Don't stop pretraining! Efficiently building specialised language models in resource-constrained settings.

Sven Najem-Meyer[1], Frédéric Kaplan[1], Matteo Romanello[2]

[1]Digital Humanities Laboratory, EPFL, Lausanne, Switzerland
[2]Institute of Archeology and Classical Studies, University of Lausanne, Lausanne, Switzerland
{sven.najem-meyer,frederic.kaplan}@epfl.ch
matteo.romanello@unil.ch

Abstract

Developing specialised language models for low-resource domains typically involves a trade-off between two specialisation strategies: adapting a general-purpose model through continued pretraining or retraining a model from scratch. While adapting preserves the model's linguistic knowledge, retraining benefits from the flexibility of an in-domain tokeniser – a potentially significant advantage when handling rare languages. This study investigates the impact of tokenisation, specialisation strategy, and pretraining data availability using classical scholarship – a multilingual, code-switching and highly domain-specific field – as a case study. Through extensive experiments, we assess whether domain-specific tokenisation improves model performance, whether character-based models provide a viable alternative to subword-based models, and which specialisation strategy is optimal given the constraints of limited pretraining data. Contrary to prior findings, our results show that in-domain tokenisation does not necessarily enhance performance. Most notably, adaptation consistently outperforms retraining, even with limited data, confirming its efficiency as the preferred strategy for resource-constrained domains. These insights provide valuable guidelines for developing specialised models in fields with limited textual resources.

1 Introduction

Transformer-based language models have achieved remarkable success through transfer learning, where models pretrained on large general-purpose corpora are fine-tuned for downstream tasks. Though relatively straightforward, this approach proves more challenging for tasks involving highly domain-specific fields or rare languages. In such settings, it might be beneficial – if not essential – to develop specialised language models (e.g. Lee et al., 2019; Chalkidis et al., 2020; Schweter et al., 2022; Yamshchikov et al., 2022). However, there is no consensus on the optimal specialisation strategy – whether to pretrain a model from scratch or to adapt an existing one.

Adapting involves further pretraining a generic model on domain-specific data. The approach has been shown to increase downstream performance (e.g. Peters et al., 2019; Gururangan et al., 2020), while preserving the broad linguistic knowledge acquired during the initial pretraining phase. However, this strategy does not grant infinite flexibility. A key obstacle to specialisation often lies in the model's predefined vocabulary. Commonly used subword tokenisation methods (e.g. Wu et al., 2016; Kudo and Richardson, 2018) tie the model to a fixed vocabulary, which can be suboptimal or utterly inappropriate for certain languages. Thus, several tokenisation-free models have been developed to address this issue. CANINE (Clark et al., 2022) and CHARFORMER (Tay et al., 2022) use a broad set of Unicode points as a vocabulary. However, while they circumvent tokenisation issues, these models often perform below their subword-based counterparts and significantly limit the maximum input sequence length.

While domain-specific tokenisation may improve model performance, modifying the model's vocabulary and embeddings requires retraining from scratch. This creates a critical trade-off between optimising tokenisation at the cost of pretrained knowledge or leveraging existing models despite suboptimal tokenisation. As the decision is constrained by the availability of pretraining data, the dilemma is particularly crucial in low-resource settings. Studies reporting superior results from retraining domain-specific models (e.g. Lee et al., 2019; Schweter et al., 2022) often have access to extensive training resources. While some studies claim better results with smaller pretraining datasets (Riemenschneider and Frank, 2023; Manjavacas Arevalo and Fonteyn, 2021), many

Proceedings of LaTeCH-CLfL 2025, pages 252–260

advocate for adapting over retraining (Konle and Jannidis, 2020; Gururangan et al., 2020), but few systematically control for tokenisation, leaving its precise role in model performance an open question.

This study examines these specialisation strategies in the context of classical scholarship, a field characterised by intense multilingualism, frequent code-switching, and highly domain-specific vocabulary. These factors, along with the extensive use of rare characters and diacritics, pose significant tokenisation challenges, particularly for texts with a high proportion of Latin and ancient Greek, which are often absent from generic multilingual models, making the field an ideal case study for tokenisation and specialisation strategies. While character-level models may offer greater adaptability, subword tokenisers often struggle with historical texts due to transcription errors, spelling variations, and morphological inconsistencies. Retraining also remains impractical given the scarcity of clean textual data. Although this study assembles the largest classics-related corpus to date, it remains constrained to 1.4B tokens for six languages, less than half the 3.3B tokens used by the first monolingual BERT (Devlin et al., 2019). This study assesses the impact of tokenisation, specialisation strategy, and the availability of domain-specific data. It asks three research questions: What are the benefits of in-domain tokenisation? Do character-based models provide a viable and more adaptable alternative to subword models? Finally, which specialisation strategy is most effective given the constraints of available data?

2 Related work

Domain- and language-specific tokenisation Rust et al. (2021) investigated the impact of tokenisation on the performance of various monolingual and multilingual language models. The authors found a beneficial impact in utilising dedicated monolingual tokenisers. More specifically, their research reveals that languages well-represented within the multilingual model's (mBERT) training data (e.g. English or Japanese) suffer minimal performance loss when compared to their monolingual counterparts. However, for languages less represented in the multilingual training data such as Finnish, the multilingual model's tokeniser performed worse than its monolingual counterpart. Consequently, mBERT performed significantly worse than the ad-hoc pretrained Finnish BERT, with an average drop of 3.8 points in F1 and accuracy across multiple downstream tasks. Detailing these analyses for large language models, Ali et al. (2024) show that the size and specificity of the vocabulary as well as the tokenisation method could account for differences of 5% to 15% in a wide variety of downstream tasks. Their study further demonstrates that while multilingual tokenisers are more efficient with a larger vocabulary (82,000 to 100,000), English monolingual tokenisers find an optimal range between 33,000 and 45,000 tokens. Table 1 provides a general comparison between models, pretraining and adapting data as well as tokenisation and specialisation strategies.

Table 1: Overview of domain-specific language models. Tok. indicates whether the tokeniser is Generic or In-Domain. Str. indicates the (best) specialisation strategy used: Re-Training or Adapting. Gen. and Spec. refer to the size of the generic and domain-specific pretraining data (in billion words).

Paper	Lang.	Tok.	Dom.	Gen.	Spec.	Str.
Devlin (2019)	Eng	G	Gen.	3.3	-	-
Liu (2019)	Eng	G	Gen.	33	-	-
Conneau (2020)	Multi	G	Gen.	250	-	-
Beltagy (2019)	Eng	ID	Science	3.3	3.17	RT
Lee (2019)	Eng	G	Medical	3.3	18	AD
Chalkidis (2020)	Eng	G	Law	3.3	2.5	AD
Manjavacas (2021)	Eng	ID	History	-	3.9	RT
Schweter (2022)	Multi	ID	History	-	30	RT
Gabay (2022)	Fra	ID	History	-	0.19	RT
Hosseini (2021)	Eng	G	History	3.3	5.4	AD
Brandsen (2021)	Dut	G	Archeo.	2.4	0.66	AD
Bamman (2020)	Lat	ID	Anc. Lg.	-	0.64	RT
Singh (2021)	Gre	ID	Anc. Lg.	-	<0.1	RT
Yamshchikov (2022)	Gre	G	Anc. Lg.	3.0	0.01	AD
Riemenschneider (2023)	Multi	ID.	Anc. Lg.	-	0.57	RT

Byte- and character-level tokenisation In an exploratory study, Choe et al. (2019) showed that byte-level language models could match the perplexity of word-level models when given the same parameter budget. Building upon their research, Clark et al. (2022) released CANINE, which shows gains over mBERT (Devlin et al., 2019) by working with characters instead of subword tokens. While different character-level tokenisation strategies have been proposed by competing models such as CHARFORMER (Tay et al., 2022) and CharacterBERT (Boukkouri et al., 2020), ByT5 (Xue et al., 2022) is the first to show that byte-level tokenisation can outperform word-level tokenisation on a wide range of tasks. The authors argue that byte-level tokenisation is more efficient than character-level tokenisation, given that it allows for a smaller vocabulary size and a more efficient use of the

model's parameters. In their experiments, ByT5 outperformed T5 (Raffel et al., 2020) on a wide range of tasks, including translation, summarisation, and question answering. Furthermore, the authors show that ByT5 is more robust to out-of-domain data than T5, suggesting that byte-level tokenisation can improve the generalisation capabilities of language models.

Domain-Specific Language Modelling Domain-specific language models have been developed for law (Chalkidis et al., 2020), biomedicine (Lee et al., 2019), science (Beltagy et al., 2019), history (Schweter et al., 2022), and classical philology (Riemenschneider and Frank, 2023). BioBERT (Lee et al., 2019), trained on 18B tokens of biomedical texts, retained BERT's vocabulary and weights, while SciBERT (Beltagy et al., 2019) explored both an adapted and a fully retrained model, demonstrating a slight superiority of in-domain tokenisation. In the legal domain, Chalkidis et al. (2020) compared retraining and adapting strategies for LegalBERT. However, the authors only assessed the performance of their adapted model on downstream legal tasks, allowing no comparison with the retrained model. Manjavacas Arevalo and Fonteyn (2021) compared the performance of a generic BERT, a historical-adapted BERT (Hosseini et al., 2021) and MacBERTh, a model pretrained on a corpus of 3.9B tokens composed of historical English exclusively. Despite sharing the same architecture, MacBERTh outperformed the two other models across a range of historical NLP tasks. This aligns with broader findings indicating that domain-specific models generally surpass generic ones within their fields (Schweter et al., 2022; Gabay et al., 2022; Konle and Jannidis, 2020; Manjavacas and Fonteyn, 2022; Gururangan et al., 2020).

Modelling classical scholarship and ancient languages In the field of classical studies, the development of domain-specific language models is made particularly crucial by the under-representation – if not the complete absence – of ancient languages in generic pretraining corpora. This absence is especially problematic in the case of ancient Greek, which exhibits a complex morphology, a rich inflectional system, and profuse usage of diacritics which radically distinguishes it from modern, simplified Greek. Recent years have therefore seen several efforts to develop language models tailored to ancient languages. Bamman and Burns (2020) released a LatinBERT which outperformed the state of the art. Most interestingly, these results are obtained with a relatively small pretraining corpus of 640M tokens gathered from diverse sources, showing that a dataset roughly one-fourth the size of BERT's could be leveraged to train a model achieving state-of-the-art performance. For ancient Greek, two notable studies stand out. Yamshchikov et al. (2022) adapt a modern Greek BERT to ancient Greek, while Singh et al. (2021) leverage online available corpora to train an ancient Greek model from scratch. Both studies show that language-specific models outperform generic monolingual and multilingual models on ancient Greek NLP tasks. Finally, in a more recent study, Riemenschneider and Frank (2023) released a collection of BERT- and T5-based ancient Greek and trilingual (Latin, Greek and English) models geared towards philology. Trained on slightly more data than the previous studies, the authors demonstrate that their models outperform the former models by a considerable margin across a range of philology-related tasks.

3 Pretraining data

In order to amass sufficient in-domain pretraining data to conduct our experiments, numerous classics-related corpora are gathered in a new Classical Scholarship Corpus (CSC). The final Classical Scholarship Corpus contains 1.4B tokens of domain-specific clean texts written in ancient Greek, Latin, English, French, German, and Italian. At the time of writing, our CSC is likely the largest corpus of clean texts gathered in the field so far. Texts are sourced through agreements with major publishers and providers or via web scraping. Hence, some corpora contain copyright-protected material. In total, 30 corpora are marshalled including notably Brill-KIEM[1], Internet Archive[2], the Corpus Thomisticum[3], Perseus and First1KGreek[4], and JSTOR[5]. The many challenges and peculiarities of classics-related data make data-cleaning a critical pre-processing step. This step notably involves the removal of documents with a high rate of optical character recognition errors. This is achieved by filtering out texts containing a low proportion (<65%) of alphanumeric characters or

[1] https://github.com/kiem-group/pdfParser
[2] https://web.archive.org/
[3] https://www.corpusthomisticum.org/
[4] https://www.opengreekandlatin.org/
[5] https://www.jstor.org/

a high proportion (>30%) of words not found in standard dictionaries. Corpora are also cleaned from recurring text spans such as headers, footers or webpage trademarks.

4 Methods

4.1 Evaluation methods

In line with Ali et al. (2024) and Rust et al. (2021), tokenisation is evaluated both intrinsically and extrinsically. Intrinsic evaluation is conducted using fertility, a widely adopted metric defined as the average number of tokens required to represent a word and measured on a 32M-tokens left-out set of the CSC. Extrinsic evaluation is established by the models' performance on downstream tasks.

These include four classics-related token classification tasks, all evaluated using macro-average F1 score, precision and recall. The first task involves Latin part-of-speech tagging with EvaLatin (Sprugnoli et al., 2020), a dataset comprising about 300,000 tokens. The second involves bibliographical entity recognition with EpiBau[6], a dataset of 1.1M English tokens annotated with ca. 37k entity mentions. Third comes multilingual named entity recognition with AjMC-NE-Corpus (Romanello and Najem-Meyer, 2024), a dataset of 111k tokens annotated with 7.3k named entities in English, German and French ($\text{AjNER}_{(\text{delenlfr})}$). Finally, text anchors recognition is evaluated with the AjMC-LL-Corpus[7], a dataset of 145k tokens annotated with 9.1k entity mentions in English, German and French (AjLR). Text anchors (lemmata) are specific to classical commentaries, and serve the purpose of linking commentary glosses to their corresponding text.

4.2 Base models

Two multilingual transformer encoders are retrained, adapted, and fine-tuned in our experiments: XLM-RoBERTa-base (Conneau et al., 2020), a subword-based, multilingual transformer encoder featuring a 250,000 SentencePiece tokeniser (Kudo and Richardson, 2018) and trained on 100 languages, including Latin and modern Greek, and CANINE-C (Clark et al., 2022), a character-based transformer encoder featuring a 40,000 Unicode points vocabulary and trained on 104 languages, including Latin and modern Greek[8]. Though CANINE's architecture necessarily differs from XLM-RoBERTa's, both models use the same 12-layers transformer stack and feature comparable parameter counts (121M vs 125M). Though these differences hamper an absolutely controlled comparison, our goal is also to provide researchers with an investigation of existing solutions and their respective upsides and shortcomings. Therefore, we also fine-tune two additional models for broader comparison purposes: hmBERT (Schweter et al., 2022), a BERT-based subword model trained on a 130GB corpus of historical texts and newspapers, including German, French, Swedish, Finnish and English, and PhilBERTa (Riemenschneider and Frank, 2023), a BERT-based model trained for classical scholarship, primarily geared towards Latin and ancient Greek, but also including English.

	hmB.	PhilB.	XLM-R	XLM-R (In-domain)		
Size (k)	33	64	250	250	82	33
Fertility	2.05	1.93	2.08	1.52	1.61	1.80

Table 2: Fertility scores of in-domain and generic tokenisers. Lower fertility scores indicate that fewer tokens are required to represent a word.

5 Experiments and results

5.1 What are the benefits of in-domain tokenisation?

Fertility To assess the benefits of specialised tokenisers, three XLM-R tokenisers are trained on the CSC: a large tokeniser of 250,000 tokens, equating XLM-R's original vocabulary size, an intermediary tokeniser of 82,000 tokens, and a small tokeniser of 33,000 tokens. Fertility scores are displayed in Table 2. As expected, fertility decreases (i.e. improves) with the domain-specificity of the tokeniser and its vocabulary size, showing that a larger specialised vocabulary requires fewer tokens to represent the same word.

5.1.1 Extrinsic evaluation

Models To evaluate the effects of tokenisation extrinsically, four XLM-R models are pretrained from scratch on in-domain data exclusively, with the different tokenisers. $\text{XLM-R}_{\text{RT-G-250}}$ is ReTrained using XLM-R's original Generic vocabulary. $\text{XLM-R}_{\text{RT-ID-(250|82|33)}}$ are ReTrained using In-Domain vocabularies of 250,000 82,000, and 33,000 tokens

[6] https://github.com/mromanello/EpibauCorpus

[7] Unpublished at time of writing as partially copyrighted.

[8] Since the authors did not release their implementation, a customised pretraining pipeline is used to train CANINE on in-domain data (See https://github.com/sven-nm/shiba-canine).

respectively. All models are pretrained for three epochs on the CSC and fine-tuned on each downstream task for 40 epochs, leaving other recommended hyperparameters unchanged.

Results Results are shown in Table 3. Surprisingly, the model retrained with the generic tokeniser (XLM-$R_{RT\text{-}G\text{-}250}$) outperforms those trained with in-domain tokenisers on all tasks, with an overall improvement of 8.4 points in F1 score over XLM-$R_{RT\text{-}ID\text{-}250}$. This result is particularly unexpected as the in-domain tokenisers are specifically designed to improve model performance on classical scholarship tasks. Interestingly, we observe a negative correlation between F1 scores and vocabulary sizes which is also incoherent with the fertility scores presented above: a better (i.e. lower) fertility usually implies a better tokeniser.

Analyses As no straightforward explanation justifies this result, further analyses are conducted. Our hypothesis is that in-domain tokenisation results in a substantially sparser token distribution, as specialised vocabularies contain more tokens fitting the precise needs of a relatively small domain-specific corpus. Hence as more tokens are used, their average frequency across the corpus diminishes. Token frequency was measured on a 300M subset of the CSC for each tokeniser and supports this hypothesis. While XLM-R's generic tokeniser only needs 95,754 of its 250,000 tokens in its vocabulary to segment the corpus, its in-domain counterpart uses 246.864 tokens. Hence, the model based on the former benefits from 6,137 token occurrences on average, while the model based on the latter must learn from a much sparser distribution of tokens, averaging to 1,701 occurrences per unique token. Furthermore, quantiles show that the generic vocabulary also leads to a much higher concentration of used tokens, with 75% of used tokens having over 1.1k occurences, versus 0.5k for XLM-R-ID-250.

While XLM-$R_{RT\text{-}ID\text{-}33}$ performs significantly better than its 250 and 82 counterparts, it still does not surpass the model based on the generic tokeniser. This result raises two considerations. First, it supports the idea that enhancing token density leads to significantly better results, especially in the case of relatively limited pretraining data. While intrinsic evaluation metrics such as fertility may provide a valuable insight on tokeniser performance in domains provided with abundant training data, the results provided here show token density to be a significantly more reliable predictor of extrinsic performance. Hence, researchers working in resource-limited environments should be advised to take this metric into account when choosing the vocabulary size of an in-domain tokeniser. Second, it shows that contrary to a generally supported claim (Rust et al., 2021; Beltagy et al., 2019; Ali et al., 2024), in-domain tokenisation does not necessarily imply better model performance. One possible explanation for this outcome is that the tokeniser's training corpus may simply be too limited in size to support the development of robust subword units. While in-domain tokenisers may lead to the best performance when given sufficient token density, they still do not outperform the generic tokeniser, suggesting that a more robust tokenisation might be obtained by training the tokeniser on larger corpora.

5.2 Do character-based models provide a viable and more adaptable alternative to subword models?

This second series of experiments provides a comparison between generic and adapted versions of CANINE (character-based) and XLM-R (subword-based). Generic versions (XLM-R and CANINE) use the checkpoints provided by each model's authors. Adapted versions (XLM-R_{AD} and CANINE-C_{AD}) are further pretrained on the CSC for three epochs. The last adapting checkpoints were shown by pre-tests to yield the best downstream results and are therefore fine-tuned for 40 epochs on each downstream task. Though XLM-R largely outperforms CANINE-C on all tasks, the latter shows much higher gains from adaptation, with improvements up to 15% F1 score for $AjNER_{en}$. Though the model's performance is still lower than XLM-R's, the gap is significantly reduced. Interestingly, adaptation significantly degrades the performance of CANINE-C on lemma recognition, while it generally benefits on all other tasks and models. Error analysis shows this effect to be due only to a significant precision drop on greek-only entities.

Discussion It remains to be seen whether the higher adaptability of CANINE-C is due to its character-based tokenisation or to other factors such as the model's architecture or pretraining objectives, which are not controlled in these experiments. However, these results confirm the model's claimed adaptability across languages (Clark et al., 2022) and suggest that researchers thoroughly de-

Model	EpiBau	EvaLat.	AjLR	AjNER$_{de}$	AjNER$_{fr}$	AjNER$_{en}$	Avg
hmBERT	0.847	0.934	0.889	0.904	0.835	0.846	0.876
PhilBERTa	0.781	0.925	0.619	0.775	0.602	0.690	0.732
CANINE-C	0.729	0.890	0.749	0.809	0.712	0.616	0.751
CANINE-C$_{AD}$	0.794	0.899	0.708	0.824	0.789	0.766	0.796
XLM-R	0.854	0.944	0.875	0.907	0.856	0.838	0.879
XLM-R$_{RT\text{-}G\text{-}250}$	0.818	0.912	0.807	0.879	0.802	0.794	0.835
XLM-R$_{RT\text{-}ID\text{-}250}$	0.788	0.895	0.668	0.809	0.722	0.683	0.761
XLM-R$_{RT\text{-}ID\text{-}82}$	0.769	0.900	0.668	0.824	0.783	0.735	0.780
XLM-R$_{RT\text{-}ID\text{-}33}$	0.787	0.905	0.761	0.848	0.814	0.795	0.818
XLM-R$_{RT\text{-}G\text{-}250\text{-}300M}$	0.623	0.684	0.578	0.670	0.587	0.542	0.614
XLM-R$_{RT\text{-}G\text{-}250\text{-}600M}$	0.734	0.771	0.711	0.786	0.701	0.687	0.732
XLM-R$_{AD}$	0.844	0.948	0.896	**0.935**	0.871	0.869	0.894
XLM-R$_{AD\text{-}EP5}$	**0.868**	**0.952**	0.896	0.924	**0.895**	**0.886**	**0.903**
XLM-R$_{AD\text{-}300M}$	0.860	0.948	0.897	0.911	0.886	0.867	0.895
XLM-R$_{AD\text{-}600M}$	0.858	0.947	**0.909**	0.923	0.886	0.875	0.900

Table 3: F1 scores of all models across downstream tasks. Results are reported for models with three epochs of pretraining. The average F1 score is equally weighted across all tasks. The best results across all models are highlighted in bold.

prived of generic subword models usable in their research field may find significant benefits in adapting CANINE-C to their domain. However, in the current state, CANINE remains significantly less capable than XLM-R.

5.3 Which specialisation strategy is most effective given the constraints of available data?

The goal of this last series of experiments is to determine whether retraining or adapting yields best results depending on the quantity of available data. Although limited to the case of classics, these experiments may provide valuable insights for other domains with similar characteristics.

Models To address the question, six variants of XLM-R are trained, each being either ReTrained or ADapted on 300M, 600M or 1.4B tokens (XLM-R$_{(RT|AD)\text{-}(300M|600M|\emptyset)}$). As the generic tokeniser has been shown to yield the best results in the first research question, it is used for the three retrained models, also allowing for a fairer comparison with adapted models, as the latter necessarily keeps the model's original vocabulary. In the experiments involving a subset of the pretraining data, model checkpoints are compared after an equal number of training steps as opposed to an equal number of epochs. This method is chosen in order to keep the amount of pretraining tokens the only changing variable.

Results XLM-R$_{AD}$ outperforms all other models trained on the entirety of CSC by a significant margin. This result shows the superiority of adapted models over both retrained and generic models. As XLM-R$_{AD}$ performs best, it is also further pretrained for two additional epochs, reaching a total of five epochs (XLM-R$_{AD\text{-}EP5}$), showing an overall improvement in performance and producing the best model overall. Table 3 also shows the results of models pretrained on 300M, 600M, and 1.4B tokens. Surprisingly, results show that 300M and 600M models yield even better results than the model trained on the entire corpus (XLM-R$_{AD}$). This unexpected outcome might be due to the fact that models are here compared at an equal number of training steps, and not at an equal number of training epochs. This implies that data-ablated models have been exposed to fewer distinct examples but have encountered these examples with greater frequency. When compared with a model trained for an approximately equal number of epochs on the entire corpus, the latter overtakes the former. Hence, the model trained on the entire corpus continues to improve after three epochs

and finally yields the best results at five epochs, which corresponds approximately to the number of epochs run by XLM-R$_{\text{AD-300M}}$. In any case, the observed differences are very small, and lead to the more reasonable conclusion that the model's performance is not significantly affected[9] by the amount of in-domain data it is further pretrained on. This very encouraging result suggests that researchers working in resource-constrained environments can still benefit from adapting models to their domain, even if they only have access to a small amount of data.

This is not the case with retrained models, whose performance pronouncedly drops when further pretrained on ablated data. This result is consistent with the trend observed in recent years, which shows that the results of pretrained models are very sensitive to the amount of data they are pretrained on. Hence, while adapting XLM-R with 1.4B as opposed to 300M tokens causes the model's average performance to drop 0.1%, retraining, the same deprivation implies a remarkable drop of 22.9% F1 score on average.

6 General discussion

The importance of tokenisation The conclusions of these experiments are multifaceted. First, experiments confirm previous findings on the critical role of tokenisation (Ali et al., 2024; Rust et al., 2021). Our experiments compare the performance of two XLM-R models pretrained on the same data but using two distinct tokenisers of equal size: a generic and a domain-specific tokeniser. As shown in Table 3, our results reveals differences ranging up to 7% on average downstream F1 scores. This substantial difference underscores the necessity of a meticulous examination of this oft-overlooked stage of model development.

Second, unlike previous studies, these experiments illustrate that in-domain tokenisation does not necessarily lead to better performance. Analyses indicate that intrinsic tokenisation evaluation methods relying on fertility do not correlate with downstream results. On the contrary, in a resource-limited environment, lower (i.e. better) fertility also leads to a lower average token frequency and less performant models. We argue that in low-resource settings, tokenisation should balance input sequence length and token-type density. While high-fertility, small-sized tokenisers produce longer input sequences by breaking words into smaller subwords, they also enable more frequent representation of each token within the corpus, which correlates with model performance. This study therefore advocates the adoption of token density as a novel intrinsic evaluation metric.

Although smaller in-domain tokenisers consistently yield better results than their larger in-domain counterparts, they still do not surpass the performance of a larger, generic tokeniser. This improved performance may be attributable to the substantially larger size training corpus, which could favour a more robust and efficient vocabulary. This hypothesis is left for future research.

The potential of character-based models The second series of experiments consistently shows that CANINE-C is outperformed by XLM-R across all downstream tasks. However, the limits of this comparison must be highlighted. Notably, the two models are pretrained on different corpora with distinct training objectives and exhibit slight architectural differences, with CANINE-C incorporating downsampling and upsampling convolutional layers around its central transformer blocks. Thus, the findings merely indicate that even in a domain where tokenisation is suboptimal, XLM-R achieves better performance than CANINE-C.

Second, the results demonstrate CANINE-C's strong adaptability to the domain, with the adapted model yielding an average improvement of 5% in F1 score over the generic model. This improvement is substantial when compared to XLM-R's average improvement of 1.5% in F1 score. Researchers working with highly specific or underrepresented languages not covered by large multilingual models may therefore find character-based models like CANINE-C advantageous when adapted to their domain.

The superiority of adapted models Finally, the third series of experiments shows an undeniable superiority of adapted models over retrained and generic models, regardless of the amount of available pretraining data. This result aligns with the principle that adaptation preserves the extensive linguistic knowledge embedded in the generic base-model, a solid foundation difficult to replicate when training from scratch on limited resources. Although retraining has shown success

[9] McNemar's tests show average differences above 0.038% in F1 score to be statiscally significant ($p < .05$). Thus, the difference between XLM-R$_{\text{AD-EP5}}$ and XLM-R$_{\text{AD-600M}}$ is not statistically significant, as is the difference between XLM-R$_{\text{AD}}$ and XLM-R$_{\text{AD-300M}}$.

in domain-specific fields with ample pretraining corpora, such as biomedical or legal domains, resource-constrained fields appear to benefit most by leveraging the power of scale utilised by the model during its original pretraining. The findings also reveal that an adaptation corpus of 300M tokens already achieves 75% of the overall performance gains, indicating adaptation as an efficient and resource-effective specialisation strategy. As the superiority of adaptation over retraining is especially evident in data-ablation scenarios, it suggests that researchers and practitioners working with limited pretraining data should prioritise this approach. Moreover, adaptation may offer the only viable pathway to specialising large language models, an approach also left for future work.

7 Conclusion

This study provides a comprehensive analysis of the impact of tokenisation and specialisation strategies on the performance of language models in the field of classical scholarship. Our results show that in-domain tokenisation does not necessarily lead to better model performance in a resource-constrained environment, and that token density is a more reliable predictor of extrinsic performance. Our experiments also show that character-based models can offer a viable alternative to subword models, especially when adapted. Finally, we show that adaptation is the most effective specialisation strategy in a resource-constrained environment, and that even relatively small adaptation corpora can yield significant performance gains. These findings provide valuable insights for researchers working in resource-limited environments and highlight the importance of tokenisation and specialisation strategies in the development of large language models. We leave the investigation of our findings in other historical domains for future work, make our models available to the research community[10].

Acknowledgments

This research has been supported by the Swiss National Science Foundation under an Ambizione grant PZ00P1_186033.

[10]See `XLM-R-for-classics*` models at https://huggingface.co/sven-nm/.

References

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Schulze Buschhoff, Charvi Jain, Alexander Arno Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. 2024. Tokenizer choice for llm training: Negligible or crucial? *Preprint*, arXiv:2310.08754.

David Bamman and Patrick J. Burns. 2020. Latin bert: A contextual language model for classical philology. *arXiv:2009.10053 [cs]*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv:1903.10676 [cs]*.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Junichi Tsujii. 2020. Characterbert: Reconciling elmo and bert for word-level open-vocabulary representations from characters. *Preprint*, arXiv:2010.10392.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *Preprint*, arXiv:2010.02559.

Dokook Choe, Rami Al-Rfou, Mandy Guo, Heeyoung Lee, and Noah Constant. 2019. Bridging the gap for tokenizer-free language models. *Preprint*, arXiv:1908.10322.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805 [cs]*.

Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette, and Benoît Sagot. 2022. 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*, pages 3367–3374, Marseille, France. European Language Resources Association.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey,

and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *Preprint*, arXiv:2004.10964.

Kasra Hosseini, Kaspar Beelen, Giovanni Colavizza, and Mariona Coll Ardanuy. 2021. Neural language models for nineteenth-century english. *Journal of Open Humanities Data*, 7(0).

Leonard Konle and Fotis Jannidis. 2020. Domain and task adaptive pretraining for language models. In *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *Preprint*, arXiv:1808.06226.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, page btz682.

Enrique Manjavacas and Lauren Fonteyn. 2022. Adapting vs. pre-training language models for historical languages. *Journal of Data Mining & Digital Humanities*, NLP4DH(Digital humanities in languages).

Enrique Manjavacas Arevalo and Lauren Fonteyn. 2021. 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*, pages 23–36, NIT Silchar, India. NLP Association of India (NLPAI).

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. *arXiv:1903.05987 [cs]*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv:1910.10683 [cs, stat]*.

Frederick Riemenschneider and Anette Frank. 2023. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.

Matteo Romanello and Sven Najem-Meyer. 2024. A named entity-annotated corpus of 19th century classical commentaries. *Journal of Open Humanities Data*, 10(1):1.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. *Preprint*, arXiv:2012.15613.

Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmbert: Historical multilingual language models for named entity recognition. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, volume 3180 of *CEUR Workshop Proceedings*, pages 1109–1129, Bologna, Italy. CEUR.

Pranaydeep Singh, Gorik Rutten, and Els Lefever. 2021. A pilot study for bert language modelling and morphological analysis for ancient and medieval greek. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 128–137, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

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).

Yi Tay, Vinh Q. Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2022. Charformer: Fast character transformers via gradient-based subword tokenization. *Preprint*, arXiv:2106.12672.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144 [cs]*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Preprint*, arXiv:2105.13626.

Ivan Yamshchikov, Alexey Tikhonov, Yorgos Pantis, Charlotte Schubert, and Jürgen Jost. 2022. Bert in plutarch's shadows. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6071–6080, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.