PortBERT: Navigating the Depths of Portuguese Language Models

Raphael Scheible-Schmitt[1,2,3], **Henry He**[1], **Armando B. Mendes** [2]

[1]School of Computation, Information and Technology, Technical University of Munich, Munich, Germany
[2]IS[2]E - Intelligent Systems, Science and Engineering, LIACC polo on Azores University, Ponta Delgada, Portugal
[3]Institute of General Practice, Faculty of Medicine and Medical Center, University of Freiburg
Correspondence: raphael.scheible@tum.de

Abstract

Transformer models dominate modern NLP, but efficient, language-specific models remain scarce. In Portuguese, most focus on scale or accuracy, often neglecting training and deployment efficiency. In the present work, we introduce PortBERT, a family of RoBERTa-based language models for Portuguese, designed to balance performance and efficiency. Trained from scratch on over 450 GB of deduplicated and filtered mC4 and OSCAR23 from CulturaX using fairseq, PortBERT leverages byte-level BPE tokenization and stable pre-training routines across both GPU and TPU processors. We release two variants, $\text{PortBERT}_{\text{base}}$ and $\text{PortBERT}_{\text{large}}$, and evaluate them on ExtraGLUE, a suite of translated GLUE and SuperGLUE tasks. Both models perform competitively, matching or surpassing existing monolingual and multilingual models. Beyond accuracy, we report training and inference times as well as fine-tuning throughput, providing practical insights into model efficiency. PortBERT thus complements prior work by addressing the underexplored dimension of compute-performance tradeoffs in Portuguese NLP. We release all models on Huggingface and provide fairseq checkpoints to support further research and applications.

1 Introduction

The development of neural language models has profoundly shaped natural language processing (NLP), particularly through the advent of transformer-based architectures such as BERT (Devlin et al., 2019) and its optimized variant RoBERTa (Liu et al., 2019). These models, which learn contextualized word representations via self-supervised pretraining, have become foundational across a wide range of NLP tasks. While early efforts prioritized English or multilingual solutions, research has shown that language-specific pretraining on high-quality, monolingual corpora often yields superior results for the target language (Delobelle et al., 2020; Scheible et al., 2024).

In Portuguese NLP, monolingual transformer models such as BERTimbau (Souza et al., 2020) and AlBERTina (Rodrigues et al., 2023) have marked important milestones. More recently, multilingual alternatives like XLM-RoBERTa (Chan, 2020) and EuroBERT (Boizard et al., 2025) have demonstrated strong cross-lingual performance by scaling up to billions of parameters. EuroBERT, in particular, follows the "Modern BERT" framework (Warner et al., 2024), which revisits encoder-based models with streamlined design and improved training efficiency. While decoder-only models continue to dominate general-purpose NLP, these developments show that encoder-based masked language models (MLMs) remain competitive and relevant.

However, many of these advancements come at considerable computational cost. As NLP systems move closer to real-world applications, ranging from chatbots and document pipelines to tasks such as named entity recognition, sentence classification, or part-of-speech tagging, efficiency becomes a central concern. Models deployed in production must often meet strict requirements in terms of latency, memory usage, and energy consumption. Prior work has shown that compact transformer models can offer significant speed-ups with minimal impact on performance (Sanh et al., 2020; Jiao et al., 2020). Yet, most Portuguese models focus primarily on accuracy, offering limited insight into training efficiency, hardware utilization, or deployment tradeoffs.

To address this gap, we introduce PortBERT, a family of RoBERTa-based encoder models tailored for Portuguese. PortBERT is trained from scratch on over 450 GB of deduplicated text from CulturaX (Nguyen et al., 2023), combining data from mC4 (Xue et al., 2021) and OSCAR23 (Jansen

Proceedings of the Workshop on Beyond English: NLP for all Languages in an Era of LLM from Texts associated with RANLP 2025, pages 59–71, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-105-9-008

et al., 2022). Following Scheible et al. (2024), we construct a byte-level BPE vocabulary with 52k tokens using Hugging Face's tokenizer tools, which helps improve token efficiency and compression, an effect observed in prior work on Dutch and German (Delobelle et al., 2020; Scheible et al., 2024).

Pretraining is performed using the fairseq framework: the base variant is trained on 8 NVIDIA A40 GPUs, and the large variant on a TPUv4-128 pod. PortBERT retains the standard RoBERTa architecture without architectural modifications like sparse attention or extended context. Instead, it emphasizes a balanced design that prioritizes pretraining efficiency, inference throughput, and downstream accuracy. While not designed to match the scale of models like EuroBERT (Boizard et al., 2025) or decoder-based LLMs, PortBERT offers a robust, reproducible, and accessible alternative for practical Portuguese NLP.

The main contributions of this study are:

- We provide two variants, $\text{PortBERT}_{\text{base}}$ and $\text{PortBERT}_{\text{large}}$, trained respectively on GPUs and a TPUv4 pod, and release both models under an open-source license.
- We evaluate PortBERT on the ExtraGLUE benchmark, showing that both models perform competitively.
- We report training time for both pretraining and downstream fine-tuning, and include throughput metrics for fine-tuning to support transparent evaluation of efficiency.

2 Related works

In recent years, a growing number of transformer-based language models have been developed for Portuguese. These include both monolingual models trained specifically on Portuguese corpora and multilingual models that support a wide range of languages. Table 1 summarizes these models, their architectures, and training data sources.

BERTimbau (Souza et al., 2020) was one of the first monolingual BERT-style models for Portuguese, available in base and large versions. It was trained on a mix of BrWaC (Wagner Filho et al., 2018), Portuguese Wikipedia, and a news corpus using whole-word masking (WWM) over one million steps.

AiBERTa[1] (Miquelina et al., 2022; Santos et al., 2025a) follows a RoBERTa-style architecture and is trained on a curated subset of Portuguese periodical websites archived in `Arquivo.pt`, a national web archive. These periodicals range from national newspapers like Público to smaller regional outlets, providing well-written and structurally consistent Portuguese text.

AlBERTina (Rodrigues et al., 2023) adopts the ALBERT architecture (Lan et al., 2020), introducing parameter-sharing and embedding factorization. The models were trained on the January 2023 version of OSCAR, as well as DCEP, Europarl, and ParlamentoPT. Separate variants exist for Brazilian and European Portuguese.

RoBERTa PT (Santos et al., 2021) was trained on 10 million English and 10 million Portuguese sentences from the OSCAR corpus. Despite its bilingual setup and relatively small training corpus, the model is widely cited and has been evaluated in various Portuguese NLP tasks.

RoBERTaCrawlPT and RoBERTaLexPT (Garcia et al., 2024) are both RoBERTa-based models developed for Portuguese. RoBERTaCrawlPT uses CrawlPT, a combined corpus comprising BrWaC, CC100-PT, and OSCAR23-PT. RoBERTaLexPT targets legal-domain applications and adds LegalPT, a corpus aggregating diverse legal documents totaling up to 125 GiB.

Among multilingual models, XLM-RoBERTa (Chan, 2020) can be used for Portuguese tasks. It is trained on 2.5 TB of filtered Common Crawl data in over 100 languages, including Portuguese.

EuroBERT (Boizard et al., 2025) is a more recent multilingual encoder model that spans 15 European languages, including Portuguese. It follows the Modern BERT architecture (Warner et al., 2024), with design choices optimized for scalability and efficiency. Its training data includes CulturaX (Nguyen et al., 2023), FineWeb (Penedo et al., 2024), EuroLLM (Martins et al., 2024), and code-related corpora such as The Stack v2 (Lozhkov et al., 2024) and Proof-Pile-2 (Azerbayev et al., 2024).

While many Portuguese models report strong downstream performance, few document training efficiency or hardware usage. PortBERT complements this work by offering initial insights into these often underreported aspects.

[1] https://huggingface.co/AiBERTa/aiberta-d-2000M-random

Model	Architecture	Language(s)	Training Data Sources
BERTimbau	BERT	1	BrWaC, Wikipedia, news corpora
AiBERTa	RoBERTa	1	Arquivo.pt (Portuguese periodicals)
AlBERTina PTPT/PTBR	ALBERT	1	OSCAR 23, DCEP, Europarl, ParlamentoPT
RoBERTa PT	RoBERTa	2	OSCAR (10M sentences each language)
$\text{RoBERTaCrawlPT}_{\text{base}}$	RoBERTa	1	CrawlPT (brWaC, CC100-PT, OSCAR23-PT)
$\text{RoBERTaLexPT}_{\text{base}}$	RoBERTa	1	CrawlPT, LegalPT (aggregated legal corpus)
XLM-RoBERTa	RoBERTa	100+	CommonCrawl (2.5TB, filtered)
EuroBERT	Modern BERT	15	CulturaX, FineWeb, EuroLLM, The Stack v2, Proof-Pile-2

Table 1: Overview of transformer-based language models relevant to Portuguese. The table lists architecture type, language coverage, and training data sources.

3 Methods

3.1 Corpus

To pre-train PortBERT, we used the Portuguese portions of mC4 and OSCAR23 (Jansen et al., 2022), two large-scale web corpora. The original size of Portuguese mC4 was approximately 453.1 GB, and OSCAR23 contributed 96.9 GB, totaling 550 GB of raw data. To reduce redundancy and improve quality, we relied on the deduplicated and filtered versions provided by CulturaX (Nguyen et al., 2023), which together amount to 456.6 GB, a size reduction of roughly 17% (93.4 GB). This large and diverse dataset ensures broad linguistic coverage with reduced duplication and noise compared to raw crawled corpora. CulturaX applied language identification, quality filtering, and deduplication to produce these cleaned subsets.

3.2 Pre-processing

RoBERTa employs the byte pair encoding (BPE) tokenizer originally introduced with GPT-2 (Radford et al., 2019), which processes raw text directly without requiring pre-tokenization or language-specific tools like Moses (Koehn et al., 2007). While this tokenizer was trained on English corpora, we followed the approach taken for GottBERT (Scheible et al., 2024) by training a dedicated Portuguese tokenizer. Using 40 GB of randomly sampled Portuguese corpus data, we created a 52k-token vocabulary optimized for the language. Although we did not explicitly measure the impact on file size or task performance for PortBERT, similar adaptations in Dutch (Delobelle et al., 2020) and German (Scheible et al., 2024) have demonstrated benefits in both respects. In our experience, a 40 GB sample is sufficient for the subword distribution to converge, and extending vocabulary training to the full corpus would add considerable overhead with little expected benefit.

3.3 Pre-training

Similar to GottBERT, we pre-trained the $\text{PortBERT}_{\text{base}}$ and $\text{PortBERT}_{\text{large}}$ models using the Fairseq framework. $\text{PortBERT}_{\text{large}}$ was trained on a 128-core TPUv4 pod (Jouppi et al., 2023), while $\text{PortBERT}_{\text{base}}$ was trained on a cluster of 8 NVIDIA A40 GPUs, using the same training corpus and identical optimization hyperparameters. Mixed-precision training (fp16) was disabled for the GPU setup and not supported by the TPU implementation used, ensuring that both models were trained entirely in full precision (fp32). This controlled setup enables a direct comparison of hardware-level training efficiency across compute architectures, without numerical precision optimizations acting as confounding factors. Both models were trained on Portuguese OSCAR data using the RoBERTa architecture. The $\text{PortBERT}_{\text{base}}$ model completed training in approximately 27 days (2,331,939 seconds), while $\text{PortBERT}_{\text{large}}$ required around 6.2 days (531,807 seconds). We used the standard RoBERTa pretraining schedule with 100k update steps, a batch size of 8k, a 10k-step warmup, and polynomial learning rate decay. The base model used a peak learning rate of 0.0004, and the large model 0.00015. As with GottBERT, we evaluated after each epoch and stored checkpoints throughout training. However, since the dataset size only permitted approximately four epochs, the final checkpoint coincided with the best-performing one.

3.4 Downstream Tasks

Based on the pre-trained BERT models, we fine-tuned several downstream tasks using the training scripts provided by Huggingface (Wolf et al., 2019). Hyperparameter optimization was performed via grid search, focusing on batch size and learning rate. Each task was trained for a maximum of 10 epochs, and the experiments were orchestrated using NNI (Microsoft, 2025) on NVIDIA A40 GPUs.

To assess model performance, each downstream task was fine-tuned 28 times using different combinations of batch sizes and learning rates. Since no separate test set was available, we selected the best-performing checkpoint based on validation set scores. The final performance figures reported for each model and task reflect the best result among these 28 validation-based runs. For comparison, we benchmarked our models against eleven other Portuguese language models.

We evaluated the models on ExtraGLUE (Santos et al., 2025b), a Portuguese adaptation of the GLUE benchmark. This suite consists of selected tasks from GLUE and SuperGLUE that were automatically translated into Portuguese, enabling language-specific assessment and ensuring that model performance reflects capabilities in the target language context.

To account for varying input lengths across tasks, we configured the maximum input sequence length individually per task based on the maximum observed input lengths after tokenization across all evaluated models: 192 tokens for MRPC and WNLI, 320 tokens for STS-B, and 512 tokens for RTE. This ensured full coverage of the datasets while avoiding unnecessary padding and memory overhead.

STS-B The Semantic Textual Similarity Benchmark (STS-B) task evaluates the model's ability to assess the semantic similarity between two sentences. Each sentence pair is assigned a similarity score ranging from 0 (completely dissimilar) to 5 (semantically equivalent). Following standard practice, we report the mean of Pearson and Spearman correlation coefficients between predicted and gold scores.

RTE The Recognizing Textual Entailment (RTE) task consists of binary classification, where the model must determine whether a given hypothesis logically follows from a provided premise. This task evaluates the model's capacity for inference and semantic reasoning.

WNLI The Winograd Natural Language Inference (WNLI) task is a coreference resolution challenge cast as binary entailment. It requires the model to resolve ambiguous pronouns and determine whether a hypothesis follows from a premise. Despite its small size and challenging structure, it is retained for completeness and consistency with GLUE-style benchmarks.

MRPC The Microsoft Research Paraphrase Corpus (MRPC) task is a binary classification problem where the model must decide whether two sentences are semantically equivalent. Evaluation is based on both accuracy and F1 score, reflecting the importance of both precision and recall in paraphrase detection.

3.5 Model Configurations and Properties

The number of parameters in BERT-like models varies significantly depending on their architecture (see Table 2). The base version of BERT, such as $\text{BERTimbau}_{\text{base}}$, has approximately 109 million parameters, while large versions like $\text{BERTimbau}_{\text{large}}$ expand to over 334 million. RoBERTa variants used in Portuguese NLP, such as $\text{RoBERTaCrawlPT}_{\text{base}}$ and $\text{RoBERTaLexPT}_{\text{base}}$, feature around 125 million parameters, comparable to $\text{PortBERT}_{\text{base}}$ (126M). The large PortBERT model increases this to 357 million, positioning it close to $\text{BERTimbau}_{\text{large}}$ while retaining RoBERTa's efficiency characteristics.

Multilingual models such as XLM-RoBERTa are designed for cross-lingual tasks, with the base version containing 278 million parameters and the large version 560 million. These parameter counts make them substantially larger than monolingual base models, but beneficial in zero-shot or cross-lingual scenarios (Eronen et al., 2023).

The AiBERTa and AlBERTina families offer diverse parameter ranges. All AiBERTa variants (regardless of source or domain configuration) have approximately 101 million parameters, with a smaller vocabulary size of 20,000. The AlBERTina models, in contrast, range from 138 million (100M variants) to over 1.5 billion parameters for the 1.5B variants, reflecting a significant increase in capacity and vocabulary size (up to 128,100 tokens). These models serve different use cases depending on the required balance between compute and performance.

Finally, EuroBERT models span from 210 million parameters in the 210M variant to over 2.1 billion in the 2.1B variant. They provide a scalable foundation for multilingual or European-centric tasks, emphasizing both vocabulary coverage and model depth.

Table 2: The size of the vocabulary and the size of the parameters are shown for the model types used in this study. This table does not show other design differences of the models. Values were extracted using Huggingface's transformers library. Models are sorted by number of parameters.

Model	Vocab Size	#Params
roBERTa PT	32000	68090880
AiBERTa	20000	101401344
BERTimbau$_{base}$	29794	108923136
RoBERTaLexPT$_{base}$	50265	124645632
RoBERTaCrawlPT$_{base}$	50265	124645632
PortBERT$_{base}$	52009	125985024
AlBERTina 100M PTPT	50265	138601728
AlBERTina 100M PTBR	50265	138601728
EuroBERT 210m	128256	211766016
XLM RoBERTa$_{base}$	250002	278043648
BERTimbau$_{large}$	29794	334396416
PortBERT$_{large}$	52009	357145600
XLM RoBERTa$_{large}$	250002	559890432
EuroBERT 610m	128256	607874688

4 Results

4.1 Downstream task evaluation

Table 3 presents the downstream evaluation results of all Portuguese language models across four ExtraGLUE tasks: STS-B, RTE, WNLI, and MRPC. We report task-specific metrics: Spearman and Pearson correlations for STS-B, accuracy for RTE and WNLI, and both accuracy and F1 for MRPC, alongside the average performance (AVG) across all tasks.

Among the base-sized models, RoBERTaLexPT$_{base}$ achieves the highest overall score with an AVG of 80.63, showing strong results particularly in MRPC accuracy (89.46) and F1 (92.34). Close behind is PortBERT$_{base}$, with an AVG of 80.57, outperforming all others in WNLI accuracy (60.56, tied with XLM-R) and ranking second in STS-B with a Spearman score of 87.39 and Pearson of 87.65. BERTimbau$_{base}$ shows the best performance in STS-B (88.5 mean), but underperforms slightly in WNLI, holding it back from overall top placement.

RoBERTaCrawlPT$_{base}$ and EuroBERT 210m also demonstrate robust overall performance, particularly in RTE and MRPC, with AVG scores above 79.0. Meanwhile, XLM RoBERTa$_{base}$ shows competitive results in WNLI (60.56) and MRPC F1 (91.32), though its STS-B score slightly lags behind the top contenders. Legacy models like roBERTa PT perform significantly worse, especially on semantic similarity tasks, confirming the impact of more recent training strategies and data sources.

In the large model category, XLM RoBERTa$_{large}$ emerges as the strongest overall model with an AVG of 84.01. It leads all others in STS-B (90.14 mean) and achieves the highest RTE score (82.31), although it underperforms in WNLI. EuroBERT 610m follows closely with an AVG of 83.44, showing outstanding performance in MRPC (94.2 F1, 91.91 accuracy) and the second-best RTE result (78.34).

PortBERT$_{large}$ achieves a solid overall score of 82.26, slightly ahead of BERTimbau$_{large}$ (82.23). While BERTimbau$_{large}$ does not dominate any single task, PortBERT$_{large}$ exhibits the highest WNLI accuracy (61.97). BERTimbau$_{large}$ stands out with strong STS-B scores (89.5 mean) and competitive MRPC metrics.

Overall, the results validate the effectiveness of the PortBERT models, with both the base and large variants frequently ranking among the top-performing models across tasks. The base model outperforms many existing Portuguese models on average, while the large model achieves results close to the best multilingual transformers. This indicates their robustness and applicability to a range of semantic and inference tasks in Portuguese.

4.2 Performance vs. Efficiency

To complement accuracy-based comparisons, we also assess model efficiency in terms of training and inference throughput (see Figure 1). Among the base models, several exhibit a favorable balance between performance and efficiency. Notably, roBERTa PT achieves the highest training throughput (62.1 samples/sec) and inference speed (112.7 samples/sec), but its task performance lags significantly behind all competitors, suggesting that efficiency alone is insufficient without adequate pretraining quality. In contrast, PortBERT$_{base}$ and RoBERTaCrawlPT$_{base}$ both demonstrate strong downstream performance (AVG: 80.57 and 80.48, respectively) while maintaining competitive training throughput around 25–26 samples/sec and inference throughput above 65 samples/sec. BERTimbau$_{base}$ similarly offers

Model	STS-B (Similarity)			RTE	WNLI	MRPC		AVG
	Spearman	Pearson	Mean	Acc	Acc	Acc	F1	
BERTimbau$_{large}$	89.4	89.61	89.5	75.45	59.15	88.24	91.55	82.23
EuroBERT 610m	88.46	88.59	88.52	78.34	59.15	**91.91**	**94.2**	83.44
XLM RoBERTa$_{large}$	**90.0**	**90.27**	**90.14**	**82.31**	57.75	90.44	93.31	**84.01**
PortBERT$_{large}$	88.53	88.68	88.6	72.56	**61.97**	89.46	92.39	82.26
AiBERTa	83.56	83.73	83.65	64.98	56.34	82.11	86.99	76.29
AlBERTina 100M PTBR	85.97	85.99	85.98	68.59	56.34	85.78	89.82	78.75
AlBERTina 100M PTPT	86.52	86.51	86.52	70.04	56.34	85.05	89.57	79.01
BERTimbau$_{base}$	**88.39**	**88.6**	**88.5**	70.4	56.34	87.25	90.97	80.32
EuroBERT 210m	86.54	86.62	86.58	65.7	57.75	87.25	91.0	79.14
RoBERTaCrawlPT$_{base}$	87.34	87.45	87.39	**72.56**	56.34	87.99	91.2	80.48
RoBERTaLexPT$_{base}$	86.68	86.86	86.77	69.31	59.15	**89.46**	**92.34**	**80.63**
XLM RoBERTa$_{base}$	85.75	86.09	85.92	68.23	**60.56**	87.75	91.32	79.95
PortBERT$_{base}$	87.39	87.65	87.52	68.95	**60.56**	87.75	91.13	80.57
roBERTa PT	48.06	48.51	48.29	56.68	59.15	72.06	81.79	61.04

Table 3: Evaluation results in %. STSB is reported with Spearman, Pearson, and their mean. RTE and WNLI are classification accuracy. MRPC includes accuracy and F1. The AVG score averages the six metrics: STSB Spearman, STSB Pearson, RTE Acc, WNLI Acc, MRPC Acc, MRPC F1. Bold = best, underlined = second-best per model size. Based on best epoch from 28 runs for max 10 epochs. The AVG score is computed as the unweighted mean across six metrics: STS-B Spearman, STS-B Pearson, RTE accuracy, WNLI accuracy, MRPC accuracy, and MRPC F1.

a good trade-off with strong performance (AVG: 80.32) and respectable throughput, making these three the most efficient base models when balancing quality and compute.

The large models generally exhibit higher downstream performance but at a considerable computational cost. XLM RoBERTa$_{large}$ leads in task performance (AVG: 84.01) and inference throughput (47.4 samples/sec) compared to its large-model peers. However, its training throughput is relatively low (14.9 samples/sec), indicating longer training durations. PortBERT$_{large}$ achieves an attractive efficiency-performance trade-off, with an AVG of 82.26 while maintaining higher training and inference throughput (23.3 and 70.7 samples/sec, respectively), positioning it as the most throughput-efficient large model while still achieving competitive accuracy. Meanwhile, EuroBERT-610M delivers strong performance (AVG: 83.44) but with lower throughput metrics, reflecting its high computational demands. These results suggest that while large models provide superior accuracy, the efficiency gap between well-optimized base and large models like PortBERT is narrowing. Full runtime statistics are reported in Appendix C.

5 Discussion

5.1 Efficiency and Accuracy Trade-offs

PortBERT demonstrates that efficient, monolingual transformer models remain a valuable asset in the evolving landscape of Portuguese NLP. While large multilingual encoders like XLM-RoBERTa or EuroBERT-610M offer strong performance, their high computational demands restrict practical deployment, particularly in latency-sensitive or resource-constrained settings. In contrast, PortBERT delivers competitive downstream task results while maintaining generally higher throughput compared to other strong Portuguese baselines, both during training and inference.

As shown in our efficiency analysis (Section 4.2), PortBERT$_{base}$ stands out for its balanced trade-off between accuracy and efficiency, ranking among the top performers in its class. PortBERT$_{large}$ narrows the performance gap to state-of-the-art models like XLM RoBERTa$_{large}$, while maintaining superior throughput and lower hardware demands. Our focus with PortBERT was on cost-efficient pretraining for Portuguese specifically, where zero-shot transfer is not required. In this sense, PortBERT complements large multilingual encoders such as XLM-RoBERTa by offering a more efficient option for monolingual applications.

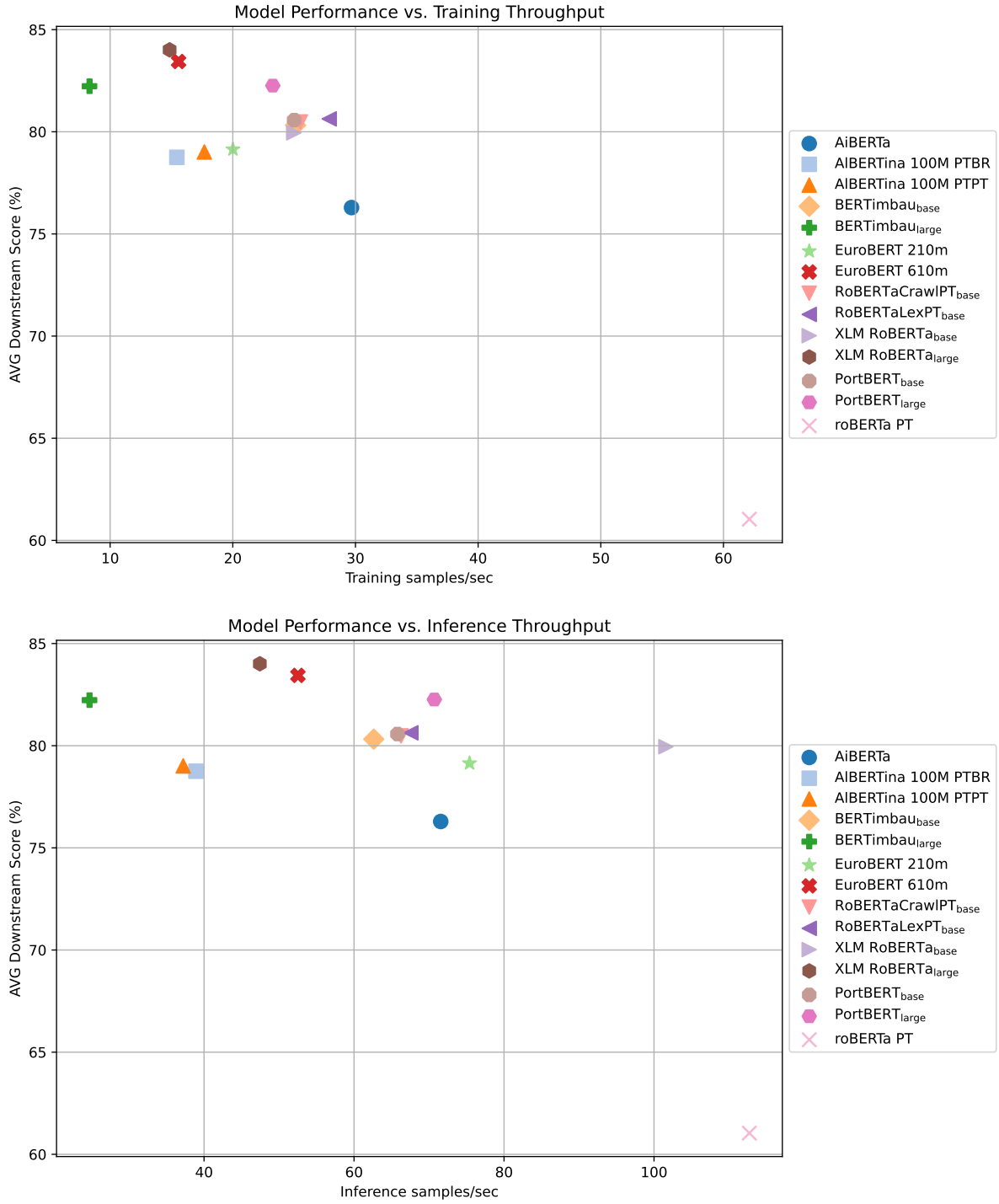

Figure 1: Performance–throughput trade-off across models. The top plot shows the relationship between average downstream score (AVG) and training throughput (samples/sec), while the bottom plot presents the same metric against inference throughput. This comparison highlights which models offer the best balance between effectiveness and computational efficiency during training and inference.

The performance differences between PortBERT and large multilingual encoders such as XLM-RoBERTa$_{large}$ are not solely attributable to the amount of training data. They also reflect architectural and training differences, including the substantially larger parameter count of XLM-RoBERTa (560M vs. 357M for PortBERT$_{large}$), its much larger multilingual vocabulary (250k vs. 52k tokens), and the use of a massive multilingual corpus (2.5TB multilingual vs. 456GB of Portuguese).

In addition to hardware throughput, PortBERT models also demonstrate strong parameter efficiency. PortBERT$_{\text{base}}$ (126M parameters) achieves higher average performance than larger models such as XLM-RoBERTa$_{\text{base}}$ (278M) and EuroBERT-210M (212M), despite having less than half their parameter count. PortBERT$_{\text{large}}$ (357M) achieves results close to XLM-RoBERTa$_{\text{large}}$ (560M) and EuroBERT-610M (608M), highlighting the impact of targeted, monolingual pretraining on recent Portuguese corpora. This makes PortBERT a compelling choice in scenarios where both accuracy and model size matter.

5.2 Training Setup and Hardware Comparisons

Beyond per-job throughput, the total pretraining time differed substantially between hardware setups. PortBERT$_{\text{base}}$, trained on 8 NVIDIA A40 GPUs, required approximately 27 days to complete 100k update steps. In contrast, PortBERT$_{\text{large}}$, trained on a TPUv4 128 pod, completed training in just over 6 days. Both models used the same batch size, corpus, and optimizer settings in full precision (fp32), allowing for a clean comparison of training performance across hardware platforms. Using GottBERT's pretraining durations as a reference, we estimate that PortBERT$_{\text{base}}$ would have taken around 1.3 days to train on comparable TPU infrastructure. This illustrates the advantage of modern TPUs for large-scale training, particularly when time is a critical factor. However, TPU-specific constraints, including limited memory flexibility and less mature tooling for PyTorch and custom workflows, can limit development. In addition, the lack of local TPU hardware forces developers to rely on cloud platforms, slowing iteration and complicating debugging.

Efficiency comparisons must also consider hardware configuration. Due to memory constraints, EuroBERT-610M and partly XLM RoBERTa$_{\text{large}}$ were trained without parallel jobs (i.e., one job per GPU), whereas PortBERT and other models used multiple parallel training jobs per GPU to maximize utilization. This difference in hardware allocation might have impacted the observed throughput and training durations, potentially skewing efficiency comparisons in this regard.

5.3 Positioning Among Existing Models

Recent large-scale efforts such as EuroBERT (Boizard et al., 2025) illustrate the scale-performance frontier in multilingual modeling. EuroBERT training consumed over 200,000 GPU hours across MI250X and MI300A clusters and leveraged cutting-edge optimization techniques such as FlashAttention (Dao, 2023). While such models raise the performance ceiling, they also require infrastructure that is out of reach for many academic or industry teams. In contrast, PortBERT was trained on commodity hardware using open-source tools, offering a transparent and efficient alternative that lowers the entry barrier for building high-quality models in any languages.

To our knowledge, PortBERT is the first RoBERTa-style Portuguese model trained on recent deduplicated and filtered corpora from CulturaX (mC4) and OSCAR23, using a fully transparent and reproducible fairseq pipeline. This positions it as a strong alternative to more resource-intensive systems, particularly for researchers and practitioners seeking open, efficient solutions.

Although decoder-only models such as GPT variants dominate general-purpose NLP, they are often unsuitable for sentence-level classification tasks due to their autoregressive nature. Encoder-based models like PortBERT offer lower inference latency and better fit for downstream classification, especially under real-world constraints.

5.4 Architectural Constraints and Training Stability

We deliberately retained the standard RoBERTa encoder architecture. Our goal was not only to establish a strong monolingual baseline, but also to enable a fair comparison of computational costs with GottBERT, which was trained on a comparable TPU setup. Introducing architectural modifications such as sparse or FlashAttention would have shifted the baseline and made this comparison meaningless.

Like GeistBERT (Scheible-Schmitt and Frei, 2025), PortBERT prioritizes practical usability over raw scale. Although it does not achieve top performance on every benchmark, it remains consistently strong across tasks, making it a compelling option in the accuracy-efficiency trade-off. PortBERT could also be adapted for longer inputs using architectures such as Longformer (Beltagy et al., 2020) or Nyströmformer (Xiong et al., 2021), though at the cost of increased training complexity.

During pretraining, we did not apply WWM, as stable support for it was missing in the fairseq

TPU implementation. As with GottBERT, we encountered TPU-specific constraints: the lack of dynamic memory allocation required processing the corpus as a continuous token stream, deviating from RoBERTa's dynamic sentence-sampling strategy. We were also constrained to 32-bit precision due to unstable 16-bit support in fairseq's TPU implementation, increasing memory use and runtime. To ensure stability under these conditions, we used conservative learning rates. For comparability, we deliberately applied the same pre-processing and training constraints to the GPU-based base model, even though the GPU setup would have supported dynamic sampling and mixed precision.

5.5 Final Remarks

Ultimately, PortBERT is a step toward sustainable and accessible language modeling for Portuguese. It illustrates that thoughtful model design, combined with optimized pretraining and recent corpora, can yield strong models without relying on large-scale infrastructure. Future work may explore quantized or distilled versions for mobile deployment and domain-specific continued pretraining to further expand applicability or even continue pretraining with a more diverse corpus using WWM similar to Scheible-Schmitt and Frei (2025).

6 Conclusion

We presented PortBERT, a family of RoBERTa-based language models for Portuguese, pre-trained on recent large-scale corpora (mC4 and OSCAR23). While not state-of-the-art on all benchmarks, PortBERT models achieve strong downstream performance and demonstrate notable efficiency in training and inference. To support reproducibility and downstream adoption, we release both Huggingface-compatible models and fairseq checkpoints. These resources enable further pretraining, fine-tuning, or adaptation for longer contexts and domain-specific tasks. PortBERT offers an efficient and accessible foundation for Portuguese NLP.

Acknowledgments

We gratefully acknowledge the support of Google's TensorFlow Research Cloud (TFRC) for providing access to Cloud TPUs, which enabled efficient pretraining of PortBERT$_{\text{large}}$. We also thank Nora Limbourg, our Google Cloud Customer Engineer, for her valuable technical assistance and coordination throughout the project.

R.S. would also like to thank Bruno & Suzi, as well as all members and friends of the Best Spot Azores Diving Center, including Alberto & Simona, Arturo, João & Claudia, Maëlle & Elias, Maria, Oliver, Paula, Raquel, Ruben, Sara and Vasco. Their kindness, presence, and community spirit provided strength and stability in a time of personal challenge. It is always a pleasure to dive with us.

Limitations

This work has several limitations. First, although we used deduplicated and filtered corpora from CulturaX (mC4 and OSCAR23), we did not apply deduplication across all possible data sources or levels of granularity. Residual duplication or noise may therefore remain in the training data.

Second, PortBERT was trained exclusively on web-based Portuguese text, without explicit control for dialectal variation (e.g., Brazilian vs. European Portuguese) or domain-specific content. As a result, the model's performance on underrepresented dialects or specialized registers (e.g., legal, medical, or informal language) may be suboptimal without further fine-tuning.

Third, while we aimed for stable and reproducible training configurations across both GPU and TPU platforms, we opted for conservative learning rates and default precision settings to ensure stability, particularly on TPUs where dynamic memory allocation and mixed precision remain limited in fairseq. We did not explore extensive hyperparameter tuning in regard of the peak learning rate and did not apply WWM, which could potentially yield further gains.

Fourth, we did not include a detailed error analysis of model predictions. While such an analysis could provide additional insights into systematic failure modes, our focus in this work was on efficiency and establishing strong baselines for Portuguese NLP.

Lastly, our evaluation is focused on the ExtraGLUE benchmark. While this provides a useful proxy for general NLP performance in Portuguese, it does not capture the full range of downstream tasks or real-world deployment settings. Moreover, ExtraGLUE does not offer a held-out test set with a submission server, which limits the ability to conduct blind evaluations and compare models in a

standardized manner.

Ethical Considerations

As with any large-scale language model, PortBERT is susceptible to inheriting and reproducing biases present in its training data. While we apply deduplication techniques to reduce noise and redundancy, deeper societal, cultural, and representational biases may persist. This is particularly relevant for downstream applications in sensitive domains such as healthcare, education, or public administration, where biased outputs could reinforce inequality or misinformation.

Training on large-scale web-based corpora also introduces privacy concerns. Although the dataset is filtered and preprocessed, models may inadvertently memorize and surface sensitive or personal information. Careful handling is necessary when deploying PortBERT in real-world applications, especially those involving user data or decision-making contexts.

Finally, despite efforts to balance performance and efficiency, pretraining transformer models on GPUs and TPUs consumes substantial computational resources. The associated energy usage and environmental impact underline the importance of developing sustainable training practices and promoting model reuse.

References

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen Marcus McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2024. Llemma: An open language model for mathematics. In *The Twelfth International Conference on Learning Representations*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M. Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, João Alves, Kevin El-Haddad, Manuel Faysse, Maxime Peyrard, Nuno M. Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. 2025. Eurobert: Scaling multilingual encoders for european languages.

Branden Chan. 2020. XLM-RoBERTa: The multilingual alternative for non-english NLP. Library Catalog: towardsdatascience.com.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. *arXiv:2001.06286 [cs]*. ArXiv: 2001.06286.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023. Zero-shot cross-lingual transfer language selection using linguistic similarity. *Information Processing & Management*, 60(3):103250.

Eduardo A. S. Garcia, Nadia F. F. Silva, Felipe Siqueira, Hidelberg O. Albuquerque, Juliana R. S. Gomes, Ellen Souza, and Eliomar A. Lima. 2024. RoBERTaLexPT: A legal RoBERTa model pretrained with deduplication for Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 374–383, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Tim Jansen, Yangling Tong, Victoria Zevallos, and Pedro Ortiz Suarez. 2022. Perplexed by quality: A perplexity-based method for adult and harmful content detection in multilingual heterogeneous web data.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Norman P. Jouppi, George Kurian, Sheng Li, Peter Ma, Rahul Nagarajan, Lifeng Nai, Nishant Patil, Suvinay Subramanian, Andy Swing, Brian Towles, Cliff Young, Xiang Zhou, Zongwei Zhou, and David Patterson. 2023. TPU v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut.

2020. Albert: A lite bert for self-supervised learning of language representations.

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

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, Tianyang Liu, Max Tian, Denis Kocetkov, Arthur Zucker, Younes Belkada, Zijian Wang, Qian Liu, Dmitry Abulkhanov, Indraneil Paul, Zhuang Li, Wen-Ding Li, Megan Risdal, Jia Li, Jian Zhu, Terry Yue Zhuo, Evgenii Zheltonozhskii, Nii Osae Osae Dade, Wenhao Yu, Lucas Krauß, Naman Jain, Yixuan Su, Xuanli He, Manan Dey, Edoardo Abati, Yekun Chai, Niklas Muennighoff, Xiangru Tang, Muhtasham Oblokulov, Christopher Akiki, Marc Marone, Chenghao Mou, Mayank Mishra, Alex Gu, Binyuan Hui, Tri Dao, Armel Zebaze, Olivier Dehaene, Nicolas Patry, Canwen Xu, Julian McAuley, Han Hu, Torsten Scholak, Sebastien Paquet, Jennifer Robinson, Carolyn Jane Anderson, Nicolas Chapados, Mostofa Patwary, Nima Tajbakhsh, Yacine Jernite, Carlos Muñoz Ferrandis, Lingming Zhang, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2024. Starcoder 2 and the stack v2: The next generation.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe.

Microsoft. 2025. Neural network intelligence. `https://github.com/microsoft/nni/`. Accessed: 2025-05-01.

Nuno Miquelina, Paulo Quaresma, and Vítor Beires Nogueira. 2022. Generating a european portuguese bert based model using content from arquivo.pt archive. In *Intelligent Data Engineering and Automated Learning – IDEAL 2022*, pages 280–288, Cham. Springer International Publishing.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. *Advancing Neural Encoding of Portuguese with Transformer Albertina PT-**, page 441–453. Springer Nature Switzerland.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Daniel Santos, Nuno Miquelina, Daniela Schmidt, Paulo Quaresma, and Vítor Beires Nogueira. 2025a. Performance evaluation of nlp models for european portuguese: Multi-gpu/multi-node configurations and optimization techniques. In *Algorithms and Architectures for Parallel Processing*, pages 298–314, Singapore. Springer Nature Singapore.

Daniel Santos, Nuno Miquelina, Daniela Schmidt, Paulo Quaresma, and Vítor Beires Nogueira. 2025b. Performance Evaluation of NLP Models for European Portuguese: Multi-GPU/Multi-node Configurations and Optimization Techniques. In *Algorithms and Architectures for Parallel Processing*, pages 298–314, Singapore. Springer Nature.

Rodrigo Santos, João Rodrigues, António Branco, and Rui Vaz. 2021. Neural text categorization with transformers for learning portuguese as a second language. In *Progress in Artificial Intelligence*, pages 715–726, Cham. Springer International Publishing.

Raphael Scheible, Johann Frei, Fabian Thomczyk, Henry He, Patric Tippmann, Jochen Knaus, Victor Jaravine, Frank Kramer, and Martin Boeker. 2024. GottBERT: a pure German language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21237–21250, Miami, Florida, USA. Association for Computational Linguistics.

Raphael Scheible-Schmitt and Johann Frei. 2025. Geistbert: Breathing life into german nlp.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Jorge A Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brwac corpus: A new open resource for brazilian portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. 2021. Nyströmformer: A nyström-based algorithm for approximating self-attention.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Parameters

The parameter space for our grid search is listed in Table 4. In addition, Table 5 shows the parameters of the best models (selection based on validation set) of the respective tasks. We include these details to support reproducibility of our downstream results.

Parameter	Values
Learning Rate	7e-5, 5e-5, 2e-5, 1e-5, 7e-6, 5e-6, 1e-6
Batch Size	16, 32, 48, 64
Epochs	10

Table 4: Hyperparameters used in the grid search of the downstream tasks.

B Perplexity

During pretraining, model perplexity was tracked on a test set after each optimization step and on a validation set at every checkpoint (see Figure 2). The models exhibited a plateau in their perplexity curves, brief for the base models, but more prolonged for the large ones. Some training curves also showed temporary spikes, which may appear as divergence if not interpreted with context. Across both models, convergence occurred gradually and stabilized by around 30k steps. In contrast, the validation perplexity decreased steadily across both models without showing pronounced plateaus, stabilizing at lower values by the end of training. This results from the limited number of validation checkpoints (three intermediate epochs and a final checkpoint at 100k steps), which yield a coarser view of the learning dynamics.

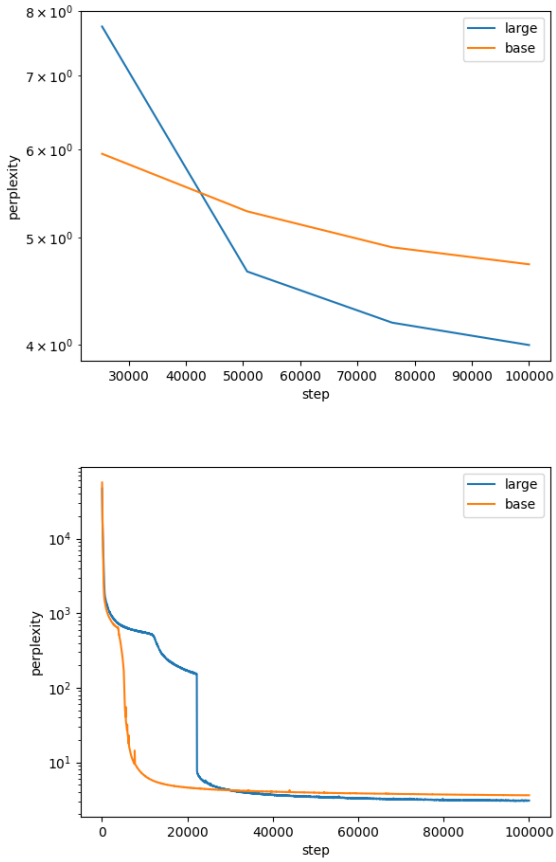

Figure 2: Perplexity of the PortBERT models. Top based on a validation at the checkpoints. Bottom based on the validation of each optimization cycle during the training.

C Efficiency Measurements

Tables 6 and 7 report detailed runtime statistics for all models and tasks. Table 6 provides a task-level breakdown of training and inference times, while Table 7 compares model-level efficiency metrics, including throughput and per-epoch timing. All models were fine-tuned using Huggingface Transformers (v4.52.3) on NVIDIA A40 GPUs.

Task	Training Time	Inference Time
MRPC	157:04	00:38
RTE	241:46	00:57
STSB	314:24	02:25
WNLI	25:30	00:08

Table 6: Computation time in hours and minutes for the downstream tasks, summing up to 1549 hours and 29 minutes, which corresponds to approximately 64.6 days of GPU usage.

Model	STS-B		RTE		WNLI		MRPC	
	BS	LR	BS	LR	BS	LR	BS	LR
EuroBERT 210m	16	7 E-06	64	2 E-05	32	2 E-05	32	7 E-06
XLM RoBERTa$_{\text{large}}$	64	2 E-05	32	1 E-05	16	7 E-05	64	2 E-05
AlBERTina 100M PTPT	64	7 E-05	32	2 E-05	16	2 E-05	48	5 E-05
AlBERTina 100M PTBR	64	5 E-05	16	1 E-05	32	1 E-06	48	7 E-05
AiBERTa	32	2 E-05	32	1 E-05	32	7 E-05	32	5 E-05
EuroBERT 610m	16	1 E-05	16	7 E-06	32	1 E-05	16	5 E-06
XLM RoBERTa$_{\text{base}}$	16	1 E-05	32	2 E-05	64	2 E-05	16	2 E-05
roBERTa PT	32	7 E-05	32	1 E-05	48	5 E-06	32	7 E-05
RoBERTaCrawlPT$_{\text{base}}$	48	7 E-05	64	7 E-05	48	1 E-06	48	7 E-05
BERTimbau$_{\text{large}}$	32	2 E-05	16	7 E-05	32	7 E-06	16	5 E-05
BERTimbau$_{\text{base}}$	48	5 E-05	16	1 E-05	48	7 E-05	32	7 E-05
PortBERT$_{\text{base}}$	48	7 E-05	16	1 E-05	16	1 E-06	64	1 E-05
RoBERTaLexPT$_{\text{base}}$	48	5 E-05	48	5 E-05	32	7 E-06	64	2 E-05
PortBERT$_{\text{large}}$	16	2 E-05	16	7 E-06	32	7 E-06	16	7 E-06

Table 5: Hyperparameters of the best downstream task models for each task and pre-trained model. BS refers to batch size, and LR denotes the learning rate.

Model	Train Time (s)	Train/s	Time/Epoch (s)	Eval Time (s)	Eval/s
AiBERTa$_{\text{2000M}}$	1306.47	29.68	142.39	7.24	71.56
AlBERTina$_{\text{PTBR}}$	2906.82	15.44	309.08	15.85	38.92
AlBERTina$_{\text{PTPT}}$	2800.95	17.68	300.35	17.54	37.22
BERTimbau$_{\text{base}}$	1499.94	25.12	152.88	9.15	62.63
BERTimbau$_{\text{large}}$	4406.49	8.32	484.90	21.44	24.73
EuroBERT$_{\text{210M}}$	1777.84	20.01	181.90	6.60	75.40
EuroBERT$_{\text{610M}}$	2498.58	15.58	254.26	12.80	52.52
RoBERTaCrawlPT$_{\text{base}}$	1682.64	25.51	171.76	9.15	66.28
RoBERTaLexPT$_{\text{base}}$	1457.99	27.86	149.42	8.84	67.58
XLM-RoBERTa$_{\text{base}}$	1440.55	24.97	152.49	4.86	101.59
XLM-RoBERTa$_{\text{large}}$	2139.34	14.85	233.65	10.46	47.44
PortBERT$_{\text{base}}$	1524.59	25.00	160.29	8.96	65.79
PortBERT$_{\text{large}}$	2389.63	23.26	264.74	15.09	70.71
roBERTa PT	635.46	62.11	79.02	4.82	112.71

Table 7: Training and inference efficiency of all evaluated models. Metrics include total training time, training samples per second, average time per epoch, total evaluation time, and evaluation throughput.