



BABYLMS

FOR SAMPLE-EFFICIENT LANGUAGE MODELLING OF ANCIENT GREEK

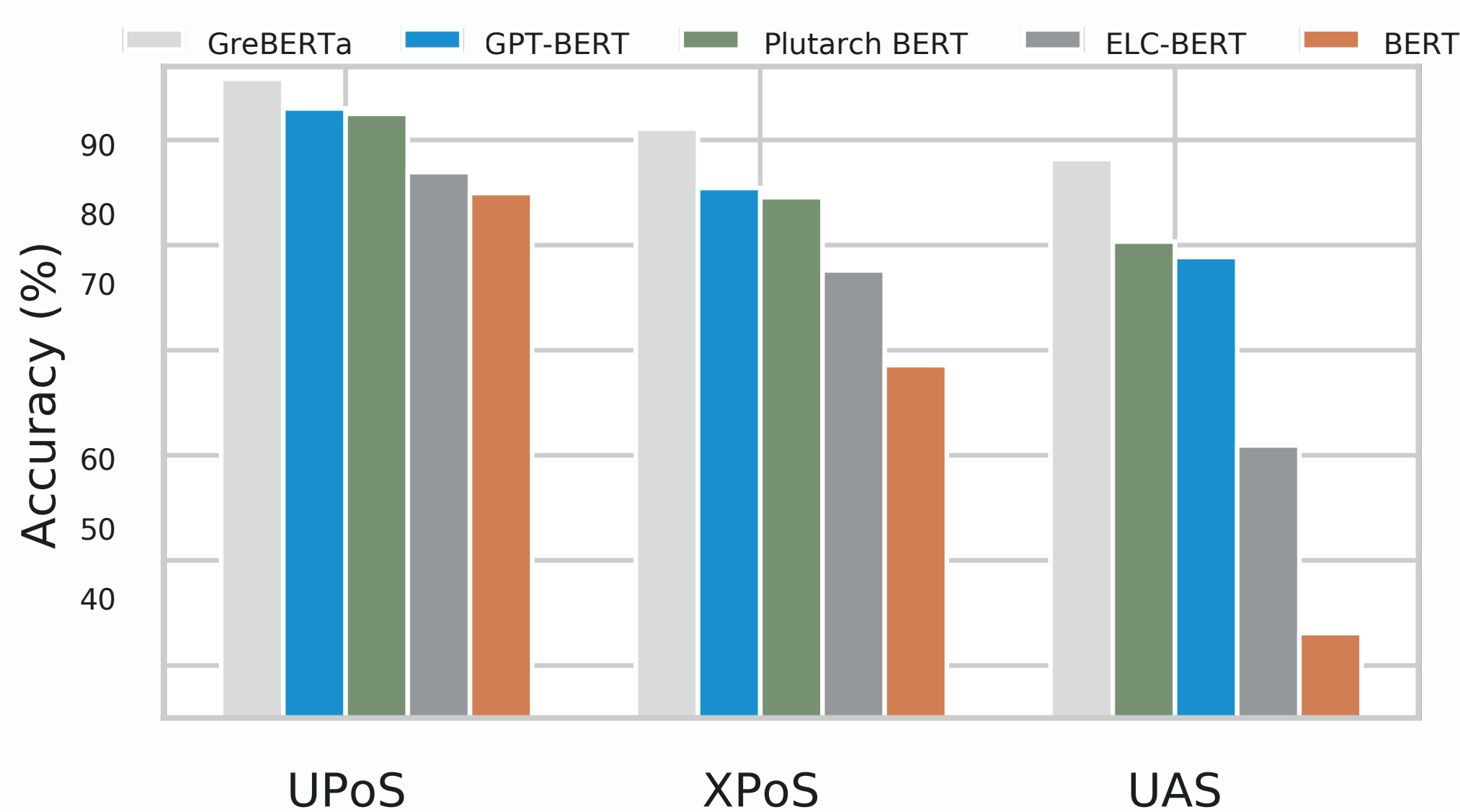
INTRODUCTION

- Ancient Greek, despite its cultural and historical significance to Western society, has limited publicly available datasets.
- The BabyLM Challenge is a shared task for developing sample-efficient architectures, such as ELC-BERT and GPT-BERT.
- We train these BabyLMs and a baseline BERT model on limited Ancient Greek data.
- The models' performance is compared across three downstream Natural Language Understanding (NLU) tasks.
- We compare our results with previous Ancient Greek language models trained on significantly more data, which we cast as 'skylines'.

OBJECTIVES

1. Investigate the impact of including diacritics in the pretraining of an Ancient Greek language model.
2. Ablate the modifications to ELC-BERT to determine which lead to improved downstream performance.
3. Determine the optimal causal-to-masked ratio for pretraining an Ancient Greek GPT-BERT.
4. **Investigate the suitability of sample-efficient BabyLM architectures for the low-resource context of Ancient Greek.**

Downstream Performance Across All NLU Tasks



CONCLUSIONS

- GPT-BERT is more sample efficient than ELC-BERT and BERT when trained on the same limited Ancient Greek corpus.
- Although GPT-BERT outperforms one of the skyline models on PoS tagging, it is outperformed by the other skyline models across all downstream tasks.
- **Future research in Ancient Greek language modelling should pursue larger, high-quality datasets as well as more sample-efficient architectures.**

MODELS

1 BERT

A Transformer-based language model trained using Masked Language Modelling (MLM), which captures bidirectional context.

2 ELC-BERT

A BabyLM that employs a layer-weighting mechanism that assigns learnable weights to each layer's output.

3 GPT-BERT

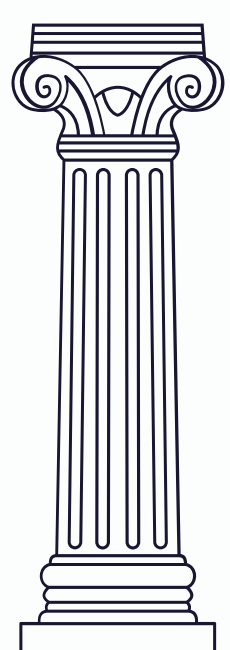
A hybrid BabyLM that combines the MLM training objective (from BERT) and the Causal Language Modelling objective (from GPT).

METHODOLOGY

- Curate and preprocess a collection of digitized Ancient Greek texts.
- Split the processed corpus into training, validation, and test sets.
- Train a custom Ancient Greek tokenizer.
- Pretrain our models on the training split of the curated Ancient Greek corpus (~25M words).
- Finetune our models for each task using the Universal Dependencies Ancient Greek Perseus treebank.

RESULTS

- **GPT-BERT significantly outperforms both ELC-BERT and BERT across all downstream NLU tasks.**
- The improved sample efficiency of GPT-BERT leads to greater performance gains on more difficult tasks.
- GPT-BERT marginally outperforms a skyline model, Plutarch BERT, on UPoS and XPoS tagging.
- Another skyline model, GreBERTa, substantially outperforms GPT-BERT across all NLU tasks..



University of Cape Town
Department of Computer Science
TEL: +27 (0)21 650 2663
Website: <https://sit.uct.ac.za/>

Dimitrios Dalakas
Kimon Christelis
Stylios Dalakas

Authors

✉ dlkdim001@myuct.ac.za
✉ chrkim008@myuct.ac.za
✉ dlksty001@myuct.ac.za

Supervisor

Francois Meyer
✉ francois.meyer@uct.ac.za

