Global vs. Local Sentence Embeddings for Brazilian Portuguese: Revisiting Monolingual Models in the Age of Foundation Models

Matheus Peixoto and Guilherme Silva
Postgraduate Program in Computer Science, Federal University of Ouro Preto
35.400-000 – Ouro Preto – MG – Brazil
{matheus.peixoto,guilherme.lopes}@aluno.ufop.edu.br

Giacomo Figueredo
Department of Language Studies, Federal University of Ouro Preto
35.420-000 – Mariana – MG – Brazil
giacomo.figueredo@ufop.edu.br

Pedro Silva and Eduardo J. S. Luz
Department of Computing, Federal University of Ouro Preto
35.400-000 – Ouro Preto – MG – Brazil
{silvap,eduluz}@ufop.edu.br

Abstract

The choice between large-scale, multilingual, foundation models and specialized monolingual models for languages like Brazilian Portuguese (PT-BR) presents a complex trade-off between generalization and specialization. Throughout this paper, we use *global* as shorthand for *multilingual* foundation models and *local* for *Portuguese-specialized* models. This paper investigates this trade-off through an empirical study across a diverse suite of tasks. We evaluate multiple families of language models under both linear probing and fine-tuning regimes. We find that monolingual encoders exhibit greater "adaptation plasticity" during fine-tuning, improving both classification and semantic similarity, whereas multilingual (global) models often lose semantic similarity structure after task-specific adaptation. In this work, we measure this effect as an STS drop relative to the frozen baseline, i.e., adaptation-induced representation drift. However, this plasticity comes at a cost: our tokenization analysis suggests that monolingual models struggle with foreign terms, whereas modern multilingual tokenizers show surprising morphological competence, challenging the common assumption that language-specific vocabularies are *inherently* superior in all settings. We conclude that the optimal model choice is a task-dependent trade-off between vocabulary coverage and adaptation flexibility.

1 Introduction

The advent of large-scale foundation models has transformed the landscape of Natural Language Processing. Models like Gemma (Team et al., 2024) and Qwen (Yang et al., 2024) are trained on web-scale multilingual data, demonstrating impressive zero-shot transfer and linear-probing capabilities across a wide range of tasks and languages, including medium-resource languages like Brazilian Portuguese (PT-BR). In this work, *global/local* refers to the *pretraining scope* (multilingual vs. Portuguese-specialized).

The prevailing approach to specialize these models is lightweight adaptation, using Parameter-Efficient Fine-Tuning (PEFT) methods like Low-Rank Adaptation (LoRA) (Hu et al., 2022) on local corpora. This practice, however, raises a critical yet under-explored question: *how effective is this adaptation, and does it hold universally across different downstream tasks*? Specifically for Brazilian Portuguese (PT-BR), established monolingual models like BERTimbau (Souza et al., 2020), though based on older architectures, were trained exclusively on Brazilian corpora. Their success solidified a long-standing assumption within the NLP community for Portuguese: that specialized, monolingual vocabularies are inherently superior for achieving state-of-the-art performance (Souza et al., 2020; Virtanen et al., 2019; Rust et al., 2021). However, the massive scale and architectural advances of modern foundation models compel us to challenge this assumption. It is an open question whether a lightly adapted global model can now outperform a deeply specialized local one.

This paper addresses this gap by conducting a comparative study. We investigate the performance of four distinct families of sentence embeddings on a comprehensive set of PT-BR tasks: Classification, Clustering, Natural Language Inference (NLI), and Semantic Textual Similarity (STS). Our

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 529–539

analysis operates under two key regimes: (i) linear probing for supervised tasks (Classification/NLI) and intrinsic similarity evaluation for STS, to measure frozen out-of-the-box representations; and (ii) a lightweight adaptation scenario to measure the impact of supervised fine-tuning. Our goal is to understand the trade-offs between global, generalist models and local, specialist ones.

We structure our investigation around the following research questions: **RQ1:** *How do monolingual and multilingual (global) embedding models compare in PT-BR under linear-probing and lightweight fine-tuning regimes across diverse tasks*? **RQ2:** *Does the adaptation of the models to their respective tasks always yield ~~to~~ significant improvements, or are there performance trade-offs between different tasks?* **RQ3:** *Which family of embeddings offers the best trade-off between performance and robustness across the evaluated PT-BR tasks?*

Our contributions are fourfold. We first present a comprehensive benchmark of diverse sentence embedding families on a suite of PT-BR tasks. Through this benchmark, we identify an "adaptation dichotomy": fine-tuning boosts classification performance but significantly degrades semantic similarity capabilities in global (multilingual) models (Figures 1a and 1b). In contrast, we provide evidence that specialized monolingual models exhibit greater "plasticity", remaining highly competitive in realistic adaptation scenarios. Finally, to explain these findings, we offer a tokenization analysis that challenges the prevailing assumption of monolingual superiority, revealing that modern multilingual tokenizers possess surprising morphological awareness, even as all models struggle with informal language. We also explicitly clarify that STS is evaluated intrinsically on frozen embeddings and that STS "forgetting" refers to post-adaptation degradation relative to that frozen baseline.

2 Related Work

Wu and Dredze (2020) analyze the representation quality of the 104 languages covered by multilingual BERT (mBERT) and compare it against monolingual and bilingual baselines. Their results reveal that mBERT does not perform uniformly across languages: it struggles significantly in languages with limited pretraining data, while languages with more data available are better handled by specialized monolingual models. Interestingly, mBERT achieves its best relative performance in medium-resource languages, where data scarcity is reduced, but monolingual models no longer offer a clear advantage.

Seeking to address the underrepresentation of Finnish in mBERT, Virtanen et al. (2019) trained a BERT-base model from scratch and reported substantial improvements across all evaluated tasks compared to mBERT. A similar limitation is observed for Brazilian Portuguese, for which Souza et al. (2020) introduced BERTimbau, leveraging the brWaC, the largest publicly available Portuguese corpus. Their monolingual model consistently outperformed mBERT, further fueling the discussion on the trade-offs between monolingual and multilingual architectures.

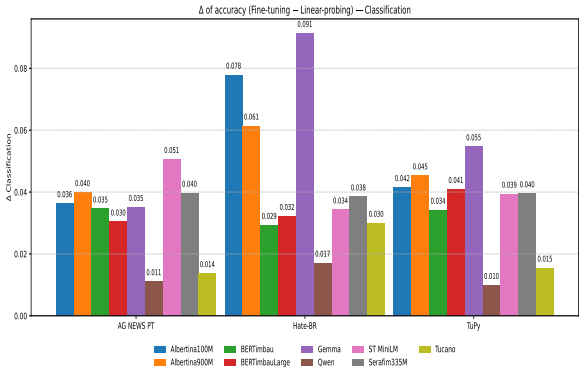
Expanding this perspective, Rust et al. (2021) show that the performance gap between monolingual and multilingual BERT models is driven less by multilingualism itself and more by factors such as pretraining data size and the quality of the tokenizer. They demonstrate that language-specific tokenizers can yield substantial downstream improvements.

Tokenization disparities are further highlighted by Petrov et al. (2023), who observe that identical content in Italian can generate up to 1.6× more tokens than in English when processed by multilingual tokenizers, a discrepancy that worsens as the language's global popularity decreases.

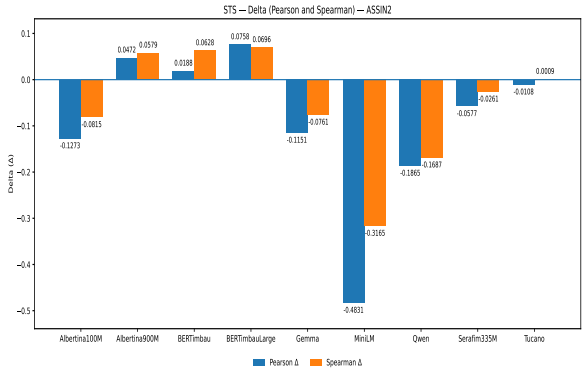
Finally, recent work has examined how adaptation procedures affect model stability. Chen et al. (2024) shows that, under instruction-tuning, multilingual models tend to lose linguistic consistency, whereas monolingual models maintain significantly greater robustness and stability throughout the adaptation process. We note that there also exist multilingual models explicitly trained for embedding quality (e.g., multilingual E5/BGE-style models); we discuss this as a limitation since our benchmark focuses on widely used general-purpose foundation models and Portuguese-specialized baselines.

2.1 Multilingual and Foundation Embeddings

Very common models are the multilingual ones like mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), which demonstrated that a single model could learn effective representations for many languages. Sentence embedding models like LaBSE (Feng et al., 2022) and the 'paraphrase-multilingual-MiniLM-L12-v2'



(a) Accuracy gain from lightweight fine-tuning over linear probing (Classification).



(b) Change in STS Spearman correlation after lightweight task adaptation, relative to the frozen baseline.

Figure 1: The core finding of our work: the adaptation dichotomy. Lightweight fine-tuning consistently improves classification (Figure 1a), but often harms STS for multilingual (global) models (Figure 1b). We interpret this as adaptation-induced representation drift (measured against the frozen baseline), while acknowledging that encoder-only and decoder-only architectures may differ in how sentence embeddings are best extracted (see Limitations).

(Reimers and Gurevych, 2019) built on this, optimizing for cross-lingual sentence similarity. More recently, foundation models like Gemma (Team et al., 2024) and Qwen (Yang et al., 2024) represent the current state-of-the-art (SOTA), trained on unprecedented scales of data. These models are the primary subject of our "global" category.

2.2 Monolingual Models for Portuguese

On the other end of the spectrum, researchers focused on specialized models for particular languages, leveraging the development of monolingual models for non-English languages, such as Portuguese, which has been crucial for achieving high performance on language-specific tasks. Early work focused on word embeddings like Word2Vec trained on local corpora. The Transformer era brought models like BERTimbau (Souza et al., 2020), a BERT model pre-trained from scratch on a large PT-BR corpus (brWaC). These models have consistently been strong baselines, demonstrating the value of deep specialization. Our work contrasts this specialist approach with modern generalist models.

Following BERTimbau's specialization, other models have emerged, each one with its main characteristics and backgrounds. Albertina (Santos et al., 2024a) is an encoder-only family based on DeBERTa, trained with a focus on different Portuguese variants; Serafim (Gomes et al., 2024) is a family of sentence encoders trained to produce Portuguese specialized embeddings for classification, clustering, and information retrieval; and, recently, Tucano (Corrêa et al., 2025), a family of Portuguese-focused decoder-only models trained on the large GigaVerbo corpus, aiming for strong Portuguese text generation. Text generation tasks were also adapted to PT-BR with Sabiá (Pires et al., 2023), Canarin (Maicon Domingues, 2023) and BODE (Garcia et al., 2024). We also acknowledge additional recent PT-oriented LLMs (e.g., Cabrita (Larcher et al., 2023), GlórIA (Lopes et al., 2024) and Gervásio (Santos et al., 2024b)), which we did not include due to scope and reproducibility constraints, and which we discuss briefly as a limitation.

2.3 Fine-tuning and Catastrophic Forgetting

The process of adapting pre-trained models is not without its challenges. Fine-tuning on a specific task can lead to "catastrophic forgetting", where the model loses some of its general capabilities (Kirkpatrick et al., 2017). In this paper, we use "catastrophic forgetting" *operationally* to denote *adaptation-induced representation drift*: a degradation in held-out STS correlation *after* supervised adaptation, measured *relative to the frozen baseline* (intrinsic similarity evaluation). This is related to stability/geometry changes under fine-tuning rather than a continual-learning setup. This connects to research on the geometry of embedding spaces and how it is altered by task-specific adaptation (Mosbach et al., 2020).

3 Experimental Setup

3.1 Models

We selected seven representative models to cover the spectrum from local specialists to global generalists, spanning both encoder-only and decoder-only architectures. Unless stated otherwise, we compute sentence embeddings by *mean pooling* the last-layer token representations using the attention mask (padding excluded). For decoder-only models, we do *not* generate text; we take the final hidden states for the input sequence and apply the same pooling.

The models are: (i) **Albertina PT-BR**: A BERT-based model developed over the DeBERTa model. Both sizes were explored, the 100M and the 900M parameters (‘PORTULAN/albertina-100m-portuguese-ptbr-encoder‘ and ‘PORTULAN/albertina-900m-portuguese-ptbr-encoder‘); (ii) **BERTimbau**: A BERT-base model pre-trained on PT-BR data. We tested both base and large BERTimbau (‘neuralmind/bert-base-portuguese-cased’ and ‘neuralmind/bert-large-portuguese-cased’); (iii) **Gemma**: A modern foundation model from Google (‘google/gemma-2-2b-it’) ; and (iv) **Qwen**: A strong multilingual embedding model from Alibaba (‘qwen/qwen3-rmbedding-0.6b’). (v) **Serafim**: An encoder focused on clustering and semantic search (‘PORTULAN/serafim-335m-portuguese-pt-sentence-encoder‘); (vi) **MiniLM**: A compact multilingual sentence transformer (‘sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2’); (vii) **Tucano**: A decoder transformer pretrained on the Portuguese dataset GigaVerbo (‘TucanoBR/Tucano-630m’).

The key properties of the seven models are summarized in Table 1.

Model Family	Architecture	Vocab Size	Parameters
Albertina PT-BR	Encoder	128k	139M 887M
BERTimbau	Encoder	29k	110M 335M
Gemma	Decoder	256k	2B
Qwen	Decoder	152k	596M
Serafim	Encoder	29k	335M
ST MiniLM	Encoder	250k	118M
Tucano	Decoder	32k	650M

Table 1: Overview of the models used in our study.

3.2 Tasks and Datasets

To ensure a comprehensive evaluation, we selected four tasks spanning different NLP capabilities, summarized in Table 2.

Task	Metric	Dataset	Test Size
Classification	Accuracy	AG News PT	7600
		TuPyE	8734
		HateBR	1400
Clustering	AMI	AG News PT	7600
		TuPyE	8734
		HateBR	1400
NLI	Accuracy	Assin2	2448
STS	Pearson	ASSIN	4000
	Spearman	Assin2	2448

Table 2: Explored tasks, metrics and datasets

The evaluation comprised four tasks: classification, clustering, NLI, and STS. For the classification and clustering tasks, we considered three datasets: (1) AG News PT, (2) TuPyE, and (3) HateBR. AG News PT is a Portuguese translation of the original AG News dataset (Maritaca-AI, 2023), which contains more than one million English news articles collected over the course of one year from the ComeToMyHead platform. The dataset is organized into four topical categories: (1) World, (2) Sports, (3) Business, and (4) Sci/Tech.

TuPyE (Oliveira et al., 2023) is a corpus created through the integration of three datasets and represents one of the largest annotated Portuguese resources for hate speech detection. It comprises 43,668 anonymized documents collected from social media and is annotated with three classes: (1) Non-aggressive, (2) Aggressive without hate, and (3) Aggressive with hate. Finally, HateBR (Vargas et al., 2022) is another hate speech corpus in PT-BR, consisting of approximately 7,000 samples evenly distributed between the hate speech and non-hate (normal) categories.

It is important to note that the HateBR dataset does not include an official train–test split. To address this, we applied a stratified random partition, allocating 20% of the data for testing and the remaining 80% for training. A fixed random seed was used to ensure reproducibility. Additionally, the TuPyE dataset provides two binary annotation columns, namely “aggressive” and “hate”. We convert these two binary labels into a single 3-way label via an explicit mapping: (aggressive=0, hate=0)→ *normal* (non-aggressive), (1,0)→ *aggressive (non-hate)*, and (1,1)→ *hate (aggressive-*

with-hate). If any (0,1) cases occur, we map them to *hate* for consistency.

For the NLI and STS tasks, we employed the ASSIN2 dataset, while the original ASSIN dataset was used exclusively for STS. The ASSIN corpus (Fonseca et al., 2016) consists of more than 10,000 sentence pairs extracted from Google News in Portugal and Brazil (not only PT-BR). ASSIN2 (Real et al., 2020), in turn, follows a similar construction methodology but contains over 10,000 sentence pairs exclusively in PT-BR. It includes human annotations for both NLI and STS.

3.3 Evaluation Regimes

Frozen Baseline (Linear Probing and Intrinsic STS). For classification and NLI, we train a simple logistic regression classifier on top of the frozen embeddings and report accuracy values ranging from 0 (all predictions incorrect) to 1 (all predictions correct), with higher values indicating better performance. In the case of standard classification, the model performs direct label prediction. For NLI, however, the logistic regression classifier is applied to pairwise embedding features derived from each premise–hypothesis pair.

For clustering, we apply the KMeans algorithm to the embeddings produced by the model and evaluate performance using the AMI (Adjusted Mutual Information) metric. AMI quantifies the similarity between two clustering assignments while correcting for agreement that may arise by chance. Its values range from -1 to 1, where 1 denotes perfect correspondence, 0 indicates similarity no better than random expectation, and negative values reflect agreement worse than would be expected by chance. Higher values indicate better performance.

Finally, for STS, we compute a similarity score directly between the embeddings. In this study, we employ both Pearson and Spearman correlation coefficients. This evaluation setup assesses the intrinsic quality of the pre-trained representation space. Higher values indicate better performance. Unless otherwise stated, all evaluations use a maximum sequence length of 512 and batch size 4 (limited by Gemma).

Lightweight Adaptation. We employ LoRA for all models in each supervised task (Classification, NLI, STS). The model used for clustering is the same as that trained for Classification. We use a consistent set of hyperparameters across all models to ensure a fair comparison. The models were trained for 10 epochs with a learning rate of $3e^{-5}$, a weight decay of 0.01, a warm-up ratio of 0.06, and using *fp16*. We use LoRA with rank r=8, α=16, and dropout 0.05. The target modules vary according to model architecture (following each checkpoint's HuggingFace naming). For encoder-style models (BERT/DeBERTa-like), we adapt attention projections and the output dense layer (e.g., `query`, `key`, `value`, `dense`). For decoder-style LLM blocks (Gemma/Qwen/Tucano-like), we adapt attention and MLP projections (e.g., `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`). In all results tables, Δ denotes the difference between the adapted model and its corresponding frozen baseline for the same task/metric.

4 Results and Discussion

The experiments were conducted on a machine equipped with 128 GB of DDR4 RAM, an NVIDIA RTX 3090 GPU, and an Intel Core i9-10900 processor. The implementation was carried out using the Python programming language. The Transformers library (4.52.4) was employed for model training alongside PyTorch (2.7.0). The lightweight adaptation was made using the PEFT library (0.18.0). The metrics were computed using SciKit-Learn. [1]

4.1 Results and the Research Questions

Our analyses are guided by the research questions outlined in this study.

RQ1: Linear-probing vs. Fine-tuning Trade-offs. Across all PT-BR tasks, monolingual models (Albertina, BERTimbau, Serafim, Tucano) consistently achieve stronger and more stable performance than global multilingual models (Gemma, Qwen, MiniLM). Although multilingual models are often competitive under linear-probing settings, they exhibit reduced reliability under lightweight fine-tuning, particularly for semantic similarity tasks.

In classification tasks (Table 3), monolingual models achieve slightly higher average accuracy under linear probing (0.847 vs. 0.834), with BERTimbauLarge performing best overall. Fine-tuning improves results for all models, but monolingual approaches still retain their advantage, with Al-

[1]For transparency and reproducibility purposes, the code used in the experiments is publicly available in https://anonymous.4open.science/r/2026-Propor-EmbeddingGemma-C0B6/src2/dataloader.py.

bertina900M leading on AG News PT (0.9366) and Serafim335M on HateBR (0.9364).

Dataset	Model	Linear-Probing	Fine-tuning	Δ
AG NEWS PT	Albertina100M	0.872	0.908	0.036
	Albertina900M	0.897	**0.937**	0.040
	BERTimbau	0.887	0.922	0.035
	BERTimbauLarge	**0.898**	0.929	0.030
	Gemma	0.887	0.922	0.035
	MiniLM	0.871	0.921	**0.051**
	Qwen	0.898	0.909	0.011
	Serafim335M	0.887	0.927	0.040
	Tucano	0.892	0.906	0.014
Hate-BR	Albertina100M	0.802	0.880	0.078
	Albertina900M	0.848	0.909	0.061
	BERTimbau	0.900	0.929	0.029
	BERTimbauLarge	**0.902**	0.934	0.032
	Gemma	0.824	0.916	**0.091**
	MiniLM	0.844	0.879	0.034
	Qwen	0.873	0.890	0.017
	Serafim335M	0.898	**0.936**	0.039
	Tucano	0.878	0.908	0.030
TuPy	Albertina100M	0.761	0.802	0.042
	Albertina900M	0.774	0.819	0.045
	BERTimbau	0.788	0.823	0.034
	BERTimbauLarge	**0.789**	**0.830**	0.041
	Gemma	0.763	0.818	**0.055**
	MiniLM	0.761	0.800	0.039
	Qwen	0.781	0.791	0.010
	Serafim335M	0.788	0.828	0.040
	Tucano	0.784	0.800	0.016

Table 3: Classification results (Accuracy) under linear-probing and fine-tuning regimes. Best scores per dataset/setting are in bold.

For clustering (Table 4), Serafim335M dominates on HateBR and TuPy, whereas Qwen achieves the best result on AG News PT, followed by BERTimbauLarge, while Serafim335M ranks eighth with considerably lower results. Although fine-tuning benefits all models, monolingual models exhibit larger gains (Δ 0.303 vs. 0.256) and higher average performance (0.498 vs. 0.430).

In the NLI task (Table 5), Serafim again stands out, achieving an accuracy of 0.839 and maintaining its dominance after fine-tuning. Although most models improve under tuning, Qwen experiences a slight performance degradation, reflected in a negative Δ of 0.011.

Finally, STS (Table 6) reveals the clearest contrast between model families. Serafim335M leads under the *frozen* regime on ASSIN and ASSIN2, but its STS correlations drop after fine-tuning (relative to the frozen baseline), consistent with adaptation-induced representation drift. In contrast, Albertina900M, BERTimbau, and BERTimbauLarge consistently improve across all metrics.

In summary, monolingual embedding models outperform global multilingual models on average under both linear-probing and lightweight fine-tuning regimes, while also exhibiting greater stability. Although global multilingual models remain competitive, fine-tuning consistently provides greater benefits to monolingual models.

Dataset	Model	Linear-Probing	Fine-tuning	Δ
AG NEWS PT	Albertina100M	0.478	0.721	0.243
	Albertina900M	0.514	**0.802**	0.289
	BERTimbau	0.521	0.733	0.212
	BERTimbauLarge	0.547	0.754	0.207
	Gemma	0.277	0.712	**0.435**
	MiniLM	0.526	0.753	0.226
	Qwen	**0.588**	0.716	0.128
	Serafim335M	0.446	0.752	0.306
	Tucano	0.542	0.697	0.156
Hate-BR	Albertina100M	0.016	0.450	0.435
	Albertina900M	0.022	0.559	0.536
	BERTimbau	0.005	0.636	0.631
	BERTimbauLarge	0.000	0.654	**0.654**
	Gemma	0.004	0.430	0.426
	MiniLM	0.152	0.457	0.305
	Qwen	0.007	0.448	0.441
	Serafim335M	**0.318**	**0.662**	0.343
	Tucano	0.020	0.426	0.406
TuPy	Albertina100M	0.002	0.148	0.147
	Albertina900M	0.002	**0.281**	**0.279**
	BERTimbau	0.012	0.177	0.164
	BERTimbauLarge	0.008	0.207	0.199
	Gemma	0.001	0.066	0.065
	MiniLM	0.007	0.177	0.171
	Qwen	0.007	0.112	0.105
	Serafim335M	**0.045**	0.191	0.146
	Tucano	0.002	0.105	0.103

Table 4: Clustering results (AMI) under linear-probing and fine-tuning regimes. Best scores per dataset/setting are in bold.

Dataset	Model	Linear-Probing	Fine-tuning	Δ
ASSIN2	Albertina100M	0.763	0.785	0.023
	Albertina900M	0.773	0.805	0.032
	BERTimbau	0.808	0.829	0.021
	BERTimbauLarge	0.772	0.800	0.029
	Gemma	0.724	0.732	0.008
	MiniLM	0.790	0.822	**0.032**
	Qwen	0.815	0.804	-0.011
	Serafim335M	**0.839**	**0.855**	0.015
	Tucano	0.769	0.779	0.011

Table 5: NLI results (Accuracy) on ASSIN2 under linear-probing and fine-tuning regimes. Best scores are in bold.

RQ2: The Adaptation Dichotomy. The main finding addressing RQ2 is a clear performance trade-off observed across multiple models. Although these models consistently improve results in classification, clustering, and NLI tasks (with only marginal degradation in a single case), they suffer a decline in STS performance. This drop is evident in both Spearman and Pearson correlation scores, as shown in the fifth and final (Δ) columns of Table 6. Importantly, these STS Δ values are computed against the frozen STS baseline (intrinsic similarity evaluation), not against the linear-probing classifier setting.

DT	Model	STS Pearson			STS Spearman		
		LP	FT	Δ	LP	FT	Δ
ASSIN	Albertina100M	0.584	0.581	-0.003	0.592	0.595	0.003
	Albertina900M	0.599	0.608	0.009	0.608	0.634	0.026
	BERTimbau	0.614	0.635	**0.021**	0.619	0.648	**0.029**
	BERTimbauLarge	0.582	0.599	0.017	0.597	0.618	0.021
	Gemma	0.395	0.340	-0.054	0.454	0.401	-0.053
	MiniLM	0.678	0.447	-0.231	0.679	0.524	-0.155
	Qwen	0.745	0.567	-0.178	0.739	0.585	-0.154
	Serafim335M	**0.809**	**0.730**	-0.079	**0.801**	**0.747**	-0.053
	Tucano	0.532	0.517	-0.015	0.541	0.527	-0.014
ASSIN 2	Albertina100M	0.614	0.487	-0.127	0.587	0.506	-0.082
	Albertina900M	0.601	0.648	0.047	0.565	0.623	0.058
	BERTimbau	0.672	0.690	0.019	0.614	0.677	0.063
	BERTimbauLarge	0.681	0.757	**0.076**	0.619	0.689	**0.070**
	Gemma	0.469	0.354	-0.115	0.466	0.390	-0.076
	MiniLM	0.772	0.289	-0.483	0.715	0.398	-0.317
	Qwen	0.805	0.618	-0.186	0.744	0.575	-0.169
	Serafim335M	**0.860**	**0.802**	-0.058	**0.832**	**0.806**	-0.026
	Tucano	0.561	0.550	-0.011	0.509	0.510	0.001

Table 6: STS results (Pearson and Spearman) under linear-probing and fine-tuning regimes. Best scores per dataset/setting are in bold. DT = dataset; LB = linear-Probing; FN = fine-tuning.

In contrast, Albertina900M and both BERTimbau variants not only achieve improvements in classification tasks but also present positive Δ values for STS tasks across both evaluated datasets. Notably, these models are the only ones to demonstrate consistent performance gains across all experimental settings. We hypothesize that BERT-based architectures benefit from a more "plastic" embedding space, which allows for task-specific specialization without significant loss of their general semantic representations. We refer to this property as "adaptation plasticity", and posit that it arises from a less constrained embedding geometry when compared to the more "rigid", cross-lingually aligned embedding spaces characteristic of global models.

We therefore use "catastrophic forgetting" as shorthand for this post-adaptation STS degradation relative to the frozen baseline, i.e., representation drift induced by supervised fine-tuning.

It is worth highlighting that, despite belonging to the same model family, Albertina100M and Albertina900M differ in several aspects. Specifically, they employ distinct tokenizers (DebertaTokenizerFast and DebertaV2TokenizerFast, respectively) and are trained on different corpora as a consequence of their size disparity (OSCAR and CulturaX, respectively) (Santos et al., 2024a). These architectural and data-related differences may be the reason for the discrepancies observed in the performances of both models.

RQ3: Which Embedding Family to Choose? Selecting an embedding family requires more than comparing performance metrics, as model size and execution time are also critical factors. High-performing models may be impractical if their runtime is prohibitive. To account for this, we report the average execution time (in seconds) over five test-set evaluations per task, as shown in Table 7. All models were evaluated with a maximum sequence length of 512 tokens and a batch size of 4 limitted by Gemma.

In classification, BERTimbauLarge achieves superior performance at the cost of higher runtime compared to MiniLM, but typically has a smaller runtime than the global multilingual models. Serafim offers comparable size and performance, with slightly higher execution time on Tupy and lower on HateBR but substantially lower runtime on AG News PT. When efficiency is paramount, BERTimbau provides near-competitive performance with significantly reduced execution time, making it a practical alternative despite being slower than MiniLM.

For clustering, BERTimbauLarge presents the best balance between performance and runtime, achieving results close to the top-performing models while remaining more efficient. In STS and NLI tasks, Serafim consistently delivers the strongest results, albeit with substantially higher execution times. While fine-tuning could be considered, there is a risk of catastrophic forgetting, which suggests linear probing as a more stable alternative.

Overall, the most effective strategy is task-specific model selection. However, when execution time constraints require a single model, BERTimbau offers a robust compromise between computational efficiency and predictive performance across tasks.

4.2 Analysis: The Plasticity vs. Rigidity Trade-off

For morphologically rich languages like PT-BR, a common assumption is that specialized language-specific tokenizers are more efficient than their general-purpose counterparts (Rust et al., 2021). Our benchmark includes a strong contingent of such specialized models (BERTimbau, Albertina, Tucano, Serafim). However, our quantitative analysis (Table 8) reveals this principle does not translate into a simple rule. We found no direct correlation between a model being monolingual and its tokenization efficiency or its downstream performance. Surprisingly, the monolingual decoder Tucano yielded the fewest tokens, while the monolingual encoder Albertina was consistently among

Model	CLS (seconds)			Clustering (seconds)			NLI (seconds)	STS (seconds)	
	AGN	HateBR	TuPy	AGN	HateBR	TuPy	ASSIN2	ASSIN	ASSIN2
Albertina100m	1190.527	28.024	430.625	33.371	7.391	38.395	71.356	31.394	18.826
Albertina900m	1800.243	59.084	488.531	90.571	14.028	76.620	125.725	63.767	34.786
BERTimbau	795.823	12.967	**233.455**	15.831	3.865	18.483	**29.450**	**12.688**	8.069
BERTimbauLarge	1039.507	22.827	322.232	31.081	5.985	31.059	51.820	22.369	13.877
Gemma	3026.848	79.020	697.281	141.704	18.039	105.376	146.326	85.862	40.190
Qwen	1501.652	54.261	498.322	61.007	12.210	67.097	131.940	57.971	35.070
Serafim335m	982.286	22.415	325.077	30.706	6.257	32.020	52.619	23.319	14.027
ST_miniLM	**629.193**	**12.589**	288.648	**13.017**	**2.826**	**14.678**	30.563	13.206	**8.047**
Tucano	1412.694	28.883	352.823	43.048	7.917	40.038	65.169	27.637	16.835

Table 7: Average time in seconds across tasks and datasets. AGN = AG News PT.

the most verbose, yet performed strongly. This finding refutes the simplistic hypothesis that monolingual focus alone guarantees a more compact or effective representation, revealing a deeper complexity we term the "Tokenization Dilemma."

Modelo	AGN	TuP	ASS	ASS2	HBR	ITW	Mean
Albertina100M	2.098	2.186	2.143	2.046	2.418	2.265	2.193
Albertina900M	1.841	1.884	1.855	1.707	2.007	1.914	1.868
BERTimbau	**1.400**	1.644	1.363	1.207	1.626	1.468	1.451
BERTimbauLarge	**1.400**	1.644	1.363	1.207	1.626	1.468	1.451
Gemma	1.416	1.491	1.458	1.187	1.536	1.429	1.420
QWen	1.686	1.749	1.730	1.529	1.874	1.725	1.716
Serafim335M	**1.400**	1.644	1.363	1.207	1.626	1.468	1.451
ST_miniLM	1.460	1.536	1.434	1.308	1.624	1.461	1.471
Tucano	1.311	**1.484**	**1.271**	**1.162**	**1.525**	**1.326**	**1.347**

Table 8: Average number of tokens per word from different models across datasets. Bold values indicate the smallest averages per column. AGN = AG News PT; TuP = TuPY; ASS = ASSIN; ASS2 = ASSIN2; HBR = HateBR.

A qualitative analysis of the tokenizers' behavior (Table 9) clarifies this dilemma by exposing their distinct strategic trade-offs. It is observed that WordPiece tokenizers (BERTimbau, Serafim) align more closely with Portuguese morphology (e.g., segmenting *preditiva* into *pred* + *itiva*) but struggle with loanwords, fragmenting a common term like *fake news* into semantically poor subunits. In contrast, SentencePiece tokenizers (Gemma, Tucano) are more token-efficient and robust to lexical borrowing, but their segmentations are often less morphologically faithful. Finally, Byte-level tokenizers (Albertina100M, Qwen) tend to oversegment Portuguese words, increasing token counts and reducing linguistic clarity.

This tokenization dilemma provides a compelling explanation for the adaptation dichotomy observed in our main results. We posit that these behaviors reflect a trade-off between the geometric properties of the models' embedding spaces.

We argue that the performance of monolingual encoders like BERTimbau and Albertina900M reflects "adaptation plasticity." Their specialized, monolingual vocabularies likely result in fewer cross-lingual constraints during pre-training. This, in turn, may lead to a less constrained and more malleable embedding geometry. We hypothesize that this flexibility allows the fine-tuning process to more effectively reorganize the space for Portuguese-centric tasks, explaining their consistent improvement on STS.

Conversely, we posit that the behavior of global models like Gemma reflects "adaptation rigidity." Their multilingual vocabularies and pre-training objectives are explicitly designed to promote strong cross-lingual alignment. We believe this creates a highly structured and constrained embedding geometry, optimized for consistency across languages. While this rigidity is beneficial for linear-probing transfer, it appears to resist task-specific adaptation. We suggest that fine-tuning acts as a disruptive force on this balanced structure, distorting the geometry and causing the observed catastrophic forgetting on STS.

4.3 Limitations and Future Work

Our study provides a quantitative analysis of model performance, focusing on *what* occurs during adaptation rather than the linguistic mechanisms underlying *why*. Although our tokenization analysis provides preliminary insights, several directions for deeper, theoretically grounded investigation remain, particularly those informed by Systemic Functional Linguistics (Figueredo et al., 2024).

First, our benchmark is intentionally practice-oriented and therefore includes both encoder-only and decoder-only architectures. As a result, part of the observed STS behavior may reflect architectural differences and sentence-embedding ex-

Model	Tokenization				
	'*pulsava*"	'*preditiva*"	'*saudosa*"	'*fundamento*"	'*fake news*"
Morphology	'puls 'a', 'va'	'predit', 'iva'	'saud', 'osa'	'fund', 'a', 'mento'	-
Albertina100M	'Ġpuls', 'ava'	'Ġpred', 'it', 'iva'	'Ġsa', 'ud', 'osa'	'Ġfundament', 'o'	'Ġfake', 'Ġnews'
Albertina900M	'_pul', 's', 'ava'	'_pre', 'di', 'tiva'	'_sau', 'd', 'osa'	'_fund', 'amento'	'_fake', '_news'
BERTimbau	'pul', '##sa', '##va'	'pred', '##itiva'	'saud', '##osa'	'funda', '##mento', 'o'	'fa', '##ke', 'ne', '##ws'
BERTimbauLarge	'pul', '##sa', '##va'	'pred', '##itiva'	'saud', '##osa'	'funda', '##mento', 'o'	'fa', '##ke', 'ne', '##ws'
Gemma	'_puls', 'ava'	'_predi', 'tiva'	'_sau', 'dos', 'a'	'_fundamento'	'_fake', '_news'
Qwen	'Ġpuls', 'ava'	'Ġpred', 'it', 'iva'	'Ġsa', 'ud', 'osa'	'Ġfund', 'amento', 'Ġo'	'Ġfake', 'Ġnews'
Serafim335M	'pul', '##sa', '##va'	'pred', '##itiva'	'saud', '##osa'	'funda', '##mento', 'o'	'fa', '##ke', 'ne', '##ws'
ST_MiniLM	'_pulsa', 'va'	'_pred', 'i', 'tiva'	'_sau', 'dos', 'a'	'_fundamento'	'_fake', '_news'
Tucano	'_puls', 'ava'	'_pred', 'itiva'	'_saud', 'osa'	'_fundamento'	'_f', 'ake', '_new', 's.'

Table 9: Qualitative tokenization of different words and expressions, showing the morphological structure of PT-BR terms. Contrary to expectations, Gemma's and MiniLM's tokenizer shows strong morphological alignment, while BERTimbau's monolingual vocabulary fails on common English loan-words.

traction choices, rather than multilinguality alone. In particular, decoder-only foundation models are not universally optimized for sentence embeddings, and alternative extraction/training strategies may alter their semantic similarity performance. Moreover, we did not include recent multilingual models explicitly trained for embedding quality (e.g., multilingual E5/BGE-style models), which could plausibly shift the multilingual curve in STS. Likewise, we did not benchmark some recent PT-focused decoder models (e.g., Cabrita/GlórIA/Gervásio-style releases), as our goal was a controlled comparison over widely used baselines with reproducible checkpoints.

Second, our datasets sample a limited set of textual registers (e.g., news and informal social media). A comprehensive understanding of model generalization requires a broader "registerial cartography" (Matthiessen, 2015), enabling the evaluation of adaptation across a wider range of situational contexts, including variations in field (e.g., technicality), tenor (e.g., formality and social distance), and mode (e.g., modalities and media).

Third, our adaptation protocol uses a fixed training budget (10 epochs) for uniformity across models. While this ensures comparability, it prevents us from isolating whether some STS degradation could be mitigated by early stopping, alternative schedules, or different adaptation budgets. A systematic analysis of training dynamics and stopping criteria would be a natural extension.

Finally, the current metric-based evaluation could be complemented by a dedicated linguistic probing suite. Such an approach would support a fine-grained analysis of the instantiation process, clarifying how fine-tuning reshapes a model's probabilistic preferences across lexical phenomena (e.g., polysemy), grammatical structures (e.g., syntax), and discourse semantics (e.g., appraisal). This would shift the analysis from documenting performance degradation on STS tasks to explaining how adaptation affects the model's capacity to map linguistic potential onto instantiation. In this direction, the notion of "adaptation plasticity" could be formalized as a measure of model sensitivity to forced subpotentialization during fine-tuning. Experimental designs could quantify how adaptation alters the *keyness* of linguistic features that characterize distinct tasks, advancing the field from performance benchmarking toward a principled understanding of how models capture the probabilistic and dynamic nature of language.

5 Conclusion

This paper presents a systematic comparison of global and local sentence-embedding models for PT-BR. Our results show that model selection entails a trade-off between performance, representational stability, and computational cost. Under linear probing, multilingual and monolingual models perform competitively. However, with lightweight fine-tuning, monolingual models consistently outperform. While dataset specialization improves performance in most tasks, it causes substantial degradation in STS, indicating adaptation-induced representation drift ("catastrophic forgetting" measured relative to the frozen STS baseline).

Tokenization analyses reveal that no tokenizer is universally superior. Despite strong downstream results, Tucano's tokenizer does not yield consistently better embeddings or task-level dominance. SentencePiece tokenizers demonstrate stronger morphological robustness and better handling of English loanwords, whereas WordPiece tokenizers produce morphologically aligned segmentations but struggle with frequent foreign terms. These

findings suggest that tokenization imposes distinct geometric constraints on the embedding space, directly influencing fine-tuning efficiency.

Acknowledgments

The authors would like to thank the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* - Brazil (CAPES) - Finance Code 001, *Fundacão de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG, grants APQ-01768-24), *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), and Universidade Federal de Ouro Preto (PROPPI/UFOP) for supporting the development of this study.

References

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian's, Malta. Association for Computational Linguistics.

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. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Nicholas Kluge Corrêa, Aniket Sen, Sophia Falk, and Shiza Fatimah. 2025. Tucano: Advancing Neural Text Generation for Portuguese. *Patterns*.

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Giacomo Figueredo, Gabriel Gomes Botelho Freitas, Laura Scaramussa Azevedo, and Lucas Alexandre Damasceno. 2024. Considerações sobre a organização do texto e da instanciação sob a perspectiva sistêmico-funcional. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 40(1):202440159824.

E Fonseca, L Santos, Marcelo Criscuolo, and S Aluisio. 2016. Assin: Avaliacao de similaridade semantica e inferencia textual. In *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal*, pages 13–15.

Gabriel Lino Garcia, Pedro Henrique Paiola, Luis Henrique Morelli, Giovani Candido, Arnaldo Cândido Júnior, Danilo Samuel Jodas, Luis C. S. Afonso, Ivan Rizzo Guilherme, Bruno Elias Penteado, and João Paulo Papa. 2024. Introducing bode: A fine-tuned large language model for portuguese prompt-based task. *Preprint*, arXiv:2401.02909.

Luís Gomes, António Branco, João Silva, João Rodrigues, and Rodrigo Santos. 2024. Open sentence embeddings for portuguese with the serafim pt* encoders family. *Preprint*, arXiv:2407.19527.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Celio Larcher, Marcos Piau, Paulo Finardi, Pedro Gengo, Piero Esposito, and Vinicius Caridá. 2023. Cabrita: closing the gap for foreign languages. *Preprint*, arXiv:2308.11878.

Ricardo Lopes, Joao Magalhaes, and David Semedo. 2024. GlórIA: A generative and open large language model for Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 441–453, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Maicon Domingues. 2023. canarim-7b (revision 08fdd2b).

Maritaca-AI. 2023. ag_news_pt.

Christian M.I.M. Matthiessen. 2015. Register in the round: registerial cartography. *Functional Linguistics*, 2(1):1–48.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

Felipe Oliveira, Victoria Reis, and Nelson Ebecken. 2023. Tupy-e: detecting hate speech in brazilian portuguese social media with a novel dataset and comprehensive analysis of models. *arXiv preprint arXiv:2312.17704*.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems*, volume 36, pages 36963–36990. Curran Associates, Inc.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. *Sabiá: Portuguese Large Language Models*, page 226–240. Springer Nature Switzerland.

Livy Real, Erick Fonseca, and Hugo Goncalo Oliveira. 2020. The assin 2 shared task: a quick overview. In *International Conference on Computational Processing of the Portuguese Language*, pages 406–412. Springer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135.

Rodrigo Santos, João Rodrigues, Luís Gomes, João Silva, António Branco, Henrique Lopes Cardoso, Tomás Freitas Osório, and Bernardo Leite. 2024a. Fostering the ecosystem of open neural encoders for portuguese with albertina pt* family. *Preprint*, arXiv:2403.01897.

Rodrigo Santos, João Silva, Luís Gomes, João Rodrigues, and António Branco. 2024b. Advancing generative ai for portuguese with open decoder gervásio pt-*. *Preprint*, arXiv:2402.18766.

Fabricio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, pages 403–417. Springer.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. Hatebr: A large expert annotated corpus of brazilian instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP (RepL4NLP)*. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.