FlexQwen: Exploring Hybrid Objectives and Text Originality for Portuguese

Miguel de Mello Carpi and Marcelo Finger
University of São Paulo, São Paulo, Brazil
{miguel,mfinger}@ime.usp.br

Abstract

While scaling laws suggest increasing model and dataset sizes for better results, efficient pre-training techniques for low-resource scenarios present unique challenges that require further investigation. This work introduces FlexQwen, a model based on the Qwen 3 architecture adapted for a hybrid causal-masked objective, and the Carolina Originality dataset, a subset of the Corpus Carolina tailored for efficient pre-training in Portuguese. We investigate two primary research questions: the influence of hybrid masked-causal modelling and the impact of text originality on model performance. Our experiments compare a high-originality *Gold* split against a length-matched control group. Results indicate that hybrid objectives may be viable for efficient training. Furthermore, we provide open access to our code, datasets, and training logs to foster further research in efficient Portuguese LLMs.

1 Introduction

Language-specific pre-training is essential for models obtaining linguistic, cultural and domain understanding (Souza et al., 2020; Rodrigues et al., 2023; Corrêa et al., 2025). Early work focused on the continued pre-training of multilingual models, but recent works explore increasing both model and dataset size, following the scaling laws (Kaplan et al., 2020). However, recent research into pre-training suggests that performance is often driven more by data quality and uniqueness than sheer volume. For instance, deduplicating training data has been shown to improve model performance and efficiency (Lee et al., 2022), while works such as Geiping and Goldstein (2023) demonstrate that competitive models can be trained on limited hardware by optimizing the training recipe and data throughput. Furthermore, techniques that harness diversity for data selection can significantly outperform random sampling in compute-constrained scenarios (Zhang et al., 2024). This is aligned with the findings of Samuel et al. (2023), who achieved results on par with BERT using only the 100 million words of the British National Corpus (BNC), emphasizing the importance of data diversity and training recipe over scale.

Motivated by these ideas, this work presents initial results of ongoing work in exploring the available techniques for efficient pre-training of large language models (LLMs) in low-resource languages. We share these results in the belief that they may benefit other research groups working on low-resource languages. The main contributions of this work are the *Carolina Originality* dataset and the `FlexQwen` model. The dataset is a subset of the Corpus Carolina (Finger et al., 2025) designed to benchmark how text originality influences pre-training efficiency in low-resource settings, while the model is an adaptation of Qwen 3 (Yang et al., 2025) modified to support a **hybrid causal-masked** objective (Charpentier and Samuel, 2024). The source code, the datasets and the training experiments settings are available in our repository[1].

In this work, our experiments are designed to answer the following research questions: **Q1** What is the influence of the hybrid masked-causal modelling training objective in a low-resource scenario with diverse texts? **Q2** Does increasing text originality above 85% make models better?

[1] https://github.com/mmcarpi/flexqwen

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 1079–1084

2 Method

For this work, we utilized the Carolina Corpus version 2.0 Bea[2], an open, human-curated corpus for Linguistics and Artificial Intelligence. It comprises approximately 1 billion tokens distributed across millions of texts of varied typology, all crawled from the Brazilian Portuguese web. Beyond its substantial volume, each document in Carolina includes rich metadata, enabling precise filtering for downstream applications. In our study, we focused on four key metadata fields: `originality`, `mime_type` (source file format), `tokens` (number of words), and `Carolina_typology`.

Introduced in version 1.2 Ada, the originality field quantifies the uniqueness of a text relative to the rest of the corpus. The score ranges from 0 to 100, where 0 indicates a near-duplicate of existing content and 100 indicates highly unique content. This value was computed using the Onion (ONe Instance ONly) algorithm (Pomikálek, 2011), as detailed in the methodology described by Serras et al. (2024)

Complementing originality is the `mime_type`, which indicates the format of the source document. This metadata is critical for quality control, as it allows us to isolate and remove documents originating from PDFs, which frequently suffer from extraction artifacts.

We also utilized the `tokens` metadata, which provides an approximate word count for each document.

Finally, we filtered by the `Carolina_typology` field. This metadata categorizes texts into distinct discourse domains. These categories represent the specific context or communicative situation in which the texts occur and were annotated to ensure the corpus encompasses multiple linguistic registers.

Construction of the Base Dataset Our initial filtering step refined the raw corpus into a cleaner *Base* dataset representing standard language usage. We applied two major exclusion criteria: removal of PDF-derived documents to mitigate extraction artifacts, and exclusion of legislative/judicial branches to avoid over-representation of "legalese".

High Originality Split We filtered the *Base* dataset for documents with an originality score ≥ 85 and sampled 0.5M documents. The threshold of 85 was chosen based on the distributional analysis of the Carolina Corpus, where it represents approximately the median originality score for the *Base* dataset. By selecting documents at or above this cut-off we ensure a dataset with sufficient volume of 220M tokens to our experiments. We refer to this high-originality split as *Gold*.

Matched Length Split Preliminary analysis of the Carolina Corpus revealed a strong positive Spearman correlation ($\rho = 0.68$) between document length and originality scores. This relantionship is characterized by shorter documents exhibiting higher variance and lower mean originality, while longer documents tending to converge toward higher scores. To isolate the impact of originality from these length-dependent dynamics, we created a control split using a 1-to-1 nearest-neighbour matching algorithm. This process selected documents from the Base dataset (excluding Gold members) that precisely mirror the token-count distribution of the *Gold* split. This resulted in two distributionally aligned training sets: *Gold* and *Matched* each containing 220M tokens, ensuring that any observed performance differences can be attributed to text originality rather than the confounding influence of document length.

Distributional Alignment To ensure the validity of our comparative analysis, we assessed the distributional alignment between the *Gold* (high originality) and *Matched* (control) splits. Although a Kolmogorov-Smirnov (KS) (Hodges, 1958) test rejected the null hypothesis ($p < 0.05$), this is expected given the test's extreme sensitivity to large sample sizes ($N \approx 100,000$). However, the effect size indicates a functionally perfect match: the KS statistic (D) was **0.009**, meaning the maximum divergence between the cumulative distribution functions is less than 1%. This result confirms that the *Matched* dataset preserves the natural variance of the *Base* corpus without introducing length bias or data duplication artifacts.

Tokenizer set Finally, we sampled *Tokenizer Train* (50M tokens) from the remaining Base texts.

[2] https://sites.usp.br/corpuscarolina

3 Model & Objective

We introduce **FlexQwen**, a hybrid model based on the Qwen3 architecture (Yang et al., 2025) adapted to support the masked-causal objective proposed by Charpentier and Samuel (2024). This approach replaces standard random masking with Masked Next-Token Prediction (MNTP) (BehnamGhader et al., 2024), where prediction labels are shifted to the $(i+1)$-th position to align with the Causal Language Modelling (CLM) target. This alignment enables a flexible forward pass: the model switches between causal (generative) and bidirectional (representation) modes solely by toggling the attention mask.

For tokenization, we trained a byte pair encoding (BPE) (Sennrich et al., 2016) tokenizer with a vocabulary of 64,000 tokens on the *Tokenizer Train* split. To optimize parameter efficiency, we tie the embedding and language modelling weights. The final architecture employs an embedding dimension $d_{model} = 768$, a feed-forward size $d_{ff} = 1536$, 12 Transformer blocks, 12 attention heads ($d_k = 64$) with 4 key-value (KV) groups for grouped-query attention (GQA) (Ainslie et al., 2023). This configuration results in approximately 110M trainable parameters.

Training For pre-training we chose the AdamW (Loshchilov and Hutter, 2018) optimizer. Following the protocol established by Brown et al. (2020), we applied a weight decay to all 2D weight matrices while excluding biases and normalization scales (Zhang and Sennrich, 2019) from decay to maintain training stability. During training we use cosine-decay (Loshchilov and Hutter, 2017) for adjusting the learning rate with a minimum learning rate of 0.1. For MNTP we combine with SpanBERT masking strategy (Joshi et al., 2020) with mean span length of 3, max span length of 10, and mask probability of 15%. Distinct of traditional BERT and SpanBERT style, we drop mask replacement with original or random tokens. The input sequence length is 512 tokens.

4 Experiments

Following the philosophy of the "Cramming" challenge proposed by Geiping and Goldstein (2023) , which seeks to maximize performance within a fixed compute budget, our experiments focus on data efficiency. We treat the 220M token limit as a fixed constraint to answer our research questions: **Q1** regarding the influence of the hybrid masked-causal modelling objective in low-resource scenarios, and **Q2** regarding the impact of text originality above the 85% threshold. To answer these, we conduct two ablation studies. For **Q1**, we compare the hybrid objective against a pure MNTP baseline. For **Q2**, we compare the *Gold* split against the *Matched* length-control split.

For **Q1** we employ two distinct hybrid-training curricula to mitigate potential forgetting of generative capabilities during the transition between objectives:

- **Shift** The model starts with the CLM objective for n steps before switching to the MNTP objective.
- **Linear** The model starts with a 1.0 probability of selecting the CLM objective, which linearly decreases to 0 over n steps as the MNTP probability increases to 1.0.

We hypothesize that the **linear** strategy will better preserve the model's generative ability compared to the hard switch observed in preliminary results. These strategies are compared against one baseline for each data split, that is a model trained solely on the MNTP objective.

To ensure a proper evaluation, we conduct a hyperparameter search using Optuna (Akiba et al., 2019) with Tree-Structured Parzen Estimator (TPE) (Watanabe, 2025). For this we use a sample of 16M tokens from each training set and use the corresponding validation loss to select the optimal configuration. For each setting, we conduct 25 trials (including 5 startup trials) to optimize: AdamW's learning rate (1×10^{-6} – 1×10^{-2}), β_1 (0.8 – 0.99), weight decay (1×10^{-5} – 1×10^{-2}), and ϵ (1×10^{-9} – 1×10^{-6}) parameters. We also optimize fraction of total number steps n used to switch or regulate the objective probabilities (0.25 – 0.75), and the warm up steps for the cosine-decay scheduler (0.0 – 0.2). Finally, each model is trained for the full 220M tokens.

5 Results

Table 1 details the optimal hyperparameters obtained. We observe that n consistently occu-

Table 1: Hyperparameter configurations and final performance metrics. n denotes the switch duration for hybrid strategies. Baseline (Masked) runs omit the n parameter. The losses are based on the 16M token subset.

		Hyperparameters					Loss Metrics		
Split	**Strategy**	**LR**	ϵ	β_1	**WD**	n	**HPO**	**Train**	**Val**
Gold	Shift	5.56e-4	7.22e-8	0.89	4.03e-3	0.68	4.68	3.77	3.71
	Linear	4.19e-4	2.97e-8	0.93	4.99e-4	0.67	5.43	3.87	3.64
	Masked	2.40e-4	8.96e-9	0.82	2.92e-4	—	5.83	3.98	3.76
Matched	Shift	6.95e-4	1.04e-9	0.80	3.31e-4	0.67	4.43	3.58	3.66
	Linear	3.47e-4	2.23e-7	0.87	7.18e-4	0.74	5.18	3.49	3.49
	Masked	3.52e-4	6.07e-9	0.80	1.37e-4	—	5.45	3.82	3.63

pies more than half of the training tokens (e.g., > 0.60 for *Gold*). Combined with the high values for learning rates, we can see that the optimization prioritized rapid loss minimization. We now proceed to analyse our research questions.

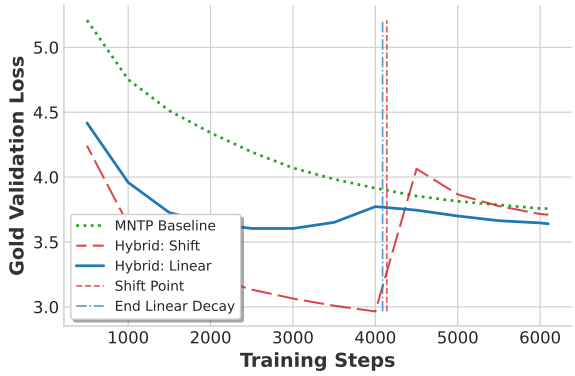

Figure 1: Validation loss on *Gold* split. The abrupt *Shift* strategy (red) causes a catastrophic loss, whereas the *Linear* decay schedule (blue) maintains stable learning, outperforming the pure MNTP baseline (green).

Impact of Hybrid Objective (Q1) Figure 1 shows that a significant degradation in performance occurs after the objective switch when using the *shift* strategy. For instance, on the *Gold* split, we can see the loss abruptly increased when the shift happened and did not recovered before the training finished. This indicates that without a cooling-down period or a more gradual mixing ratio, the model suffers from catastrophic forgetting when forced to focus solely on the masked objective. A similar trend occurs with the *linear* strategy; however it is less intense due to the model being trained progressively more on the MNTP objective. Despite this loss increase on objective switching, the Hybrid approach did succeed in lowering the MNTP loss compared to pure MNTP baselines.

Impact of Originality (Q2) The final loss values and overall optimal hyperparameters are similar across models indicating that the dynamics do not change much between the splits. Nevertheless, we notice that in the *Matched* split with the *Linear* strategy n was the highest together with the lowest loss in train and validation sets, which may indicate that the linear strategy benefits from a longer decay.

6 Conclusion

This paper presented initial steps toward developing efficient language models for Brazilian Portuguese, introducing **FlexQwen** and the **Carolina Originality** dataset.

Our ablation studies answer our research questions with two key insights. Regarding **Q2 (Originality)**, we could not confirm that a high originality (score ≥ 85) improves model performance faster than an diverse control sample. However, it is important to note that Corpus Carolina is a curated dataset and the texts have a high originality score. In this regard, this question needs more data and experiments to be fully answered.

Regarding **Q1 (Hybrid Objective)**, our results highlight the role of curriculum learning. We observed that an abrupt performance drop occurs during the objective transition. Nonetheless, the proposed *linear* schedule demonstrates that a gradual mixing of the objectives mitigates this instability, though it is not entirely eliminated.

7 Limitations

While our length-matched experiments isolate the impact of document length, the primary

limitation of this study is the small scale of pre-training (220M tokens). The findings presented here must be verified at larger compute budgets. Furthermore, while we showed that the hybrid objective is more efficient than the MNTP, it is not by a large margin. Future work should investigate different strategies for avoiding degradation and model behaviour when scaling to more tokens.

Additionally, because the Carolina Corpus encompasses multiple discourse typologies, future analysis is necessary to determine if filtering for high originality inadvertently biases the model toward specific domains. Finally, our current evaluation relies primarily on loss on held-out test sets. Future work requires fine-tuning on downstream Portuguese benchmarks, such as ASSIN 2 (Real et al., 2020) and HateBR (Vargas et al., 2022), to fully assess the practical utility of these representations.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support from the University of São Paulo and the São Paulo Research Foundation (FAPESP) (grant 2019/07665-4). Marcelo Finger was partly supported by FAPESP (grants 2023/00488-5 and 2022/11254-2) and the National Council for Scientific and Technological Development (CNPq) (grant PQ1 302963/2022-7). Computational resources from the Multi-User Equipment were used thanks to FAPESP project 19/26702-8. The authors also acknowledge the support of the National Center for High-Performance Computing in São Paulo (CENAPAD-SP), a joint project between UNICAMP and FINEP - MCTI. We are grateful to the reviewers for their insightful suggestions regarding experiments and related works. Finally, the authors used generative AI tools to assist with writing (paraphrasing and language refinement) and code development; all AI-generated suggestions were verified by the authors.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. *Preprint*, arXiv:2305.13245.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 2623–2631, New York, NY, USA. Association for Computing Machinery.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders. *Preprint*, arXiv:2404.05961.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Lucas Georges Gabriel Charpentier and David Samuel. 2024. GPT or BERT: Why not both? In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 262–283, Miami, FL, USA. Association for Computational Linguistics.

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

Marcelo Finger, Maria Clara Paixão de Sousa, Cristiane Namiuti, Vanessa Martins do Monte, Aline Silva Costa, Felipe Ribas Serras, Mariana Lourenço Sturzeneker, Miguel de Mello Carpi, Mayara Feliciano Palma, and Gabriela Alves Lachi. 2025. Building Carolina: Metadata for Provenance and Typology in a Corpus of Contemporary Brazilian Portuguese. *Cadernos de Linguística*, 6(4):e812–e812.

Jonas Geiping and Tom Goldstein. 2023. Cramming: Training a Language Model on a single GPU in one day. In *Proceedings of the 40th International Conference on Machine Learning*, pages 11117–11143. PMLR.

J. L. Hodges. 1958. The significance probability of the smirnov two-sample test. *Arkiv för Matematik*, 3(5):469–486.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020.

SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *Preprint*, arXiv:2001.08361.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating Training Data Makes Language Models Better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. *Preprint*, arXiv:1608.03983.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Jan Pomikálek. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora*. Disertační práce, Masarykova univerzita, Fakulta informatiky.

Livy Real, Erick Fonseca, and Hugo Gonçalo Oliveira. 2020. The ASSIN 2 Shared Task: A Quick Overview. In *Computational Processing of the Portuguese Language*, pages 406–412, Cham. Springer International Publishing.

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-*. In *Progress in Artificial Intelligence: 22nd EPIA Conference on Artificial Intelligence, EPIA 2023, Faial Island, Azores, September 5–8, 2023, Proceedings, Part I*, pages 441–453, Berlin, Heidelberg. Springer-Verlag.

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Felipe Ribas Serras, Mariana Sturzeneker, Miguel de Mello Carpi, Mayara Feliciano Palma, Maria Clara Ramos Morales Crespo, Aline Silva Costa, Vanessa Martins Do Monte, Cristiane Namiuti, Maria Clara Paixão de Souza, and Marcelo Finger. 2024. Exploring Computational Discernibility of Discourse Domains in Brazilian Portuguese within the Carolina Corpus. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 255–265, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

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.

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, Marseille, France. European Language Resources Association.

Shuhei Watanabe. 2025. Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance. *Preprint*, arXiv:2304.11127.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *Preprint*, arXiv:2505.09388.

Biao Zhang and Rico Sennrich. 2019. Root Mean Square Layer Normalization. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Chi Zhang, Huaping Zhong, Kuan Zhang, Chengliang Chai, Rui Wang, Xinlin Zhuang, Tianyi Bai, Qiu Jiantao, Lei Cao, Ju Fan, Ye Yuan, Guoren Wang, and Conghui He. 2024. Harnessing Diversity for Important Data Selection in Pretraining Large Language Models. In *The Thirteenth International Conference on Learning Representations*.