Towards Prompt Engineering and Large Language Models for Post-OCR correction in handwritten texts

Sávio Santos de Araújo[1], **Byron Leite Dantas Bezerra**[1], **Arthur Flor de Sousa Neto**[1]

[1]Programa de Pós-graduação em Engenharia de Computação (PPGEC)
Universidade de Pernambuco (UPE)

{savio.santos,byron.leite}@upe.br, afsn@ecomp.poli.br

Abstract. *This work explores the use of Large Language Models (LLMs) for post-OCR spelling correction in full sentences across Portuguese, French, and English. Using outputs from a state-of-the-art recognition model on the BRESSAY, RIMES, and IAM datasets, we evaluated two different zero-shot prompts. Closed LLMs, such as Gemini and GPT series, consistently outperform open-source models in reducing Character Error Rate (CER) and Word Error Rate (WER), while also offering faster inference. Despite the good accuracy of open models, their high computational demands hinder their practical use. Code is available at https://github.com/savi8sant8s/zero-shot-spelling-corrector.*

1. Introduction

Optical Character Recognition (OCR) technology is fundamental for digitizing documents, but it often produces significant errors, especially in handwritten texts. This makes post-OCR correction a critical step to ensure the quality of the digital text. While traditional spelling correction strategies have been explored for this task ([Neto et al. 2020a], [Vargas et al. 2021]), they often struggle with complex, context-dependent errors. In contrast, Large Language Models (LLMs) have emerged as a promising solution due to their deep understanding of language nuances.

This work evaluates the effectiveness of open and closed LLMs for the post-OCR correction task in Portuguese, French, and English. Our study distinguishes itself by focusing on correcting entire sentences or paragraphs, a challenge that requires broader contextual understanding. We employ a zero-shot methodology for all experiments to assess the models' performance in a more realistic and context-rich scenario.

For our evaluation, we used three benchmark handwriting recognition datasets: BRESSAY for Portuguese [Neto et al. 2024], RIMES for French [Grosicki et al. 2008] and IAM for English [Marti and Bunke 2002]. The initial OCR outputs were generated using a state-of-the-art (SOTA) recognition model [Neto et al. 2020b] to establish a consistent error baseline. Model performance was measured using Character Error Rate (CER) and Word Error Rate (WER), alongside the computational cost of inference. Our goal is to identify the most effective and practical models for this task, providing a current, multilingual analysis for the field.

2. Related Works

Recent literature shows a growing interest in using Large Language Models (LLMs) for post-OCR correction, exploring a range of both open-source models (such as LLaMA,

BART, and ByT5) and closed-source systems (like the GPT series). The most comprehensive study, by [Boros et al. 2024], evaluated 14 LLMs across 8 languages and concluded that their overall performance on the task was low. Other works have focused on specific languages, such as Vietnamese [Do et al. 2025] and Portuguese [de Araújo et al. 2024], or investigated the performance of specific models on English texts [Veninga 2024, Thomas et al. 2024]. More recently, [Zhang et al. 2024] conducted an in-depth study with GPT models on historical newspapers, highlighting the importance of prompt engineering. Similarly, [Principe et al. 2025] also achieved good results in English but noted high variability in performance depending on the prompt used.

Our work differentiates itself by performing a zero-shot analysis covering three languages (Portuguese, French, and English) with recent LLMs, both open and closed. Unlike most of the cited work, which focuses on historical or printed texts, we evaluate the models in correcting complete sentences from modern handwriting datasets. Thus, we investigate whether the most recent LLMs can overcome the context limitations and the poor overall performance highlighted by [Boros et al. 2024], offering a practical and updated perspective on the task.

3. Methodology

Our methodology follows the five-stage process. It begins with defining the correction task, moves through the selection of datasets, models, and prompts, and concludes with the execution and evaluation of the experiments.

3.1. Delimit Task

Post-OCR correction can be performed at various levels, from isolated words to entire paragraphs. While traditional methods are often limited to word-level corrections, this work targets the most complex level: the correction of sentence sets (paragraphs).

This approach was chosen to fully leverage the contextual understanding of modern LLMs. By operating on larger text blocks in a zero-shot setup, we aim to correct not only spelling errors but also issues related to punctuation, grammar, and sentence flow, providing a more comprehensive and realistic evaluation of the models' capabilities without any task-specific fine-tuning.

3.2. Select Datasets

Our evaluation is based on three handwritten document datasets with distinct characteristics: BRESSAY [Neto et al. 2024], a collection of formal Portuguese essays with diverse writing styles (from which we constructed 834 sentences using 5,916 test lines); RIMES [Grosicki et al. 2008], consisting of scripted letters in formal French (100 sentences from 779 lines); and the IAM database [Marti and Bunke 2002], composed of formal English texts known for its unconventional punctuation formatting (255 sentences from 1,861 lines). For all datasets, we generated initial OCR outputs from their test partitions using the state-of-the-art HTR-Flor model [Neto et al. 2020b]. These line-level outputs were subsequently grouped into complete sentences to provide richer contextual input for the LLMs during the correction task. Figure 1 presents examples of images present in the selected datasets

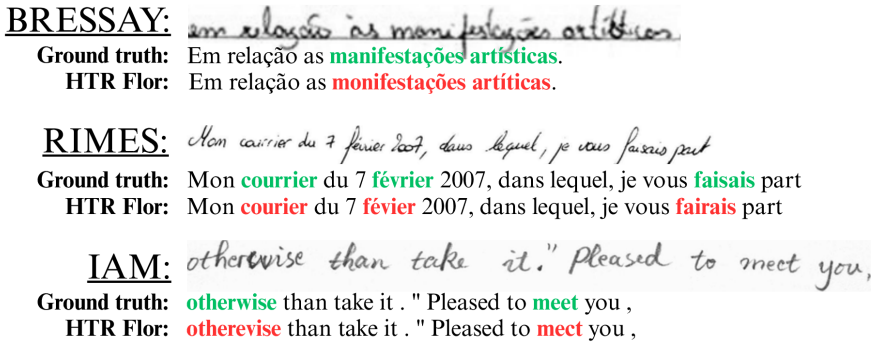

Figure 1. Example lines from the BRESSAY, RIMES and IAM datasets. The words in red are parts that the OCR missed and the words in green are the ground truth.

3.3. Select LLMs

The selection of LLMs for this study aimed to balance performance with practical constraints such as computational resources and time. A comprehensive evaluation is infeasible, considering the hundreds of thousands of open-source models available[1] and the wide variety of proprietary systems, such as Claude[2], Grok[3], and Command[4], which were outside the scope of this work.

We chose five open-source LLMs for local deployment based on their recency and popularity within the community, architectural diversity, and established performance on academic and public leaderboards. Table 1 provides a complete overview of the selected models.

Table 1. LLMs for evaluation. Closed models did not disclose some information.

Model	Total Params	Knowledge Cutoff	Open / Closed	GPU Usage
GPT-4o Mini[5]	-	2023-09	Closed	-
GPT-3.5 Turbo[6]	-	2021-08	Closed	-
Gemini 2.0 Flash[7]	-	2024-08	Closed	-
Gemini 2.0 Flash Lite[8]	-	2024-08	Closed	-
Sabia 3[Abonizio et al. 2025]	-	2024-09	Closed	-
Sabiazinho 3[Abonizio et al. 2025]	-	2025-02	Closed	-
Phi-4[Abdin et al. 2024]	14B	2024-06	Open	11.1GB
Mistral 3.1 Small[9]	24B	c. 2025	Open	16.6GB
Gemma 3[Team et al. 2025]	27B	2024-08	Open	19.3GB
Qwen 2.5[Qwen et al. 2025]	72B	2024-09	Open	48.7GB
LLaMA 4[10]	109B	2024-08	Open	64.1GB

[1] https://huggingface.co/models?pipeline_tag=text-generation
[2] https://www.anthropic.com/claude
[3] https://x.ai/grok
[4] https://cohere.com/command

The selection of proprietary models was strategic. Google and OpenAI models were chosen for their strong cost-benefit ratio, speed, and reliable multilingual performance. In contrast, Maritaca AI's models were included specifically for their specialization in Portuguese, making them ideal for evaluating the BRESSAY dataset.

During our initial screening of open-source models, several candidates were evaluated but ultimately excluded from the final analysis. The IBM Granite models (v3.1-v3.3) [11] were discarded due to a high rate of hallucination. Similarly, Deepseek-R1 [DeepSeek-AI et al. 2025] and Qwen 3[12] also exhibited significant hallucinations and unpredictable inference times, making them unsuitable for reliable benchmarking.

3.4. Select Prompts

We evaluated two distinct prompts with different philosophies, as shown in Table 2. The first prompt (1) is a concise instruction of our own design, focused on simplicity and objectivity. The second (2) is a more detailed, rule-based prompt adapted from [Principe et al. 2025], which was identified as a high-performing prompt for post-OCR correction. To ensure consistency and leverage the models' primary training data, all prompt instructions were in English, regardless of the source text's language.

A common challenge with LLMs is their tendency to add extraneous text, which complicates automatic evaluation. To mitigate this, we adopted a tagging strategy inspired by [Do et al. 2025]. Each text line was encapsulated in unique identifier tags following the format `<i.j.k> text line </i.j.k>`. This structure constrains the model's output to the defined tags and allows for reliable extraction of the corrected text using regular expressions.

Table 2. Prompts used in the experiments.

ID	Prompt
1 (ours)	Correct only obvious spelling mistakes in words within tags. Keep the number of tags the same. Do not add extra text or change correct text. Maintain the unique and historical style of the text.
2 (adapted)	Act as a document analyst specialising in OCR correction. Your task is to correct OCR errors in the text. Guidelines: 1. Ensure corrections accurately reflect the text language and conventions. 2. Keep punctuation marks from the original text, and do not add new punctuation. 3. Preserve original word splits. 4. Keep the hyphenation in the original text. 5. Do not delete words unless duplicated. 6. Do not modify the end of the text. 7. Do not correct numbers.

3.5. Predict and Evaluate

All experiments were conducted on Google Cloud Platform (GCP) in a Linux Debian 11 environment with 167GB of RAM, using an Intel Xeon CPU (12) and NVIDIA A100 SXM4 80GB. The inference method used was the zero-shot prompt, where no reference

[11] https://huggingface.co/ibm-granite
[12] https://qwenlm.github.io/blog/qwen3

example of the task to be performed was passed and the temperature 0 was used as the task aims for a deterministic result.

To evaluate the LLMs outputs with the ground truth, the CER and WER metrics were used, which are widely used to evaluate OCR and HTR systems [Sánchez et al. 2019] and were the metrics used in related works. In addition to the LLM predictions, the time spent to perform the task and GPU consumption (for open LLMs) were calculated. After this, the metrics of the original raw OCR outputs and LLM predictions were computed and the metrics of each approach were averaged, allowing to compare the performance of the models and prompts on the task.

4. Results

The complete results of our experiments are presented in Table 3, with inference performance detailed in Figure 2. The analysis reveals several key trends.

Table 3. CER and WER mean across datasets for each prompt.

Model	Prompt(s)	CER Mean	WER Mean
Baseline	–	6.60	22.62
GPT-4o Mini	1 / 2	6.13 / 6.02	14.83 / 13.89
GPT-3.5 Turbo	1 / 2	6.64 / 6.03	17.73 / 16.34
Gemini 2.0 Flash	1 / 2	**5.54** / 7.76	15.06 / 18.57
Gemini 2.0 Flash Lite	1 / 2	8.03 / 7.36	18.39 / 19.15
Sabia 3	1 / 2	11.01 / 9.46	20.53 / 19.08
Sabiazinho 3	1 / 2	12.81 / 10.47	21.35 / 23.31
Phi-4 14B	1 / 2	7.63 / 8.01	17.71 / 19.39
Mistral 3.1 24B	1 / 2	7.26 / 7.37	16.23 / 16.13
Gemma 3 27B	1 / 2	7.14 / 6.93	17.13 / 16.36
Qwen 2.5 72B	1 / 2	6.85 / 7.09	15.14 / **13.65**
LLaMA 4 109B	1 / 2	7.63 / 7.61	19.80 / 21.46
Prompt Mean	1 / 2	7.88 / 7.65	17.63 / 17.94

The results for CER highlight the difficulty of this fine-grained task. As shown in Table 3, only a few closed-source models managed to improve upon the baseline of 6.60%. Specifically, Gemini 2.0 Flash (with Prompt 1) achieved the most significant reduction (5.54%), followed closely by GPT-4o Mini. Conversely, most open-source models, along with the Sabiá models, actually increased the CER, indicating that they introduced more character-level errors than they corrected.

In contrast, the performance on WER was much more positive across the board. Nearly all models successfully reduced the WER from the baseline of 22.62%, demonstrating a strong capability for correcting whole-word errors. The open-source Qwen 2.5 72B (with Prompt 2) delivered the best performance, achieving the lowest WER of

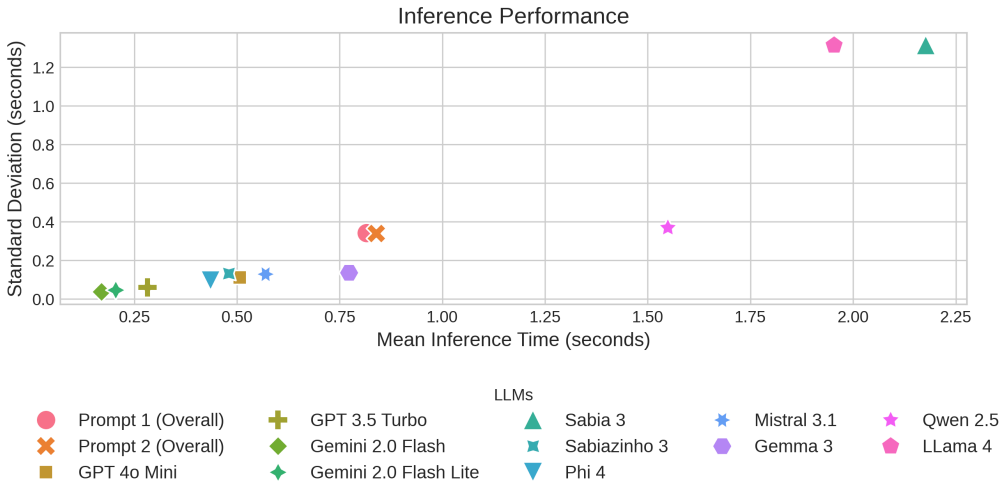

Figure 2. Inference performance of LLMs, analysing the line correction time. The graph shows the average time (in seconds) on the x-axis and the standard deviation on the y-axis. Lower values on both axes indicate better results.

13.65%. Other strong performers included GPT-4o Mini and GPT-3.5 Turbo, which also provided substantial improvements.

This discrepancy between WER and CER improvements suggests that while LLMs excel at semantic, word-level correction, they struggle with the character-level noise inherent in OCR outputs. This aligns with previous research [de Araújo et al. 2024, Veninga 2024] pointing to the limitations of standard tokenization. However, this performance must be weighed against the practical costs shown in Figure 2. The top-performing open models, like Qwen 2.5, are computationally intensive and slow. The closed models, on the other hand, offer a superior balance, providing both high accuracy and fast, efficient inference, making them more suitable for real-world deployment.

5. Conclusion

This work evaluated state-of-the-art LLMs for post-OCR correction in Portuguese, French, and English, revealing that closed-source models hold a distinct advantage. Models like Gemini 2.0 Flash and GPT-4o Mini consistently outperformed open-source alternatives by providing superior reductions in both CER and WER, alongside significantly faster and more efficient inference.

While the closed-source Gemini 2.0 Flash achieved the best CER reduction, the open-source Qwen 2.5 72B delivered the top WER reduction. This highlights a key insight: current LLMs are more proficient at correcting whole-word errors than finer, character-level noise. This aligns with previous research on the limitations of standard tokenization and suggests the potential of ”token-free” architectures, despite their high computational cost.

In summary, while the open-source ecosystem is rapidly advancing, closed-source models currently offer a more practical balance of performance and efficiency for this task. This research underscores the importance of considering both accuracy and deployment costs when selecting an LLM for real-world post-OCR applications.

Acknowledgments
This study was financed by the founding public Brazilian agencies CNPq and CAPES (Finance Code 001), and the University of Pernambuco. In addition, we acknowledge all support from Di2Win (`www.di2win.com`).

References

Abdin, M., Aneja, J., Behl, H., Bubeck, S., and et al. (2024). Phi-4 technical report.

Abonizio, H., Almeida, T. S., Laitz, T., Junior, R. M., Bonás, G. K., Nogueira, R., and Pires, R. (2025). Sabiá-3 technical report.

Boros, E., Ehrmann, M., Romanello, M., Najem-Meyer, S., and Kaplan, F. (2024). Post-correction of historical text transcripts with large language models: An exploratory study. In Bizzoni, Y., Degaetano-Ortlieb, S., Kazantseva, A., and Szpakowicz, S., editors, *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 133–159, St. Julians, Malta. Association for Computational Linguistics.

de Araújo, S. S., Bezerra, B. L. D., de Sousa Neto, A. F., and Zanchettin, C. (2024). A proposal for post-OCR spelling correction using language models. In *Latinx in AI @ NeurIPS 2024*.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., and et al. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Do, T., Tran, D. P., Vo, A., and Kim, D. (2025). Reference-based post-ocr processing with llm for precise diacritic text in historical document recognition.

Grosicki, E., Carre, M., Brodin, J.-M., and Geoffrois, E. (2008). RIMES evaluation campaign for handwritten mail processing. *ICFHR 2008 : 11th International Conference on Frontiers in Handwriting Recognition*, pages 1–6.

Marti, U.-V. and Bunke, H. (2002). The IAM-database: An English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5.

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

Neto, A. F. d. S., Bezerra, B. L. D., and Toselli, A. H. (2020a). Towards the natural language processing as spelling correction for offline handwritten text recognition systems. *Applied Sciences*, 10(21).

Neto, A. F. d. S., Bezerra, B. L. D., Toselli, A. H., and Lima, E. B. (2020b). Htr-flor: A deep learning system for offline handwritten text recognition. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 54–61.

Principe, J. P. P., Fischer, A., and Scius-Bertrand, A. (2025). Post-correction of handwriting recognition using large language models. In Buntine, W., Fjeld, M., Tran, T., Tran, M.-T., Huynh Thi Thanh, B., and Miyoshi, T., editors, *Information and Communication Technology*, pages 106–118, Singapore. Springer Nature Singapore.

Qwen, Yang, A., Yang, B., Zhang, B., and et al. (2025). Qwen2.5 technical report.

Sánchez, J. A., Romero, V., Toselli, A. H., Villegas, M., and Vidal, E. (2019). A set of benchmarks for handwritten text recognition on historical documents. *Pattern Recognition*, 94:122–134.

Team, G., Kamath, A., Ferret, J., Pathak, S., and et al. (2025). Gemma 3 technical report.

Thomas, A., Gaizauskas, R., and Lu, H. (2024). Leveraging LLMs for post-OCR correction of historical newspapers. In Sprugnoli, R. and Passarotti, M., editors, *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 116–121, Torino, Italia. ELRA and ICCL.

Vargas, D. S., de Oliveira, L. L., Moreira, V. P., Bazzo, G. T., and Lorentz, G. A. (2021). socrates - a post-ocr text correction method. In *Anais do XXXVI Simpósio Brasileiro de Bancos de Dados*, pages 61–72, Porto Alegre, RS, Brasil. SBC.

Veninga, M. (2024). Llms for ocr post-correction.

Zhang, J., Haverals, W., Naydan, M., and Kernighan, B. W. (2024). Post-ocr correction with openai's gpt models on challenging english prosody texts. In *Proceedings of the ACM Symposium on Document Engineering 2024*, DocEng '24, New York, NY, USA. Association for Computing Machinery.