Biatron: A Parameter-Efficient Small Language Model for Brazilian Portuguese with Integrated Mathematical Reasoning

Daniel Fazzioni and Maria C. X. de almeida and Anna P. V. L. B. Moreira
Anderson S. Soares and **Sávio S. T. de Oliveira** and **Fernando M. Federson**

Institute of Informatics– Federal University of Goiás (UFG)
fazzioni@egresso.ufg.br
federson@ufg.br

Abstract

The development of Small Language Models (SLMs) for Portuguese faces significant challenges in balancing parameter efficiency with specialized capabilities, particularly in mathematical reasoning domains where existing models demonstrate limited native competence. This work introduces the first model in the Biatron series, a 345-million-parameter language model specifically optimized for Brazilian Portuguese through strategic data curation rather than brute-force parameter scaling. Using a carefully designed 60-30-10 data mixture combining high-quality Portuguese text from GigaVerbo, chain-of-thought reasoning examples, and mathematical datasets, Biatron was trained on 300 billion tokens using the Megatron-LM framework, achieving 32% Model FLOP Utilization. The model attains an overall score of 0.245 (aggregate performance) on Portuguese-specific benchmarks, approaching within 1.6% of Tucano-630M's performance while utilizing 45% fewer parameters. Most significantly, Biatron achieves 7.5% Pass@1 accuracy on mathematical reasoning tasks—more than doubling the performance of Tucano-2.4B (3.5%) despite being nearly seven times smaller. These results validate that strategic data mixing can rival parameter scaling for language model development, establishing a reproducible methodology for efficient AI development in resource-constrained language contexts. To support reproducibility and further research, the final model weights, training logs, and intermediate checkpoints are publicly available[1]

1 Introduction

Large Language Models (LLMs) have established a new paradigm in artificial intelligence, demonstrating an unprecedented ability to process and generate human language across diverse domains. The development of these systems typically involves massive computational resources and extensive datasets to achieve high-level linguistic competence (Kumar, 2024; Joshua et al., 2025).

A central characteristic of this field is the scaling law, where model performance is often tied to the number of parameters, the volume of training tokens, and the architectural efficiency of transformer blocks (Kaplan et al., 2020; Hoffmann et al., 2022). Furthermore, the adaptation of these models to specific languages and the optimization of their reasoning capabilities through curated data mixtures represent critical dimensions of modern architectural design (Zheng et al., 2023; Mumuni and Mumuni, 2025).

Despite the dominance of massive models, the scientific community has identified a significant challenge in developing Small Language Models (SLMs) that maintain high performance in non-English languages, such as Portuguese, without requiring prohibitive hardware. Previous efforts to address this in the Brazilian context include models such as Tucano (Corrêa et al., 2025) and Sabiá (Pires et al., 2023; Sales Almeida et al., 2024), which utilized large-scale corpora such as GigaVerbo to establish baselines for the language. However, these earlier interventions often face trade-offs between parameter efficiency and specialized reasoning, particularly in technical domains like mathematics, leaving a gap for more compact yet capable architectures (Abdin et al., 2023; Zhang et al., 2024).

This research introduces Biatron-345M, the first member of a new family of SLMs specifically optimized for the Portuguese language. Using the Megatron-LM framework and a 300 billion token strategic data mix, combining 60% high-quality Portuguese text from GigaVerbo, 30% synthetic chain-of-thought reasoning and 10% mathematical datasets, this work demonstrates that a compact model can achieve competitive results with signifi-

[1] https://huggingface.co/Biatron/biatron-345m
https://github.com/Fazzioni/Biatron
https://api.wandb.ai/links/fazzioni/p5uymwk6

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 868–877

cantly larger architectures. The model achieves a simple arithmetic mean score of 0.245 on Portuguese-specific benchmarks, closely approaching the performance of Tucano-630M (0.249) while using nearly half the parameters and outperforming established baselines such as SmolLM2-360M, Qwen3-0.6B, and Gemma-3-270M. Additionally, Biatron-345M exhibits superior convergence during supervised fine-tuning and extends the context window to 4096 tokens—double that of comparable Tucano models. By demonstrating that careful data curation and strategic mixing can rival brute-force scaling, this work establishes a reproducible methodology for efficient, localized AI development and lays the foundation for future models in the Biatron family.

2 Related Work

The computational costs of Large Language Models have accelerated the development of Small Language Models (SLMs), typically defined as parameter-efficient architectures (<7-10B) optimized for resource-constrained environments (Nguyen et al., 2024). This shift relies on strategic data curation rather than brute-force scaling. Pioneering efforts include Microsoft's Phi family (Gunasekar et al., 2023; Javaheripi et al., 2023; Abdin et al., 2024), which validated the use of "textbook-quality" synthetic data, and Google's Gemma series (Gemma Team et al., 2024a,b), which utilized knowledge distillation to achieve state-of-the-art results. The open-source community further established efficiency benchmarks with TinyLlama (Zhang et al., 2024) and the data-centric SmolLM family (Ben Allal et al., 2025), while Alibaba's Qwen series (Bai et al., 2023; Yang et al., 2024; Qwen Team et al., 2024) demonstrated that multilingual SLMs can maintain performance in various languages. These works collectively establish SLMs as viable alternatives to massive models for practical applications.

SLMs for Portuguese

Despite global advances, Portuguese remains underrepresented in high-quality training data, creating a significant gap for model development. The scarcity of curated datasets in specialized domains—such as mathematics and structured reasoning—compounds the difficulty of training competitive models without resorting to translation, which often degrades linguistic authenticity. Furthermore, infrastructure constraints in Portuguese-speaking regions require efficient architectures that democratize AI development without requiring massive GPU clusters. Moreover, Portuguese morphological complexity and regional variations require targeted modeling approaches beyond generic multilingual solutions.

Recent efforts summarized in Table 1 have established important baselines. Decoder-only models such as Tucano (Corrêa et al., 2024b), Sabiá (Pires et al., 2023), and Glória (Lopes et al., 2024) offer strong general performance using corpora such as GigaVerbo (Silva et al., 2018), but prioritize web text over technical domains. In contrast, encoder-only architectures such as Albertina (Rodrigues et al., 2023), BERTimbau (Souza et al., 2020), and their domain-specific variants (Finardi et al., 2021; Pires et al., 2025; Campiotti et al., 2023) excel in classification but lack generative capabilities. A critical gap remains for compact generative models that integrate native mathematical reasoning and chain-of-thought capabilities during pre-training. Biatron addresses this by combining strategic data curation with efficient training to achieve competitive performance in both general understanding and reasoning tasks.

3 Model Architecture

The Biatron architecture follows the principles of modern transformer design, maintaining adequate computational efficiency for resource-constrained environments. As previously described, Biatron is a decoder-only transformer model composed of 345 million parameters, trained from scratch on Portuguese data.

The model consists of 32 transformer layers with a hidden dimension of 960. Grouped Query Attention (GQA) is employed with 15 attention heads and 5 query groups, which reduces the memory footprint and computational cost compared to standard multi-head attention while maintaining model quality. This design choice follows recent trends in efficient language model architectures that demonstrate GQA's ability to achieve performance comparable to Multi-Head Attention (MHA) with significantly reduced KV cache requirements. The feed-forward network in each layer uses a hidden dimension of 3840 (4 times the model dimension), following the standard scaling factor established in transformer architectures, with GELU activation functions. Layer normalization is applied for stable

Table 1: Overview of Small Language Models developed for Portuguese.

Model/Family	Parameters	Year	Architecture	Domain/Focus	Reference
Tucano	160M, 630M, 1.1B, 2.4B	2024	Decoder-only	General (BR)	(Corrêa et al., 2024b)
TeenyTinyLlama	160M, 460M	2024	Decoder-only	General (BR)	(Corrêa et al., 2024a)
Sabiá	7B, 65B	2023	Decoder-only	General (BR)	(Pires et al., 2023)
Glória	1.3B, 2.7B	2024	Decoder-only	General (EU-PT)	(Lopes et al., 2024)
Albertina	100M, 900M, 1.5B	2023	Encoder-only	General (BR/EU)	(Rodrigues et al., 2023)
BERTimbau	110M, 330M	2020	Encoder-only	General (BR)	(Souza et al., 2020)
DeBERTinha	40M	2023	Encoder-only	General (BR)	(Campiotti et al., 2023)
BERTaú	110M	2021	Encoder-only	Financial (BR)	(Finardi et al., 2021)
DeB3RTa	70M, 426M	2025	Encoder-only	Financial (BR)	(Pires et al., 2025)

training dynamics.

The model processes sequences of up to 4096 tokens – double the context window of comparable Tucano models (2048 tokens), enabling the capture of longer-range dependencies crucial for Portuguese text understanding. This fixed sequence length was maintained throughout the training without employing curriculum learning strategies that progressively increase the length of the context. For positional information, Rotary Position Embeddings (RoPE) are utilized with a base frequency of 10,000 and full rotation (100% of dimensions), which has proven effective in extrapolating to longer sequences than those seen during training.

The architectural design draws from established transformer variants, building on a GPT-style foundation while incorporating refinements from more recent architectures. Notably, RoPE from the Llama architecture family is adopted, which provides superior length extrapolation capabilities compared to absolute positional encodings. The use of tied embeddings – where input token embeddings and output prediction weights share parameters – reduces the total parameter count while maintaining model quality, a design choice that proves particularly effective for models in the sub-billion parameter range.

The architecture incorporates several standard optimization techniques in the Megatron-LM framework, including fused operations for cross-entropy loss computation and gradient reduction in bfloat16 precision, which collectively contribute to improved training efficiency. The key architectural specifications are resumed in Table 2.

Table 2: Architectural specifications of Biatron-345M.

Specification	Value
Parameters	345M
Layers	32
Hidden Dimension	960
Feed-Forward Dimension	3840 (4× hidden)
Attention Mechanism	Grouped Query Attention
Attention Heads	15
Query Groups	5
Activation Function	GELU
Normalization	LayerNorm
Maximum Sequence Length	4096 tokens
Vocabulary Size	32,000 tokens
Embeddings	Tied input-output
Position Encoding	RoPE (base=10k, rotation=100%)
Precision	bfloat16

4 Pre-training

Having established the architectural foundation of Biatron-345M, this section details the data curation strategies and training infrastructure. The methodology emphasizes strategic data mixing and efficient training practices to optimize performance within the constraints of limited Portuguese-language resources.

4.1 Training Data

The pre-training corpus draws from three primary data sources. The foundation is GigaVerbo, a large-scale corpus of web-scraped content, legal documents, and instructional data specifically curated for Brazilian Portuguese (Corrêa et al., 2024b). For mathematical reasoning, FineMath-4plus and InfiWebMath-4plus are incorporated, both subsets of the FineWeb dataset that provide high-quality mathematical content with step-by-step solutions (Allal et al., 2025b). To enhance chain-of-thought capabilities, the training incorporated a translated version of the Glaive AI dataset, containing ≈20 million reasoning examples (Moro, 2024).

This selection was motivated by recent findings—specifically the methodology established by the SmolLM3 team (Allal et al.,

2025a)—demonstrating that strategic data mixing can rival parameter scaling. The hypothesis was that simultaneous exposure to natural language, math, and structured chain-of-thought during pre-training would yield a more capable model than training on web text alone, potentially facilitating future alignment work through behavioral traces of verification.

Quality Filtering and Preparation

The pre-training leveraged a quality-filtered version of GigaVerbo using a learned classifier based on a fine-tuned BERTimbau-base (trained on GPT-4o annotations). A conservative strategy was adopted, selecting only samples with $> 95\%$ confidence to prioritize precision over recall. The resulting subset retained $\approx$70% of the original corpus (135 billion unique tokens, sampled for a total of 180 billion tokens seen during training). The mathematical and reasoning datasets were used without additional filtering, leveraging their native domain-specific curation.

Data Mixture and Training Corpus

The 60-30-10 distribution (Table 3) reflects a deliberate balance. The 60% allocation to GigaVerbo ensures foundational linguistic competence in Brazilian Portuguese. The 30% allocation to chain-of-thought data represents a key decision to integrate structured reasoning patterns directly into the model's representations rather than relegating them to fine-tuning. The remaining 10% addresses the critical gap in native mathematical reasoning capabilities absent in previous Portuguese models.

Inspired by staged training approaches like SmolLM2, this continuous multi-task exposure ensures that each batch contains examples from all domains. This regime allows the model to learn linguistic patterns and reasoning structures simultaneously. The final 300 billion token budget was achieved through multi-epoch training on these curated datasets, prioritizing robust learning over simple token volume scaling (Hoffmann et al., 2022).

Tokenization

A custom Byte Pair Encoding (BPE) tokenizer was developed specifically for Portuguese (trained on 50M FineWeb samples) with a vocabulary of 32,000 tokens. This avoids the inefficiencies of multilingual tokenizers, improving compression ratios and allowing longer contextual windows within the sequence budget. The vocabulary size balances computational efficiency with morphological coverage for complex Portuguese conjugations, aligning with strategies from efficient models such as SmolLM2.

4.2 Training Setup

The Pre-training was conducted using the Megatron-LM framework (Shoeybi et al., 2019) on NVIDIA H100 GPUs. Megatron-LM was selected for its fused kernels, optimized attention mechanisms, and native bfloat16 support, which provide numerical stability and reduced memory overhead compared to standard implementations. These optimizations enabled a 32% Model FLOPs Utilization (MFU), a substantial improvement over standard libraries (3.6%), making the 300B token run feasible in 792.72 hours (approximately 33 days) on a single H100.

Hyperparameters and Optimization

The configuration follows best practices for this scale, adapted for Portuguese:

Batch and Optimization: A global batch size of 512 (micro-batch 16) was used to maximize utilization. The AdamW optimizer was employed ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.1, gradient clipping 1.0). Training used a cosine learning rate schedule (max 6.0×10^{-5}, min 6.0×10^{-6}) over 152,000 steps with a warmup period of 0.1%.

Precision and Parallelism: Training used bfloat16 precision with Megatron-LM optimizations (cross-entropy fusion, CUDA graphs). To optimize resource utilization, GPU counts — ranging from 1 to 6 NVIDIA H100s — were frequently reallocated based on infrastructure availability. The setup utilized a "no_shard" data parallel strategy with overlapped gradient reduction to maximize throughput.

Implementation Details

The training infrastructure leveraged several advanced features of the Megatron-LM framework to maximize efficiency:

Attention implementation: the fused attention backend provided by Megatron-LM was utilized, which implements optimized attention kernels that reduce memory bandwidth requirements and improve computational efficiency.

Memory optimizations: activation checkpointing (gradient checkpointing) was enabled to trade computation for memory, allowing training of larger models on limited hardware. This technique

Table 3: Pre-training data mixture for Biatron-345M. Token counts are computed using our custom 32k BPE tokenizer. The GigaVerbo subset includes only samples classified as high quality with $> 95\%$ confidence by the BERTimbau-based filter. Dataset Tokens refers to the size of each dataset after filtering. Tokens Seen refers to the total number of tokens consumed during training, accounting for multi-epoch sampling.

Data Source	Proportion	Dataset Tokens	Epochs	Tokens Seen	Domain
GigaVerbo (High Quality)	60%	135.0B	1.35	180.0B	General Portuguese
Reasoning-PT-20M	30%	45.0B	2.00	90.0B	Chain-of-Thought
FineMath-4plus	5%	13.2B	1.13	15.0B	Mathematics
InfiWebMath-4plus	5%	11.8B	1.27	15.0B	Mathematics
Total	**100%**	—	—	**300.0B**	

recomputes activations during the backward pass rather than storing them, significantly reducing peak memory usage.

Communication optimization: for distributed training scenarios, Megatron-LM implements overlapped communication and computation, where gradient reduction occurs concurrently with backward pass computation, minimizing communication overhead.

The combination of these optimizations, along with careful hyperparameter tuning and data mixture design, enabled successful training of Biatron-345M to present strong empirical performance in Portuguese language understanding tasks, as well as superior mathematical ability, detailed in the following section.

5 Evaluation and Results

Having established the pre-training methodology, this section presents a comprehensive empirical evaluation of Biatron-345M's capabilities across Portuguese language understanding and mathematical reasoning tasks, demonstrating that strategic data mixing can produce competitive performance relative to significantly larger models.

5.1 Evaluation Protocol

To assess Biatron-345M across multiple dimensions, a diverse suite of benchmarks was selected: ENEM (Brazil's national high school exam), OAB (Brazilian Bar Association exam), EXAMS (university-level entrance examinations), and OpenAI MMLU Portuguese (57-subject comprehensive assessment). All language understanding benchmarks employ Multiple Choice Formulation (MCF), providing standardized evaluation and enabling direct comparison with baseline models.

To assess the impact of mathematical data inclusion during pre-training, a custom dataset of 200 synthetic word problems was developed covering addition, subtraction, multiplication, and division. Problems are formulated as natural language scenarios in Brazilian Portuguese contexts rather than abstract operations, requiring both linguistic comprehension and computational reasoning. Models must generate numerical answers directly rather than selecting from alternatives, with answers extracted using pattern matching on bracketed responses.

Baseline models include the Tucano family (160M, 630M, 1.1B, 2.4B parameters), SmolLM2-360M, Qwen3-0.6B, and Gemma-3 variants (270M, 1B-PT), enabling comparison across parameter efficiency, architectural choices, and data curation approaches.

All evaluations used the LightEval framework for standardized protocols and reproducibility. Portuguese benchmarks were evaluated in zero-shot settings using MCF, with models selecting among four to five answers depending on assessment, based solely on pre-training knowledge. Mathematical reasoning used generative assessment with 8 independent generations per problem, computing Pass@K metrics that represent the probability of finding at least one correct solution among K attempts. This generative evaluation was designed to complement MCF benchmarks, ensuring that observed performance gains reflect genuine capability rather than answer-probability exploitation. The models were supervised fine-tuning on 27,000 synthetic word problems (3 epochs over 9,000 base problems) using AdamW optimization with learning rate 5×10^{-5}, cosine decay, batch size 512 and gradient clipping at 0.8. All evaluations used bfloat16 precision on NVIDIA GPUs.

5.2 Overall Benchmark Performance

The aggregated performance across Portuguese-specific benchmarks reveals Biatron-345M's competitive positioning within the landscape of small

language models. Achieving an overall score of 0.245, Biatron-345M approaches within 1.7% of Tucano-630M's performance (0.249) despite utilizing approximately 45% fewer parameters (345M versus 630M). This result provides empirical validation for the hypothesis that strategic data curation can rival parameter scaling as a pathway to competitive model performance.

As shown in Table 4, Biatron-345M ranks third overall, trailing only Gemma-3-1B-PT (0.257, 1.0B parameters) and Tucano-630M. More significantly, Biatron-345M outperforms all models with similar or smaller parameter counts, surpassing SmolLM2-360M (0.235) by 10 points (where 1 point corresponds to 10^{-3}), Qwen3-0.6B (0.229, 600M parameters) by 16 points, and Gemma-3-270M (0.229) by 16 points.

On ENEM, Brazil's national high school examination, Biatron-345M achieves 0.216 accuracy, representing the highest score among all models in its parameter class and surpassing even the larger Tucano-630M (0.197)—an 19-point advantage particularly significant given ENEM's comprehensive assessment of general knowledge and Brazilian Portuguese language proficiency. In Portuguese MMLU (57 subjects), Biatron-345M achieves 0.248 accuracy, maintaining a substantial advantage over SmolLM2-360M (0.239). In OAB legal reasoning, Biatron-345M achieves 0.245, approaching Tucano-630M (0.247) within 2 points despite allocating 40% of training tokens to mathematical and reasoning domains rather than additional Portuguese text.

5.3 Mathematical Reasoning: A Novel Strength

The strategic inclusion of mathematical datasets during pre-training yields the most dramatic performance differentials observed in this evaluation. Following supervised fine-tuning, the models were evaluated on 200 held-out problems using Pass@K metrics.

Among models pre-trained without mathematical data, specifically, the entire Tucano family, Biatron-345M demonstrates markedly superior performance. At Pass@1, Biatron-345M achieves 7.5% accuracy, more than doubling Tucano-2.4B's 3.5% despite being nearly seven times smaller. This advantage widens at Pass@8: Biatron-345M reaches 19.0%, exceeding Tucano-2.4B's 14.0% by 5 percentage points. Table 5 presents the complete results.

The multilingual Qwen3-0.6B, likely benefiting from extensive mathematical data across multiple languages, substantially outperforms all Portuguese-specific models with 89.5% Pass@1 accuracy, establishing an upper bound. Analysis by arithmetic operation reveals consistent advantages for Biatron-345M, with particularly striking results in multiplication: Biatron-345M achieves 10.9% Pass@1 accuracy, where all Tucano models fail completely (0.0%). For subtraction and division, Biatron-345M achieves 6.4% and 7.7%, respectively, matching or substantially exceeding Tucano-2.4B despite the seven-fold parameter disadvantage.

These results provide compelling empirical validation for incorporating mathematical datasets during pre-training rather than relying exclusively on post-training adaptation, establishing a new baseline for Portuguese-specific architectures in mathematical reasoning.

Beyond final performance, Biatron-345M demonstrates superior training dynamics during supervised fine-tuning, achieving a lower validation loss more rapidly than all Tucano variants, including Tucano-2.4B. This accelerated convergence suggests that the pre-training data mixture –particularly chain-of-thought reasoning examples— provides favorable initialization for instruction following and structured reasoning tasks.

5.4 Linguistic Competence and Extended Context

To assess linguistic competence on morphologically challenging aspects of Portuguese and validate mathematical capabilities in few-shot settings, models were evaluated on two additional tasks. These tasks were selected as generative evaluations that complement the MCF-based benchmarks in Section 5.2, providing stronger evidence that observed performance advantages reflect genuine linguistic and mathematical competence rather than answer-probability artifacts inherent to multiple-choice formats. Figure 1 presents the accuracy versus the size of the model for the completion of verb conjugation (5,994 examples) and the addition of arithmetic with a few-shots.

In verb conjugation, Biatron-345M achieves 50.5% accuracy, positioning between Tucano-630M (53.0%) and Tucano-160M (42.7%), as shown in Figure 1(a). The 2.5 percentage point gap relative to Tucano-630M is substantially smaller than the performance advantages observed in math-

Table 4: Aggregate performance across Portuguese benchmarks (zero-shot MCF). Scores represent accuracy.

Model	Params	OAB	ENEM	MMLU-PT	EXAMS	Overall
Gemma-3-1B-PT	1.0B	0.243	0.199	**0.262**	**0.250**	**0.257**
Tucano-630M	630M	**0.247**	0.197	0.254	0.226	0.249
Biatron-345M	**345M**	0.245	**0.216**	0.248	0.224	0.245
SmolLM2-360M	360M	0.231	0.201	0.239	0.213	0.235
Tucano-160M	160M	0.229	0.209	0.234	0.222	0.231
Qwen3-0.6B	600M	0.230	0.207	0.231	0.222	0.229
Gemma-3-270M	270M	0.230	0.203	0.231	0.220	0.229

Bold indicates best score per column. Benchmarks use 4 to 5-choice MCF depending on assessment. Scores below the 0.20 to 0.25 random baseline reflect the difficulty of these benchmarks for sub-1B models across the board.

Table 5: Math reasoning performance with Pass@K metric. Best Portuguese-specific model results (excluding Qwen3) are shown in bold.

Model	Params	Pass@1	Pass@2	Pass@4	Pass@8
Qwen3-0.6B	600M	0.895	0.905	0.905	0.910
Biatron-345M	**345M**	**0.075**	**0.095**	**0.120**	**0.190**
Tucano-2.4B	2.4B	0.035	0.070	0.095	0.140
Gemma-3-270M	270M	0.050	0.075	0.095	0.110
Tucano-1.1B	1.1B	0.005	0.015	0.040	0.060
Tucano-630M	630M	0.010	0.010	0.030	0.060
Tucano-160M	160M	0.000	0.005	0.010	0.020

ematical reasoning (7.5% versus 1.0% Pass@1), indicating that strategic inclusion of specialized domains does not severely compromise core language modeling capabilities.

In the evaluation of arithmetic few-shots (Figure 1(b)), Biatron-345M demonstrates superior scaling behavior, maintaining competitive performance that reinforces the mathematical capabilities observed in supervised fine-tuning. Notably, the model outperforms substantially larger Tucano models and breaks the expected linear relationship between parameter count and accuracy, providing visual confirmation of the effectiveness of pre-training data mixture.

Biatron-345M's extended 4,096-token context window—double the 2,048-token capacity of comparably-sized Tucano models—enables processing longer documents and maintaining coherence over extended passages, critical for document question-answering, long-form generation, and multi-turn dialog. Synthesizing results reveal that Biatron-345M delivers performance within 1.6% of Tucano-630M while requiring 45% fewer parameters, translating to reduced inference costs, lower memory requirements, and feasibility for consumer hardware deployment – critical for democratizing access in resource-constrained environments.

6 Conclusion

This work introduced Biatron-345M, a 345-million-parameter language model optimized for Brazilian Portuguese that demonstrates strategic data curation as a viable alternative to brute-force parameter scaling. Through a carefully designed 60-30-10 data mixture that combines high-quality Portuguese text, chain-of-thought reasoning, and mathematical datasets, Biatron-345M achieves competitive performance with significantly larger models while introducing novel capabilities previously absent in Portuguese-specific architectures.

6.1 Summary of Contributions

Biatron-345M achieves an overall score of 0.245 on Portuguese benchmarks, approaching within 1.6% of Tucano-630M (0.249) while using 45% fewer parameters. The model demonstrates particularly strong performance on ENEM (0.216), outperforming all models in its parameter class and validating the hypothesis that strategic data mixing can rival parameter scaling.

The most significant finding emerges in mathematical reasoning, where Biatron-345M achieves 7.5% Pass@1 accuracy—more than doubling Tucano-2.4B's performance (3.5%) despite being nearly 7× smaller. This represents a fundamental advance for Portuguese NLP, as existing Portuguese-specific models have traditionally

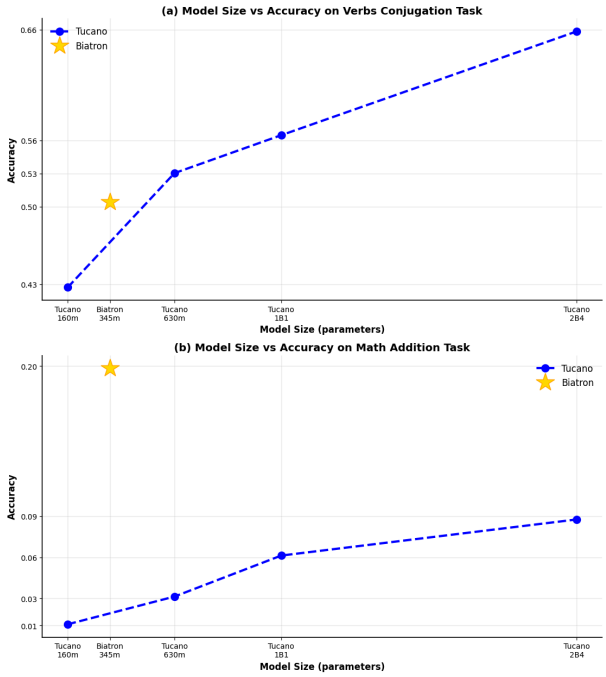

Figure 1: Model performance scaling across linguistic and mathematical tasks. **(a)** Verb conjugation accuracy demonstrates expected parameter scaling, with Biatron-345M positioned between Tucano-630M and Tucano-160M. **(b)** Few-shot arithmetic addition reveals Biatron-345M's superior mathematical reasoning, outperforming substantially larger Tucano models and confirming the effectiveness of pre-training data mixture.

lacked native mathematical reasoning capabilities. The performance advantage is particularly pronounced in multiplication (10.9% Pass@1), where larger models without mathematical pre-training fail completely.

The primary contributions of this research are: **(1)** Biatron-345M demonstrates that 345 million well-trained parameters can compete with models nearly twice their size through strategic data curation, establishing new benchmarks for parameter efficiency in Portuguese language models; **(2)** The strategic inclusion of mathematical datasets during pre-training successfully introduces numerical reasoning capabilities absent in existing Portuguese models, validating the approach of specialized data integration rather than post-hoc fine-tuning alone; **(3)** The training methodology achieves 32% Model FLOP Utilization on NVIDIA H100 GPUs, demonstrating that academic research groups can develop competitive language models within typical resource constraints; **(4)** Biatron-345M establishes the first member of a new model family and provides a reproducible framework for the development of efficient language models in resource-constrained language contexts.

6.2 Limitations and Future Directions

Several limitations warrant acknowledgment. Although Biatron-345M substantially outperforms existing Portuguese models in mathematical reasoning, its absolute performance (7.5% Pass@1) remains well below multilingual models such as Qwen3-0.6B (89.5%), indicating substantial room for improvement. The model was trained exclusively on Brazilian Portuguese data with some multi-epoch processing, and its performance on European Portuguese and other regional variants remains uncharacterized. The chain-of-thought reasoning dataset consists of machine-translated English examples, which may introduce linguistic artifacts. Finally, the mathematical evaluation focused on basic arithmetic operations, leaving the more advanced reasoning capabilities unassessed.

Promising research directions include expanding the Biatron family with larger variants to clarify whether performance advantages of strategic data mixing persist at scale and incorporating more sophisticated mathematical reasoning datasets beyond basic arithmetic. Developing native Portuguese chain-of-thought reasoning data would eliminate translation artifacts while better reflecting Brazilian cultural contexts. The development of conversational and instruction-tuned variants, the application of alignment techniques such as RLHF using Brazilian Portuguese preference data, and the systematic assessment of long-context performance would enhance practical usability. Finally, model quantization and compression techniques could enable deployment on consumer hardware, furthering the democratization objectives motivating this work.

6.3 Final Remarks

This research demonstrates that careful data curation and strategic mixture design can produce language models that compete effectively with significantly larger architectures while introducing novel capabilities. Biatron-345M establishes a reproducible methodology for developing language technologies in relatively under-resourced contexts, showing that academic research groups can develop competitive models without industrial-scale computational resources. By demonstrating that compact, efficient models can deliver competitive performance and unique capabilities through data-centric approaches, this work contributes to the

democratization of artificial intelligence development across linguistic and geographic boundaries. The Biatron family represents an ongoing commitment to advancing Portuguese language technologies through open and reproducible research that serves the needs of diverse Portuguese-speaking communities.

Acknowledgments

We gratefully acknowledge the TucanoBR project and its authors for their significant contributions to Brazilian Portuguese LLM research. Their release of the GigaVerbo dataset and the Tucano model family provided essential resources that supported this work. We also thank the Center of Excellence in Artificial Intelligence (CEIA) at the Federal University of Goiás (UFG) for providing the computational infrastructure that made this research possible.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint*.

Maribeth Abdin, Yuxuan Wang, Alex Fang, Tri Nguyen, Yao Zhang, Thomas Scialom, Lianmin Yu, and 1 others. 2023. Phi-1: Textbooks are all you need. *arXiv preprint*.

Loubna Ben Allal, Elie Bakouch, Anton Lozhkov, Leandro von Werra, and Thomas Wolf. 2025a. SmolLM3: smol, multilingual, long-context reasoner. Hugging Face Blog post. Accessed: 2026-01-06.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025b. SmolLM2: When Smol Goes Big – Data-Centric Training of a Small Language Model. *Preprint*, arXiv:2502.02737. ArXiv preprint arXiv:2502.02737.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, and 1 others. 2023. Qwen technical report. *arXiv preprint*. Qwen Team, Alibaba Group.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, and 1 others. 2025. SmolLLM2: When smol goes big – data-centric training of a small language model. *arXiv preprint*.

Israel Campiotti, Matheus Rodrigues, Yuri Albuquerque, Rafael Azevedo, and Alyson Andrade. 2023. Debertinha: A multistep approach to adapt DeBERTaV3 xsmall for brazilian portuguese natural language processing tasks. *arXiv preprint*.

Nicholas Kluge Corrêa, Sophia Falk, Shiza Fatimah, Aniket Sen, and Nythamar de Oliveira. 2024a. TeenyTinyLlama: Open-source tiny language models trained in brazilian portuguese. *arXiv preprint*.

Nicholas Kluge Corrêa, Aniket Sen, Sophia Falk, and Shiza Fatimah. 2024b. Tucano: Advancing neural text generation for portuguese. *arXiv preprint*.

Nicholas Kluge Corrêa, Aniket Sen, Sophia Falk, and Shiza Fatimah. 2025. Tucano: Advancing neural text generation for portuguese. *Patterns*, 6(11):101325.

Paulo Finardi, José Dié Viegas, Gustavo T. Ferreira, Alex F. Mansano, and Vinicius F. Caridá. 2021. Bertaú: Itaú bert for digital customer service. *arXiv preprint*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, and 1 others. 2024a. Gemma: Open models based on gemini research and technology. *arXiv preprint*.

Gemma Team, Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, and 1 others. 2024b. Gemma 2: Improving open language models at a practical size. *arXiv preprint*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, and 1 others. 2023. Textbooks are all you need. *arXiv preprint*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, and 1 others. 2022. Training compute-optimal large language models. *arXiv preprint*.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, and Jyoti Aneja. 2023. Phi-2: The surprising power of small language models. Microsoft Research Blog.

Jonah Vincent Joshua, Diosdado Asumu Ndong Andeme, and Manuel Martin Ela Ndong. 2025. Large language models: Advances, challenges, and future directions. *International Journal of Advanced Trends in Computer Science and Engineering*, 14(2):55–57.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, and 1 others. 2020. Scaling laws for neural language models. *arXiv preprint*.

Pranjal Kumar. 2024. Large language models (llms): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, 57:260.

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 Linguistics.

Cássio Moro. 2024. reasoning-v1-20m-portuguese. https://huggingface.co/datasets/cnmoro/reasoning-v1-20m-portuguese. Accessed: 2026-01-02.

Alhassan Mumuni and Fuseini Mumuni. 2025. Large language models for artificial general intelligence (AGI): A survey of foundational principles and approaches. *arXiv preprint*.

Chien Van Nguyen, Xuan Shen, Ryan Aponte, Yu Xia, and 1 others. 2024. A survey of small language models. *arXiv preprint*.

Henrique Pires, Luis Paucar, and João P. Carvalho. 2025. Deb3rta: A transformer-based model for the portuguese financial domain. *Big Data and Cognitive Computing*, 9(3):51. Financial domain-specific DeBERTa model for Portuguese.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Intelligent Systems, BRACIS 2023*, volume 14197 of *Lecture Notes in Computer Science*, pages 226–240, Cham. Springer.

Qwen Team, An Yang, Baosong Yang, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint*.

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-*. *arXiv preprint*.

Thales Sales Almeida, Hugo Queiroz Abonizio, Rodrigo Nogueira, and Ramon Pires. 2024. Sabiá-2: A new generation of portuguese large language models. *arXiv preprint*.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Arthur Silva, Nathan Hartmann, Erick Fonseca, and Sandra Aluísio. 2018. Gigaverbo: A large-scale dataset of brazilian portuguese verb inflections. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, and 1 others. 2024. Qwen2 technical report. *arXiv preprint*. Qwen Team, Alibaba Group.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. TinyLLaMA: An open-source small language model. *arXiv preprint*.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, and 1 others. 2023. Large language models for scientific synthesis, inference and explanation. *arXiv preprint*.