A proposal for post-OCR spelling correction using Language Models

Sávio Santos de Araújo
Universidade de Pernambuco
Garanhuns, Pernambuco, Brazil
savio.santos@upe.br

Byron Leite Dantas Bezerra
Universidade de Pernambuco
Recife, Pernambuco, Brazil
byron.leite@upe.br

Arthur Flor de Sousa Neto
Universidade de Pernambuco
Recife, Pernambuco, Brazil
afsn@ecomp.poli.br

Cleber Zanchettin
Universidade Federal de Pernambuco
Recife, Pernambuco, Brazil
cz@cin.ufpe.br

Abstract

This work explores the use of Language Models (LMs) to correct residual errors in texts extracted by OCR and HTR (Handwritten Text Recognition) systems. We propose a general approach but utilize the images from Brazilian handwritten essays of the BRESSAY dataset as a use case. Two standard LMs (Bart and ByT5) and two LLMs (LLama 1 and LLama 2) were evaluated in this context. The results indicate that the smaller LMs outperformed the LLMs in terms of error rate reduction (CER and WER). Traditional correction methods, such as Symspell and Norvig, were influential in some cases but fell short of the results obtained by the LMs. ByT5 with byte-level tokenization improved CER and WER, proving performance for texts with high noise. As a result, smaller LMs, after fine-tuning, are more efficient and cheaper for post-OCR corrections. We identify and propose promising future studies involving correction at broader levels of context, such as paragraphs. Code is available at https://github.com/savi8sant8s/ptbr-post-ocr-sc-llm.

1 Introduction

Accuracy in text extraction using optical character recognition (OCR) techniques is beneficial for ensuring and promoting the preservation of archives in different contexts. The more accurate the OCR result, the lower the cost, time, and need for human intervention to review and correct residual errors. This, in turn, increases efficiency and optimizes the use of computational, financial, and human resources. Therefore, using tools that act on these residual errors is one way to improve the reliability of these solutions.

Using spelling correctors in the post-OCR stage improves word and character recognition rates in OCR and HTR (Handwritten Text Recognition) systems. Statistical spelling correction approaches such as N-Gram, Symspell, and Norvig are the traditional strategies for this task (1) and (2). Furthermore, new approaches also utilize neural networks with the Transformers architecture (3), such as LLama, Bart, and ByT5, were explored in (1), (2), (4), and (5), respectively.

The recent intensification of studies and applications of open Large Language Models (LLMs) such as LLama (6) (7) (8) and Mistral (9) have shown the potential of these big models for various applications. Recent works such as (4) and (10) have explored the use of LLMs in the text correction task to identify whether they can outperform statistical approaches and models with fewer parameters.

Latinx in AI @ NeurIPS-24 (LXAI 2024).

This paper explores the potential of open Language Models (LMs) to correct residual errors in texts extracted by different OCR/HTR systems. We use three well-established optical models (Bluche (11), Flor (12), and Puigcerver (13)), the The Azure OCR system (v. 3.2) developed by Microsoft [1], and four state-of-the-art HTR systems submitted to the *ICDAR 2024 Competition on Handwritten Text Recognition in Brazilian Essays - BRESSAY* (14). Besides state-of-the-art results, all these OCR/HTR systems have spelling errors in the text extraction. We also compared the performance of the LMs with traditional correction methods.

The results revealed the potential of using smaller LMs instead of LLMs, as they achieved superior performance and less computational costs, especially when the documents have complex layouts and high recognition error rates.

The main contributions of this work are:

1. We proposed a fine-tuning approach to train the LMs on how to correct texts extracted from OCR/HTR systems;
2. We proposed a methodology to prepare the dataset and prompts to train the models with input texts considering real post-OCR errors and simulated synthetic texts;
3. We showed the impact of LMs in different OCR/HTR systems, highlighting when this approach can improve the recognition results, even considering modern LLMs;
4. We demonstrated the benefits of LMs against traditional spelling correction methods.

2 Related works

Studies aiming to use LLMs for various purposes have recently been boosted by the availability of open-source models with relatively smaller sizes compared to existing larger but efficient models, such as LLaMa (6) (7) (8) and Mistral (9), which provide ways to adapt them to different purposes at a small cost in time and using machines with relatively affordable computing power. This has made it possible to make comparisons of these models on different tasks and compare their performance against closed models, such as OpenAI's GPT models [2], one of the most famous and widely used LLMs on the market today.

In addition, the research (15) evaluates Chinese text correction, Chinese spelling correction, and English grammar correction using the *few-shot* training method with OpenAI's closed LLM GPT-3.5 Turbo. With this, the research aims to gather insights into the meaning of correction in the era of LLMs and its implications for various natural language processing applications. Our work differs from the above research since we focus on different languages (Portuguese in this paper) and spelling correction.

A third research (16) analyzes the performance of OpenAI's GPT 3.5 and GPT 4 models in the task of grammatical correction of texts compared with the same functionalities provided by Google Docs and Microsoft Word office tools. This previous work considered the Portuguese language. However, it focuses on grammatical correction and uses only closed models in its analysis, unlike our proposal, which focuses on post-OCR spelling correction with open models.

The most recent research (17) proposes reducing the research gap by analyzing the LLMs as grammatical error correctors. The research applies several evaluation criteria analyzing the efficiency of three LLMs - one open (LLaMa 2) and two closed (GPT 3.5 and GPT 4) - in the task in question. We focus on spelling corrections using only open models.

Another work (18) evaluated the use of LLMs for post-OCR correction of historical texts transcribed in several European languages (English, French, German). The study used the *zero-shot* and *few-shot* techniques to perform the corrections using the LLMs. It demonstrated that the LLMs performed poorly in the applied scenario in any challenges. Our approach focuses on the fine-tuning technique to specialize a model and attempt to overcome the limitations of the techniques used in the above study.

The study of (10) evaluates the usage of an LLM and a relatively minor LM for post-OCR correction in English. The LLama (LLM) and ByT5 (LM) models were used to compare which would perform

[1] `https://azure.microsoft.com/products/ai-services/ai-vision/`

[2] `https://platform.openai.com/docs/models`

best in the challenge, and the texts were converted to lowercase to obtain the best performance. The work demonstrated that the ByT5 model performed better than a larger model. It also identified that using the *few-shot* technique was insufficient for LLama to perform well in this task. Moreover, the ByT5 model performed best using a context of 50 characters (no more and no less) and without fine-tuning. Our paper focuses on the fine-tuning technique to confirm the impact of specializing in an LLM for correction without converting the data and with a more variable context of text length.

3 Methodology

As presented in Figure 1, the methodology adopted in our work was based on five stages presented in the following sections.

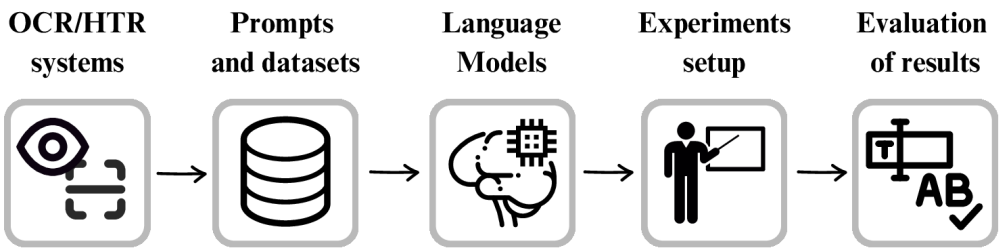


Figure 1: Methodology process.

3.1 Selected OCR/HTR systems

The following OCR/HTR systems were selected in our analysis to produce the recognition results:

- Three well-established optical models highly used to recognize handwritten lines in several datasets, including the BRESSAY dataset: Bluche (11), Flor (12), and Puigcerver (13);
- Four state-of-the-art HTR systems submitted to the *ICDAR 2024 Competition on Handwritten Text Recognition in Brazilian Essays - BRESSAY* (14): LITIS, LTU, LTU Ensemble, Pero and Demokritos;
- The Azure OCR system developed by Microsoft 1.

The HTR systems developed on ICDAR 2024 receive the names of their teams. The Demokritos team (Athens, Greece) comprised C. Vossos, K. Palaiologos, and E. Sarafoglou from the National Centre of Scientific Research Demokritos. The LITIS team (Rouen, France) comprised L. Hamdi[a], T. Simon[a], T. Constum[a], T. Paquet[a], P. Tranouez[a], and C. Chatelain[b] from the University of Rouen Normandy[a] and INSA Rouen Normandy[b]. The LTU team (Luleå, Sweden) comprised S. Corbillé, C. Liu, and E. H. B. Smith from the Luleå University of Technology (*Luleå Tekniska Universitet*). The Pero team (Brno, Czech Republic) comprised M. Kišš, M. Hradiš, K. Beneš, and J. Kohút from the Faculty of Information Technology, Brno University of Technology.

3.2 Prompts and datasets

To fine-tune the LMs, training prompts were created containing instructions (intention to correct the text), input (text with errors), and output (corrected text). The experiments used post-OCR texts and simulated synthetic texts. Table 1 presents the used prompts templates to fine-tune the LMs.

The real prompts were created from the BRESSAY dataset (19), which contains images of handwritten essays with content, structural rules, and themes similar to those used in the Brazilian national high school exam. For each OCR/HTR system described in Section 3.1, we run it on the BRESSAY dataset to produce a new dataset for each model containing the post-OCR outputs in the line-level text format. For training purposes, only the training and validation partitions were used. The test partition of the BRESSAY dataset was used to evaluate the OCR/HTR systems in this work.

Models	Template
Bart Pt and ByT5 Pt	Correct: {input}
Sabiá and Gervásio	### Instruction: {instruction} ### Input: {input} ### Response:

Table 1: Prompt templates used in the fine-tuning process. Gervásio and Sabiá require a more complete prompt. The LMs Bart and ByT5 can use a smaller prompt.

The main challenges OCR/HTR solutions faced in recognizing text from this dataset were variations in the handwriting of the authors who wrote the essays, paper textures, ink quality, different image resolutions, erasures, inconsistencies, and different styles. Figure 2 presents samples of images of lines that confirm this variety of characteristics present in the dataset.

In addition, synthetic prompts were created using the Essay-BR dataset (20), a corpus of texts written in the Brazilian national exam model format with a textual structure similar to those found in the BRESSAY dataset. According to (21), insertions, deletions, substitutions, and transpositions are the most frequent errors in OCR systems' output. Therefore, these errors were simulated in the synthetic prompts using the Python NoisOCR library [3].

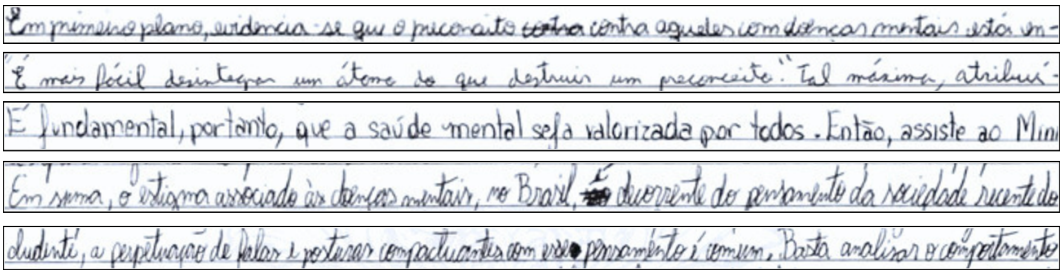


Figure 2: Line level examples from the BRESSAY dataset (19).

3.3 Selected LMs

Four criteria were used to select the LMs for the first experiment: be pre-trained in Brazilian Portuguese, be state-of-the-art models, have your base models cited in previous works, and be open for use and training. Thus, four LMs (two LLMs and two standard LMs) were selected. Table 2 presents the models, and each LM is explained below.

Language Model	Base model / Year	Params	Train dataset
Sabiá	LLama 1 (2023)	7B	ClueWeb 2022 (22)
Gervásio	LLama 2 (2023)	7B	ExtraGLUE (23)
Bart Base Portuguese	Bart (2019)	139M	BrWac Corpus (24)
ByT5 Small Portuguese	ByT5 (2021)	330M	Squad v1.1 (25)

Table 2: LMs used in this work. All of them were trained from scratch or fine-tuned with a Portuguese dataset on models trained primarily in English. In this work, fine-tuning was done in each LM to teach them how to make post-OCR spelling corrections in Portuguese manuscripts.

3.3.1 Byt5 Small Portuguese (ByT5)

ByT5 Portuguese[4] is a pre-trained language model on top of Google's ByT5 model with 330 million parameters for the Portuguese language, using the Squad Portuguese v1.1 dataset (25). This pre-trained model was publicly available on Hugging Face [5].

[3] https://github.com/savi8sant8s/noisocr

[4] https://huggingface.co/pierreguillou/byt5-small-qa-squad-v1.1-portuguese

[5] https://huggingface.co/models

3.3.2 Bart Base Portuguese (Bart)

Bart Portuguese[6] is a pre-trained language model on top of Meta's Bart model with 139 million parameters for the Portuguese language, also using the BrWac Corpus. This pre-trained model was made publicly available on Hugging Face5.

3.3.3 Sabiá (LLaMa 1)

Sabiá (26) is a set of Large Language Models pre-trained in Portuguese by the company Maritaca AI on top of the LLaMa from Meta and GPT-J models from EleutherAI using the ClueWeb 2022 corpus (22). According to the company, its best model, Sabiá-65B (about 65 billion parameters), performs equivalent to OpenAI's GPT-3.5 Turbo closed model. The selected model for fine-tuning was Sabia-7b, which was made openly available to the public and was trained on the first version of LLaMa, with 7 billion parameters, and trained in Portuguese with a subset of ClueWeb 2022.

3.3.4 Gervásio (LLaMa 2)

Gervásio (27) is a project carried out by the University of Lisbon, Portugal, that trained Meta's 7 billion parameter LLaMa 2 model for the Portuguese language. The model was trained using the ExtraGLUE (23) dataset, which is a translated version of the Glue (General Language Understanding Evaluation) dataset, widely used to evaluate natural language understanding challenges. Two models were available. Gervásio 7B PTBR [7] was chosen.

3.4 Experiments setup

All experiments were run in a Linux environment with an Nvidia RTX 4060 TI GPU and 16 GB of GPU memory. The work used two types of fine-tuning: complete and quantized (28). Complete fine-tuning was used on the selected language models that were relatively smaller (with millions of parameters) because it was possible to train them quickly and completely without having to reduce the precision of the data. Quantized fine-tuning was performed on the selected LLMs (with billions of parameters) due to the computational cost required to fine-tune them completely. The Q-Lora technique (29) with 4-bit quantization was used, reducing the computational capacity and time needed for training but preserving performance.

The image lines of training and validation partitions of the BRESSAY dataset were processed through each OCR/HTR system described in Section 3.1. Then, the recognition results were used to fine-tune the LMs and learn how to fix the recognition errors produced by each recognizer.

Each prompted dataset created from the outputs of some recognizer has 24,164 rows composed of the particular recognition errors produced by such a system. All subsets were merged with a synthetic dataset created from Essay-BR to create a unified dataset. This resulted in a total training dataset with 232,635 prompts.

The LMs were fine-tuned using the unified dataset, where 90% was used for training and 10% for validation. Complete fine-tuning was done using ByT5 Portuguese and Bart Portuguese models trained for ten epochs. Quantized fine-tuning was done on Sabiá and Gervásio, who were trained for one epoch due to the longer training time after increased training data. After preliminary experiments, the learning rate for all fine-tunings was set as 0.0001, the maximum token length was 400 for ByT5 (tokenizing at the byte level generates a larger output), 250 for the others, and the temperature was defined as 0.0001 for Sabiá and Gervásio.

The Symspell, Norvig [8] and N-Gram (30) algorithms were included for comparison purposes with the four LMs. A word frequency dictionary (for Symspell and Norvig) and a word dictionary (for N-Gram) were created for the algorithms. Norvig has a similar correction strategy for Symspell, but the Symspell authors demonstrated that they overcame it.

To compare the efficiency of the fine-tuned LMs and the traditional approaches in the spelling correction task across multiple challenges, we submit the image lines of the BRESSAY's test partition

[6] `https://huggingface.co/adalbertojunior/bart-base-portuguese`

[7] `https://huggingface.co/PORTULAN/gervasio-7b-portuguese-ptbr-decoder`

[8] https://norvig.com/spell-correct.html

OCR	Baseline	ByT5 Pt	Bart Pt	Symspell	Norvig	N-Gram	Sabiá	Gervásio
Bluche	18,02	**15,97**	19,65	17,14	18,16	20,63	27,34	24,31
Flor	10,26	**9,03**	10,92	9,84	10,9	11,91	14,4	13,43
Puigcerver	**8,94**	9,00	10,28	9,23	10,02	10,19	12,57	11,32
LITIS	**4,62**	5,01	5,11	5,23	5,93	5,36	5,75	5,46
LTU	**3,35**	3,66	4,01	3,9	4,61	4,20	4,59	4,37
LTU ENS.	**3,2**	3,51	3,82	3,77	4,48	4,05	4,44	4,21
Pero	**2,88**	3,24	3,53	3,36	4,18	3,63	4,46	3,73
Demokritos	8,21	**7,63**	9,00	8,13	9,03	9,68	11,57	11,68
Azure	12,43	**11,98**	12,45	12,76	13,68	13,67	13,42	13,05

Table 3: CER metrics of the Baseline (based on the OCR output only) and all the correction approaches. The rates in bold are the best results achieved (the lower, the better).

OCR	Baseline	ByT5 Pt	Bart Pt	Symspell	Norvig	N-Gram	Sabiá	Gervásio
Bluche	48,56	**29,11**	32,89	39,69	42,85	41,83	43,05	40,08
Flor	31,58	**18,25**	20,13	25,99	29,43	26,54	24,9	24,3
Puigcerver	25,07	**18,35**	19,22	23,41	26,32	22,63	22,39	21,8
LITIS	11,91	11,62	**11,31**	14,35	17,19	12,61	11,85	12,03
LTU	10,51	9,03	**8,91**	12,37	15,16	10,88	9,74	9,83
LTU ENS.	10,12	8,74	**8,58**	12,08	14,84	10,57	9,48	9,40
Pero	9,39	8,30	**8,11**	11,07	14,32	9,95	9,11	8,94
Demokritos	25,12	**15,99**	17,14	22,16	25,07	22,01	20,29	20,09
Azure	21,83	**17,54**	17,87	22,63	26,34	22,39	19,28	19,22

Table 4: WER metrics of the Baseline (based on the OCR output only) and all the correction approaches. The rates in bold are the best results achieved (the lower, the better).

to each OCR/HTR system described in section 3.1 and post-process the text outputs produced, in order to measure the rates of each spelling correction approach previously discussed.

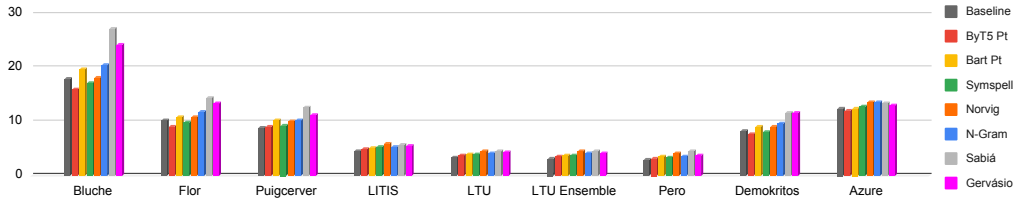
3.5 Evaluation of results

Two metrics were used to evaluate whether language models can correct the remaining errors in OCR output, thus improving the quality of the extracted texts: Character Error Rate (CER) and Word Error Rate (WER), which are the most common evaluation metrics for OCR and HTR systems (31). The metrics use the Edit Distance algorithm (32), which calculates the similarity between two strings by counting the minimum number of edits (insertions, deletions, and substitutions) to transform one string into another.

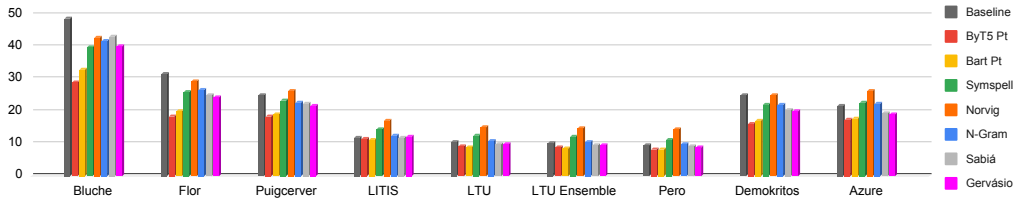
$$CER = \frac{\sum_{i=1}^{N} ED(C_{G_i}, C_{R_i})}{\sum_{i=1}^{N} |C_{G_i}|} \tag{1}$$

$$WER = \frac{\sum_{i=1}^{N} ED(W_{G_i}, W_{R_i})}{\sum_{i=1}^{N} |W_{G_i}|} \tag{2}$$

The CER metric (Equation 1) calculates the minimum number of edits at the character level, where $ED(C_{R_i}, C_{G_i})$ represents the edit distance between the recognized character C_{R_i} and the ground truth character C_{G_i} for the i-th data point. Meanwhile, the WER metric (Equation 2) makes the same assessment but at the word level, where $ED(W_{R_i}, W_{G_i})$ denotes the edit distance between the recognized word W_{R_i} and the ground truth word W_{G_i} for the i-th data point. The terms $|C_{G_i}|$ and $|W_{G_i}|$ indicate the number of characters and words, respectively, in the ground truth sequences for each i-th data point. The variable N represents the total number of samples considered in the evaluation. They are used to compare the metrics before and after fine-tuning and evaluate the models in the spelling correction task.



(a) CER metrics



(b) WER metrics

Figure 3: Grouped bar chart of CER and WER metrics for baselines (outputs of OCR solutions) and correction approaches (the lower, the better).

After the predictions were collected, the CER and WER metrics of the baselines (predictions of the OCR/HTR systems) of each fine-tuned LM and for traditional correction approaches were calculated for each line. Finally, the average of the metrics was calculated to compare the results.

4 Results and Discussion

Tables 3 and 4 and Figure 3 present the CER and WER metrics of the OCR/HTR systems isolated (baseline), the fine-tuned LMs, and the traditional spelling correction methods. Additionally, the Appendix A presents an example of an image line with errors and the suggestions of each LM and traditional correction method. Below we present the main findings of the paper.

Traditional Approaches Still Hold Value
Surprisingly, traditional correction algorithms such as Symspell performed competitively against state-of-the-art language models in several cases. For instance, it significantly reduced CER in noisy outputs from the Bluche and Demokritos systems. While advanced language models are expanding the possibilities of text post-processing, the traditional algorithms still are competitive, when simpler and more affordable solutions are needed.

Performance on Noisy Scenarios
ByT5 Portuguese consistently outperformed other models in scenarios where OCR systems produced noisy outputs, such as with the Bluche and Demokritos datasets. In these challenging environments, the model obtained the lowest CER and WER. The performance presented by the byte-level tokenization models as ByT5 offers a significant advantage in handling the variability and complexity found in handwritten texts, especially when dealing with high levels of noise.

Efficiency of Smaller Models
The superior performance of smaller models like ByT5 and Bart Portuguese when compared to LLMs such as Sabiá and Gervásio is one of the interesting findings from this study. Despite the computational advantages of smaller models, they delivered comparable, if not better, results in most of the scenarios. The models ByT5 Portuguese and Bart Portuguese got lower WERs across different OCR systems. The results show that model size is not always indicative of performance in post-OCR correction tasks.

Limitations in Correcting High-Accuracy OCR Outputs
Considering the performance of OCR systems such as LTU and Pero — where the baseline CER and WER were already low — none of the correction methods, including the fine-tuned language

models, showed significant improvements. This suggests that for high-accuracy OCR outputs, the current generation of correction models may have reached a plateau, and further advancements may require new approaches that go beyond line-level corrections, potentially incorporating sentence- or paragraph-level context to achieve more meaningful improvements.

The Advantage of Byte-Level Tokenization
Another interesting insight from the results is the advantage of byte-level tokenization in handling complex error patterns of handwritten texts. The ByT5 model consistently outperformed other models. It is an interesting contrast to models using word- or subword-level tokenization, which primarily improved WER. The experiments show that byte-level tokenization offers better granularity to handle the character-level distortions common in OCR information. This property is particularly important for post-OCR corrections in noisy and diverse datasets such as BRESSAY.

Scalability and Computational Cost Efficiency
The relation between scalability and computational cost-efficiency is also interesting. The insight is clear when comparing smaller models like ByT5 and Bart to larger language models such as Sabiá and Gervásio. The model ByT5 achieved superior performance in correcting OCR outputs but it required significantly fewer computational resources. The computational cost advantage of using smaller but well-fine-tuned models is important to practical applications because it is the reality of scaling up to real-world data volumes. In resource-constrained environments or when deploying models for widespread use, smaller models like ByT5 offer a more sustainable and scalable solution without compromising accuracy.

5 Conclusion

We investigated four language models to improve the quality of text recognition after the OCR stage. We used Portuguese manuscripts as a use case for this approach. To teach LMs correct texts, we created a dataset with real texts with errors from the BRESSAY dataset and synthetic texts with errors from Essay-BR.

The results indicate that the small LMs Bart and ByT5 stood out compared to the bigger LLMs, highlighting the ByT5 model, the only LM that reduced the CER metric and had better results on the WER metric. The other LMs tested, including Bart, only managed to improve the WER metric and did so in all experiments performed.

Regarding the traditional correction methods, Symspell, followed by Norvig, reduced the CER metric in some challenges but below the best results obtained by ByT5. N-Gram, on the other hand, was unable to reduce this rate. In the WER metric, N-Gram, Symspell, and Norvig improved the rate in some challenges, but the results were below those obtained by the language models.

In the CER metric, ByT5 was the only LM able to improve the rate, obtaining better results in 5 out of 9 challenges submitted. ByT5 also stood out in the WER metric, followed by Bart, where these models improved the WER metrics in all challenges.

After all these analyses, it was possible to conclude that the language models obtained better results when submitted to texts with more errors. It was also identified that the LMs with word/subword-level tokenization only improved the WER metric. In contrast, the ByT5 model with byte-level tokenization improved both rates, confirming that the model is ideal for spelling correction tasks.

The smaller language models obtained better results than the LLMs, standing out for their low computational training cost and rapid adaptation to the data, indicating that adding, for example, the best model from the ByT5 experiment to perform spelling corrections in OCR models/systems is beneficial. Text recognition activities that generate medium from high noise level OCR outputs can benefit from using the model to improve output quality.

Future work could explore using models with a byte-level tokenization architecture and prompt datasets richer in single words and noise variations. Explore different languages is also important. Furthermore, exploring correction at the sentence, paragraph, and page level —instead of the line level explored in this work — could help language models generate more refined outputs by increasing their understanding of context.

Acknowledgments. This study was financed by the founding public agencies: CNPq, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and FACEPE (APQ-1216-1.03/22). In addition, we acknowledge all support from Di2Win (`www.di2win.com`) during the development of this work.

References

[1] A. F. d. S. Neto, B. L. D. Bezerra, and A. H. Toselli, "Towards the natural language processing as spelling correction for offline handwritten text recognition systems," *Applied Sciences*, vol. 10, no. 21, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/21/7711

[2] D. S. V. et al., "socrates - a post-ocr text correction method," in *Anais do XXXVI Simpósio Brasileiro de Bancos de Dados*. Porto Alegre, RS, Brasil: SBC, 2021, pp. 61–72. [Online]. Available: https://sol.sbc.org.br/index.php/sbbd/article/view/17866

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[4] A. Thomas, R. Gaizauskas, and H. Lu, "Leveraging LLMs for post-OCR correction of historical newspapers," in *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, R. Sprugnoli and M. Passarotti, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 116–121. [Online]. Available: https://aclanthology.org/2024.lt4hala-1.14

[5] A. Maheshwari, N. Singh, A. Krishna, and G. Ramakrishnan, "A benchmark and dataset for post-ocr text correction in sanskrit," 2022. [Online]. Available: https://arxiv.org/abs/2211.07980

[6] H. T. et al., "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[7] ——, "Llama 2: Open foundation and fine-tuned chat models," 2023. [Online]. Available: https://arxiv.org/abs/2307.09288

[8] A. D. et al., "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[9] A. Q. J. et al., "Mistral 7b," 2023. [Online]. Available: https://arxiv.org/abs/2310.06825

[10] M. Veninga, "Llms for ocr post-correction," July 2024. [Online]. Available: http://essay.utwente.nl/102117/

[11] T. Bluche and R. Messina, "Gated convolutional recurrent neural networks for multilingual handwriting recognition," *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pp. 646–651, 11 2017.

[12] A. F. S. Neto, B. L. D. Bezerra, A. H. Toselli, and E. B. Lima, "A robust handwritten recognition system for learning on different data restriction scenarios," *Pattern Recognition Letters*, vol. 1, pp. 1–7, 4 2022.

[13] J. Puigcerver, "Are multidimensional recurrent layers really necessary for handwritten text recognition?" *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pp. 67–72, 11 2017.

[14] A. F. S. Neto, B. L. D. Bezerra, S. S. Araújo, W. M. A. S. Souza, K. F. Alves, M. F. Oliveira, S. V. S. Lins, H. J. F. Hazin, P. H. V. Rocha, and A. H. Toselli, "Icdar 2024 competition on handwritten text recognition in brazilian essays – bressay," in *Document Analysis and Recognition - ICDAR 2024*, E. H. Barney Smith, M. Liwicki, and L. Peng, Eds. Cham: Springer Nature Switzerland, 2024, pp. 345–362.

[15] X. Zhang, X. Zhang, C. Yang, H. Yan, and X. Qiu, "Does correction remain a problem for large language models?" 2023. [Online]. Available: https://arxiv.org/abs/2308.01776

[16] M. C. Penteado and F. Perez, "Evaluating gpt-3.5 and gpt-4 on grammatical error correction for brazilian portuguese," 2023. [Online]. Available: https://arxiv.org/abs/2306.15788

[17] M. Kobayashi, M. Mita, and M. Komachi, "Large language models are state-of-the-art evaluator for grammatical error correction," 2024. [Online]. Available: https://arxiv.org/abs/2403.17540

[18] E. Boros, M. Ehrmann, M. Romanello, S. Najem-Meyer, and F. Kaplan, "Post-correction of historical text transcripts with large language models: An exploratory study," *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, p. 133–159, feb 2024. [Online]. Available: https://infoscience.epfl.ch/handle/20.500.14299/203985

[19] A. Neto, B. Bezerra, S. Araujo, W. Souza, K. Alves, M. Oliveira, S. Lins, H. Hazin, P. Rocha, and A. Toselli, "Bressay: A brazilian portuguese dataset for offline handwritten text recognition," in *18th International Conference on Document Analysis and Recognition (ICDAR)*. Springer, September 2024.

[20] J. Marinho, R. Anchiêta, and R. Moura, "Essay-br: a brazilian corpus of essays," in *Anais do III Dataset Showcase Workshop*. Online: Sociedade Brasileira de Computação, 2021, pp. 53–64. [Online]. Available: https://sol.sbc.org.br/index.php/dsw/article/view/17414

[21] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Communications of the ACM*, vol. 7, no. 3, pp. 171–176, 1964.

[22] A. Overwijk, C. Xiong, X. Liu, C. VandenBerg, and J. Callan, "Clueweb22: 10 billion web documents with visual and semantic information," 2022.

[23] T. O. et al., "Portulan extraglue datasets and models: Kick-starting a benchmark for the neural processing of portuguese," 2024.

[24] J. A. Wagner Filho, R. Wilkens, M. Idiart, and A. Villavicencio, "The brWaC corpus: A new open resource for Brazilian Portuguese," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://aclanthology.org/L18-1686

[25] E. da Silva, J. Laterza, and T. Faleiros, "New state-of-the-art for question answering on portuguese squad v1.1," in *Anais do X Symposium on Knowledge Discovery, Mining and Learning*. Porto Alegre, RS, Brasil: SBC, 2022, pp. 98–105. [Online]. Available: https://sol.sbc.org.br/index.php/kdmile/article/view/24974

[26] R. Pires, H. Abonizio, T. S. Almeida, and R. Nogueira, *Sabiá: Portuguese Large Language Models*. Springer Nature Switzerland, 2023, p. 226–240. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-45392-2_15

[27] R. Santos, J. Silva, L. Gomes, J. Rodrigues, and A. Branco, "Advancing generative ai for portuguese with open decoder gervásio pt*," 2024. [Online]. Available: https://arxiv.org/abs/2402.18766

[28] P. Molino, Y. Dudin, and S. S. Miryala, "Ludwig: a type-based declarative deep learning toolbox," 2019. [Online]. Available: https://arxiv.org/abs/1909.07930

[29] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," 2023. [Online]. Available: https://arxiv.org/abs/2305.14314

[30] P. Stefanovič, O. Kurasova, and R. Štrimaitis, "The n-grams based text similarity detection approach using self-organizing maps and similarity measures," *Applied Sciences*, vol. 9, no. 9, 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/9/1870

[31] J. A. Sánchez, V. Romero, A. H. Toselli, M. Villegas, and E. Vidal, "A set of benchmarks for handwritten text recognition on historical documents," *Pattern Recognition*, vol. 94, pp. 122–134, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320319302006

[32] E. S. Ristad and P. N. Yianilos, "Learning string-edit distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 5, pp. 522–532, May 1998.

A Examples of corrections

Figure 4: Image used by OCR solutions to recognize the text. The ground truth is: **perpetuar-se-ão no dia-a-dia brasileiro**.

OCR + Corrector	**Output**	**OCR + Corrector**	**Output**
Bluche	resrsteza-e se no dia-à do brasileiro.	**Flor**	persetuar-se-ão no dia a-dia brasileiro.
+ ByT5	restringe-se no dia-a-dio brasileiro.	+ ByT5	perpetuar-se-ão no dia-a-dia brasileiro.
+ BART	resiste-se se no dia-a-dia do brasileiro.	+ BART	perspetuar-se-ão no dia a-dia brasileiro.
+ Sabia	respeita-se no dia-a-dia do brasileiro.	+ Sabia	persistir-se-ão no dia a dia brasileiro.
+ Gervasio	ressente-se se no dia-à do brasileiro.	+ Gervasio	persetuar-se-ão no dia a dia brasileiro.
+ Symspell	resrsteza-e se no dia do brasileiro.	+ Symspell	persetuar-se-ão no dia adia brasileiro.
+ Norvig	resrsteza-e se no dia-à do brasileiro.	+ Norvig	perpetuar-se-ão no dia a-dia brasileiro.
+ Ngram	resiste se no dia- do brasileiro.	+ Ngram	perpetuar-se no dia dia-dia brasileiro.
Puigcerver	ressatvar-se-ação no dia à dia brasileira.	**Pero**	perpetuar-se-ão no dia-a-dia brasileiro.
+ ByT5	ressaltar-se-ão no dia à dia brasileira.	+ ByT5	perpetuar-se-ão no dia-a-dia brasileiro.
+ BART	ressativar-se-ão no dia à dia brasileiro.	+ BART	perpetuar-se-ão no dia-a-dia brasileiro.
+ Sabia	ressaltar-se-ão no dia a dia brasileiro.	+ Sabia	perpetuar-se-ão no dia-a-dia brasileiro.
+ Gervasio	ressentir-se-ão no dia à dia brasileira.	+ Gervasio	perpetuar-se-ão no dia-a-dia brasileiro.
+ Symspell	ressatvar-se-ação no dia à dia brasileira.	+ Symspell	perpetuar-se-ão no dia-a-dia brasileiro.
+ Norvig	ressaltar-se-ação no dia à dia brasileira.	+ Norvig	perpetuar-se-ão no dia-a-dia brasileiro.
+ Ngram	ressaltar-se no dia à dia brasileira.	+ Ngram	perpetuar-se no dia-a-dia brasileiro.
LTU	perpetuar-se-ão no dia-a-dia brasileiro.	**Demokritos**	rerpetaar-se-ão no dia- a-dia brasileiro.
+ ByT5	perpetuar-se-ão no dia-a-dia brasileiro.	+ ByT5	repetir-se-ão no dia-a-dia brasileiro.
+ BART	perpetuar-se-ão no dia-a-dia brasileiro.	+ BART	realizar-se-ão no dia-a-dia brasileiro.
+ Sabia	perpetuar-se-ão no dia-a-dia brasileiro.	+ Sabia	repetir-se-ão no dia-a-dia brasileiro.
+ Gervasio	perpetuar-se-ão no dia-a-dia brasileiro.	+ Gervasio	rerpetuar-se-ão no dia-a-dia brasileiro.
+ Symspell	perpetuar-se-ão no dia-a-dia brasileiro.	+ Symspell	rerpetaar-se-ão no dia- adia brasileiro.
+ Norvig	perpetuar-se-ão no dia-a-dia brasileiro.	+ Norvig	perpetuar-se-ão no dia- a-dia brasileiro.
+ Ngram	perpetuar-se no dia-a-dia brasileiro.	+ Ngram	tornar-se-ão no dia- dia-dia brasileiro.
LTU Ens.	perpetuar-se-ão no dia-a-dia brasileiro.	**Azure**	ão no dia-
+ ByT5	perpetuar-se-ão no dia-a-dia brasileiro.	+ ByT5	ção no dia.
+ BART	perpetuar-se-ão no dia-a-dia brasileiro.	+ BART	ão no dia-
+ Sabia	perpetuar-se-ão no dia-a-dia brasileiro.	+ Sabia	são no dia-
+ Gervasio	perpetuar-se-ão no dia-a-dia brasileiro.	+ Gervasio	ão no dia.
+ Symspell	perpetuar-se-ão no dia-a-dia brasileiro.	+ Symspell	ão no dia-
+ Norvig	perpetuar-se-ão no dia-a-dia brasileiro.	+ Norvig	ão no dia-
+ Ngram	perpetuar-se no dia-a-dia brasileiro.	+ Ngram	ão no dia-
LITIS	perpetuar-se-ão no dia-a-dia brasileiro.		
+ ByT5	perpetuar-se-ão no dia-a-dia brasileiro.		
+ BART	perpetuar-se-ão no dia-a-dia brasileiro.		
+ Sabia	perpetuar-se-ão no dia-a-dia brasileiro.		
+ Gervasio	perpetuar-se-ão no dia-a-dia brasileiro.		
+ Symspell	perpetuar-se-ão no dia-a-dia brasileiro.		
+ Norvig	perpetuar-se-ão no dia-a-dia brasileiro.		
+ Ngram	perpetuar-se no dia-a-dia brasileiro.		

Table 5: Example of corrections from spelling correction approaches.