication, the bulk of available texts comes from the the period of the *Postclassica* corpus, which means that they are both socio-temporally close to the context of *Classica*, and have had cultural influence on *Nova*. On the other hand, a more complex view of dating could need to be considered, because the historical tradition of hand-copying manuscripts has lead to the phenomenon that in certain cases only more recent copies of classical manuscripts are preserved and exhibit medieval features introduced by the scribes [18].

Considering the models, *BERT medieval multilingual* appears to perform best for the historical context of interest (the Classical period, represented by the *Classica* corpus), which is also observable by the 75% threshold line for *bias* in Fig. 2. It is a surprise that *BERT adapter pfeiffer* performs relatively well on the *Classica* corpus, because it is supposedly the only model that is trained only on contemporarily written Latin texts (i.e. Wikipedia in Latin, as indicated in Table 2). When it comes to the evaluations in the *Nova* context, a reason why four models – namely *Cicero Similis*, *LaBerta*, *RoBERTa mesdieval* and *Roberta Latin cased3* – appear indistinguishable could be due to the high number of more recent Latin corpora being used for training.

Arguably, the results of *Latin BERT* – and *the Internet Archive* respectively – show that unlike the case of printed documents where optical character recognition (OCR) might have caused a breakthrough in terms of mass availability of training corpora, this is yet to happen with historical documents and, respectively, handwritten text recognition (HTR). Yet, despite a failure to generate exact meaningful words that could be captured by the method employed here, such models are still valuable for the approximations they are able to produce in under-resourced contexts where there is nothing better available.

6 Conclusion

The proposed evaluation approach builds on the central role of words in both MLMs and Corpus Linguistics, particularly expressed in the affordances of tools like SketchEngine. Whereas the method is computationally robust, the effect size measures in the statistical analysis limit conclusions to little beyond what is visible from Fig. 2.

An open question remains whether narrowing down the choice of word tokens considered in the evaluation could result in a more clear distinction between model performance. An example of doing this could be making a distinction between word types, even if this means simply content vs functional words. More targeted word selection could be attempted by masking and quantifying the specificity of – for example – content words only. Alternatively, a lemmatized vocabulary could be considered as a way to capture also less common word forms in the calculation of adequacy. The particular case of *Latin BERT* raises the question whether in future research the current evaluation method should be adapted towards mitigating transcription noise, or keep it as it is as a way to account for orthographic variation.

Further experiments need to compare other periods, particularly involving the content-rich preiods that correspond to the *Romana Postclassica* and *Medievalis* periods in the *LatinISE* corpus. Particularly interesting are comparisons involving the linguistic variation in Medieval Latin.

As previously suggested, the widely adopted "don't stop pre-training" principle [13] carries risks of historical bias (e.g. through anachronisms) [9; 23; 25]. One way to train models that are historically-aware is through the pre-training of dedicated models to historical periods. This is due to another principle – "what is done is done" – reminding that historical memory can develop only in the direction of time [20]. That is, models trained on medieval texts are expected to have some historical memory of previous periods, but not of subsequent periods. This is why models with a medieval cut-off date for their corpus are also of interest. Yet, *BERT medieval multilingual*, assumingly from its name, gets closer to such idea and this might be a possible explanation for its better performance on *Classica* sentences than on *Nova* ones. In our analysis, despite naming, we found no conclusive evidence that such historically-aware models exist, as summarised in Table 2.

Finally, even though in recent years attention has shifted away from MLM autoencoders towards next token autoregression models, the reemergence of the fill-in-the-middle paradigm [4; 17] reaffirms that some form of mask-filling approaches will continue to have their relevance also in future.

References

[1] Baack, Stefan. "A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl". In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 2199–2208. DOI: 10.1145/3630106.3659033.

[2] Bamman, David and Burns, Patrick J. "Latin BERT: A Contextual Language Model for Classical Philology". 2020. DOI: 10.48550/arXiv.2009.10053.

[3] Bamman, David and Smith, David. "Extracting two thousand years of latin from a million book library". en. In: *Journal on Computing and Cultural Heritage 5*, no. 1 (2012), pp. 1–13. DOI: 10.1145/2160165.2160167.

[4] Bavarian, Mohammad, Jun, Heewoo, Tezak, Nikolas, Schulman, John, McLeavey, Christine, Tworek, Jerry, and Chen, Mark. "Efficient Training of Language Models to Fill in the Middle". 2022. DOI: 10.48550/arXiv.2207.14255.

[5] Beck, Christin and Köllner, Marisa. "GHisBERT – Training BERT from scratch for lexical semantic investigations across historical German language stages". In: *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, ed. by Nina Tahmasebi, Syrielle Montariol, Haim Dubossarsky, Andrey Kutuzov, Simon Hengchen, David Alfter, Francesco Periti, and Pierluigi Cassotti. Singapore: Association for Computational Linguistics, 2023, pp. 33–45. DOI: 10.18653/v1/2023.lchange-1.4.

[6] Cassotti, Pierluigi, Siciliani, Lucia, DeGemmis, Marco, Semeraro, Giovanni, and Basile, Pierpaolo. "XL-LEXEME: WiC Pretrained Model for Cross-Lingual LEXical sEMantic changE". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 1577–1585. DOI: 10.18653/v1/2023.acl-short.135.

[7] Cook, Todd G. "What Would Cicero Write? Examining Critical Textual Decisions with a Language Model". en. In: *Ciceroniana on line 5*, no. 2 (2021), pp. 285–296. DOI: 10.13135/2532-5353/6523.

[8] Cuscito, Miriam, Ferrara, Alfio, and Ruskov, Martin. “How BERT Speaks Shakespearean English? Evaluating Historical Bias in Contextual Language Models”. en. In: *Proceedings of the 3rd Workshop on Artificial Intelligence for Cultural Heritage*. Bolzano, Italy: CEUR, 2024, pp. 14–21. URL: https://ceur-ws.org/Vol-3865/02_paper.pdf.

[9] Cuscito, Miriam, Ferrara, Alfio, and Ruskov, Martin. “Shakespeare did not know our vocabulary: Measuring the Historical Adequacy of LLMs”. en. In: *CyberHumanities*. Florence, Italy: IEEE, 2025.

[10] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[11] Gabay, Simon, Ortiz Suarez, Pedro, Bartz, Alexandre, Chagué, Alix, Bawden, Rachel, Gambette, Philippe, and Sagot, Benoît. “From FreEM to D’AlemBERT: a Large Corpus and a Language Model for Early Modern French”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: ELRA, 2022, pp. 3367–3374. URL: https://aclanthology.org/2022.lrec-1.359.

[12] Gibert, Ona de et al. “A New Massive Multilingual Dataset for High-Performance Language Technologies”. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, ed. by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. Torino, Italia: ELRA and ICCL, 2024, pp. 1116–1128. URL: https://aclanthology.org/2024.lrec-main.100/.

[13] Gururangan, Suchin, Marasović, Ana, Swayamdipta, Swabha, Lo, Kyle, Beltagy, Iz, Downey, Doug, and Smith, Noah A. “Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks”. 2020. DOI: 10.48550/arXiv.2004.10964.

[14] Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, and Stoyanov, Veselin. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. 2019. DOI: 10.48550/arXiv.1907.11692.

[15] Manjavacas, Enrique and Fonteyn, Lauren. “MacBERTh: Development and Evaluation of a Historically Pre-trained Language Model for English (1450-1950)”. In: *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, ed. by Mika Hämäläinen, Khalid Alnajjar, Niko Partanen, and Jack Rueter. NIT Silchar, India: NLP Association of India (NLPAI), 2021, pp. 23–36. URL: https://aclanthology.org/2021.nlp4dh-1.4/.

[16] McGillivray, Barbara and Kilgarriff, Adam. “Tools for historical corpus research, and a corpus of Latin”. In: *New Methods in Historical Corpus Linguistics*, ed. by Martin Bennett, Martin Durrell, Silke Scheible, and Richard J Witt. Korpuslinguistik und Interdisziplinäre Perspektiven auf Sprache. Narr, 2013, pp. 247–256. ISBN: 9783823367604.

[17] Nguyen, Anh, Karampatziakis, Nikos, and Chen, Weizhu. “Meet in the Middle: A New Pre-training Paradigm”. In: *Advances in Neural Information Processing Systems*, ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 5079–5091. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/105fdc31cc9eb927cc5a0110f4031287-Paper-Conference.pdf.

[18] Omayio, Enock Osoro, Indu, Sreedevi, and Panda, Jeebananda. “Historical manuscript dating: traditional and current trends”. en. In: *Multimedia Tools and Applications 81*, no. 22 (2022), pp. 31573–31602. ISSN: 1573-7721. DOI: 10.1007/s11042-022-12927-8.

[19] Palmero Aprosio, Alessio, Menini, Stefano, and Tonelli, Sara. “BERToldo, the Historical BERT for Italian”. In: *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, ed. by Rachele Sprugnoli and Marco Passarotti. Marseille, France: ELRA, 2022, pp. 68–72. URL: https://aclanthology.org/2022.lt4hala-1.10.

[20] Periti, Francesco, Ferrara, Alfio, Montanelli, Stefano, and Ruskov, Martin. “What is Done is Done: an Incremental Approach to Semantic Shift Detection”. In: *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 33–43. DOI: 10.18653/v1/2022.lchange-1.4.

[21] Pfeiffer, Jonas, Vulić, Ivan, Gurevych, Iryna, and Ruder, Sebastian. “MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, 2020, pp. 7654–7673. DOI: 10.18653/v1/2020.emnlp-main.617.

[22] Reimers, Nils and Gurevych, Iryna. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. 2019. DOI: 10.48550/arXiv.1908.10084.

[23] Riemenschneider, Frederick and Frank, Anette. “Exploring Large Language Models for Classical Philology”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 15181–15199. DOI: 10.18653/v1/2023.acl-long.846.

[24] Roelli, Philipp and Ctibor, Jan. “A new version of Corpus Corporum, the latin full-text database and tool”. In: *Bulletin du Cange (Archivium latinitatis medii aevi) 80* (2024), pp. 251–266. DOI: 10.5167/UZH-265929.

[25] Underwood, Ted, Nelson, Laura K., and Wilkens, Matthew. “Can Language Models Represent the Past without Anachronism?” 2025. DOI: 10.48550/arXiv.2505.00030.

A Further Models

The following models for Latin have been identified on HuggingFace:

- http://hf.co/AdapterHub/bert-base-multilingual-cased_la_wiki_pfeiffer
- http://hf.co/bowphs/LaBerta [23]
- http://hf.co/bowphs/PhilBerta [23]
- http://hf.co/Cicciokr/XLM-Roberta-Base-Latin-Uncased
- http://hf.co/ClassCat/roberta-base-latin-v2
- http://hf.co/cook/cicero-similis
- http://hf.co/HPLT/hplt_bert_base_la [12]
- http://hf.co/LuisAVasquez/simple-latin-bert-uncased

- http://hf.co/magistermilitum/BERT_medieval_multilingual
- http://hf.co/magistermilitum/RoBERTa_medieval
- http://hf.co/pnadel/LatinBERT [2]
- http://hf.co/pstroe/roberta-base-latin-cased
- http://hf.co/pstroe/roberta-base-latin-cased2
- http://hf.co/pstroe/roberta-base-latin-cased3