

Image-Informed Post-OCR Correction with Vision-Language Models

Alan Thomas^{1,2,*}, Xianyuan Liu^{1,2,*}, Haiping Lu^{1,2}, Robert Gaizauskas²

¹Centre for Machine Intelligence, University of Sheffield

²School of Computer Science, University of Sheffield

* Equal contribution

{alan.thomas, xianyuan.liu, h.lu, r.gaizauskas}@sheffield.ac.uk

Abstract

Improving the quality of OCR transcriptions is crucial for researchers working with digital archives, as OCR errors hinder readability and downstream NLP tasks. While prior work has focused on seq2seq neural machine translation and, more recently, prompt-based approaches with large language models (LLMs), these methods rely solely on text, overlooking the visual characteristics of source documents. In this work, we explore the potential of vision-language models (VLMs) for post-OCR correction to evaluate the impact of integrating visual information. We conduct experiments on the BLN600 dataset of 19th-century newspaper articles, comparing Llama 3.2 Vision and Qwen 2.5 VL to their text-only LLM counterparts. Our results show that VLMs significantly outperform LLMs, achieving approximately 90% reduction in character and word error rates, with the optimal setting being one that takes as input both document image and prior noisy OCR. This suggests that VLM-based post-OCR correction can be a viable and effective approach for correcting back catalogues of poorly OCR'ed historical documents and improving the usability of digital archives.

1 Introduction

Digital archives are essential for preserving and providing access to historical documents. Optical Character Recognition (OCR) technology is used to convert scanned document images into text, which is then stored in databases such as Chronicling America¹, Gale Primary Sources², and Trove³. These archives are invaluable resources for researchers, facilitating large-scale text mining and the exploration of cultural and historical records.

Historical documents present unique challenges for OCR due to poor paper quality, fading, ink

spills, outdated typefaces, and complex layouts, resulting in OCR errors (Holley, 2009). In turn, this hinders text retrieval (Chiron et al., 2017a) and negatively impacts NLP tasks (van Strien et al., 2020). To address this, post-OCR correction has become a critical task in refining noisy OCR output and restoring it to its intended form.

Prior to the recent large language model (LLM) revolution, state-of-the-art approaches framed post-OCR correction as a seq2seq neural machine translation task (Nguyen et al., 2021). More recently, LLMs have been explored as a solution via a prompt-based approach (Thomas et al., 2024; Boros et al., 2024). While these methods have demonstrated progress, they rely solely on text input and lack crucial visual context from source images. As a result, they can suffer from under-correction, leaving OCR errors intact, or over-correction, introducing new errors.

With the emergence of large vision-language models (VLMs) (Yin et al., 2024), integrating visual information presents a promising direction for improving post-OCR correction. By incorporating both document images and OCR text, a multimodal approach can leverage visual cues for a more accurate and informed correction. While going back to the document image might appear to be “cheating”, and no longer “post-OCR text correction”, there is a distinction to be made between a process that transcribes afresh from a document image vs one that starts with a noisy transcription together with the image and uses both to derive an improved transcription. Each has distinct advantages and disadvantages. The latter is still properly called “post-OCR correction” and is what we pursue here.

Specifically, in this work, we explore the potential of VLMs for post-OCR correction and compare their performance to text-only LLMs in both prompting and supervised fine-tuning settings. We conduct experiments on BLN600, an open-access dataset of 19th-century British newspaper articles

¹<https://chroniclingamerica.loc.gov/>

²<https://www.gale.com/primary-sources>

³<https://trove.nla.gov.au/>

(Booth et al., 2024), which includes source images, OCR text and manually re-keyed human transcriptions. To create paired training data, we align OCR text sequences with corresponding image patches by leveraging the ALTO XML schema⁴, which encodes words with spatial coordinates and is widely used in digital historical archives. Our results demonstrate that VLMs with multimodal input are more than twice as effective as LLMs with text input only, reducing character and word error rates by approximately 90%. This positions them as a highly effective solution for enhancing digital archives.

2 Related Work

OCR technology has improved significantly over the years. Recent state-of-the-art systems such as Tesseract (Smith, 2007), TrOCR (Li et al., 2023), and PaddleOCR (Du et al., 2020) achieve strong performance on general-purpose OCR tasks but often struggle with the diverse challenges posed by historical documents from different time periods. To address this, specialized OCR systems have been developed. Transkribus (Kahle et al., 2017) provides models trained on manuscript and print archives, and EffOCR (Carlson et al., 2024) focuses on sample-efficient OCR training. Despite these advances, post-OCR correction remains a practical and cost-effective solution for improving the OCR quality of large-scale archives without full reprocessing of the original documents.

Post-OCR correction methods can be broadly classified into three types: manual, isolated-word, and context-dependent (Nguyen et al., 2021). Manual correction is highly accurate but incurs significant costs in time and labour. Isolated-word approaches correct each word independently, while context-dependent approaches leverage surrounding text for more accurate corrections.

Various approaches have been explored for post-OCR correction of historical texts across multiple languages. ICDAR held competitions on post-OCR correction in 2017 and 2019, providing parallel corpora for several European languages (Chiron et al., 2017b; Rigaud et al., 2019). Amrhein and Clematide (2018) won the 2017 edition with an ensemble of statistical and neural machine translation models. The 2019 edition was won by Clova AI using a pre-trained multilingual BERT

(Beshirov et al., 2025), while Ramirez-Orta et al. (2022) later achieved a new state-of-the-art performance by combining corrections from character-level models with a voting scheme. Fine-tuning pre-trained encoder-decoder models has become prevalent in post-OCR correction, with models such as BART and ByT5 proving effective across languages such as English, French, Sanskrit, Swedish, and Dutch (Soper et al., 2021; Maheshwari et al., 2022; Löfgren and Dannélls, 2024; Guan et al., 2024; Debaene et al., 2025). Recent data-centric studies demonstrate that post-OCR performance can be improved by carefully constructing synthetic data (Guan and Greene, 2024).

With the strong performance of LLMs across NLP tasks and the increasing availability of open-source models, recent work has explored their potential for post-OCR correction, motivated by promising results on historical records (Kim et al., 2025) and handwritten documents (Crosilla et al., 2025). Boros et al. (2024) evaluated fourteen LLMs for post-correcting historical transcriptions under zero-shot and few-shot prompting, finding that performance remains limited without fine-tuning. Manrique-Gomez et al. (2024) have applied LLM-based pipelines to historical document collections, including GPT-4-based corrections for 19th-century Latin American newspapers. Instruction-tuned Llama models have also shown advantages over fine-tuned seq2seq baselines such as BART in low-data settings (Thomas et al., 2024). In addition, PleIAs released OCRonos⁵, a Llama 3 model fine-tuned for correcting OCR errors.

Multimodal approaches to post-OCR correction remain relatively underexplored but have shown promising results, with recent work demonstrating the potential of vision-language models for historical document transcription (Greif et al., 2025; Semnani et al., 2025). Studies have incorporated visual information from document images in different ways, including glyph embeddings of OCR characters (Chen and Zhou, 2023), combining TrOCR with CharBERT for handwritten text correction (Chen and Ströbel, 2024), and integrating CLIP visual features with a ByT5 decoder for historical German texts (Wu, 2023). These approaches consistently outperform text-only baselines by leveraging visual cues alongside textual context.

The rapid development of VLMs, following releases such as GPT-4V and open-source systems

⁴https://en.wikipedia.org/wiki/Analyzed_Layout_and_Text_Object

⁵<https://huggingface.co/PleIAs/OCRonos>

like LLaVA, has further expanded opportunities for multimodal document understanding (Yin et al., 2024). In particular, some recent VLMs, such as Qwen2.5 VL (Bai et al., 2025), not only perform OCR but, as part of their grounding capability, provide bounding boxes for “objects”, including text boxes and potentially individual words, within complex page layouts. Nonetheless, although VLMs exhibit emergent OCR capabilities, benchmarks such as OCRBench show that they still struggle with noisy or blurry text (Liu et al., 2024), suggesting the need for methods that effectively combine visual information with existing OCR transcriptions.

3 Methodology

In this section, we outline our methodology for image-informed post-OCR correction with VLMs. We define the task and its motivation, followed by an overview of data preparation, model selection, and evaluation metrics.

3.1 Problem Formulation

Most off-the-shelf OCR systems are built for general-purpose use and struggle with the various issues posed by historical document image scans. Traditional approaches to post-OCR correction rely solely on OCR text as input, attempting to fix errors without considering the visual factors that caused them. Without structural or contextual cues from the original image, these methods often fail, either under-correcting, allowing errors to persist, or over-correcting, introducing new mistakes. Additionally, the evolution of language over time poses a challenge for models that are largely trained on modern text. By incorporating the source image, we address these issues by leveraging both textual and visual features to inform corrections.

Many historical archives were digitized long ago, and newer, more advanced OCR engines may produce higher-quality transcriptions. However, this is often impractical as reprocessing entire collections can be costly and time-consuming, and regenerates the transcription independently of the existing archive, discarding the word level grounding metadata on which current search and display systems rely. A particular issue is that providers of large-scale digitized historical text archives have a substantial investment in user interface and search technology layered on top of their digitized archives. This technology carries out search over the OCR’ed

text and exploits word level bounding box coordinates, generated by the OCR engine during transcription and linked with terms in the OCR text, to display highlighted portions of the original document image containing the search terms. This search and display functionality is essential for giving historians direct access to primary sources and for enabling human mitigation of noisy OCR, at least for errors of commission.

While re-doing OCR starting from the source image using a state-of-the-art VLM tuned for OCR may produce high-quality transcription, the bounding box information for each word is also needed to support search and display. As indicated in the previous section, VLM-based OCR systems are beginning to include “grounding” or “localization” information, but there is currently little evidence of its accuracy. Furthermore, it is unlikely to be directly compatible, in terms of format and coordinate reference system, with existing provider systems and exploiting it would require a significant integration effort that may prove financially infeasible.

An alternative, less disruptive and hence less costly approach is to improve the quality of available OCR outputs at the sentence level, “inheriting” the existing image grounding information from the initial OCR process. This avoids major updates to the existing structure of digital archives or existing search methods. The process can be applied piecemeal or globally across an archive and repeated over time, ensuring continuous iterative improvements in text accuracy.

Thus, it is worth exploring the potential of VLMs not just to reprocess entire multimillion document archives from scratch but to improve post-OCR correction. Our specific proposals on how to do this follow.

3.2 Data Preparation

BLN600 is an open-access dataset of 19th-century newspaper texts derived from the British Library Newspapers collection, which spans over 300 years of UK newspaper publishing in the UK (Gale, 2025). The dataset provides OCR text and source images for 600 articles alongside high-quality manually re-keyed transcriptions (Booth et al., 2024), comprising approximately 300K tokens and 1.7M characters in both OCR and ground-truth text.

As we require a larger dataset for training, we split BLN600 into smaller paired units of image patch, OCR text and ground truth. OCR engines

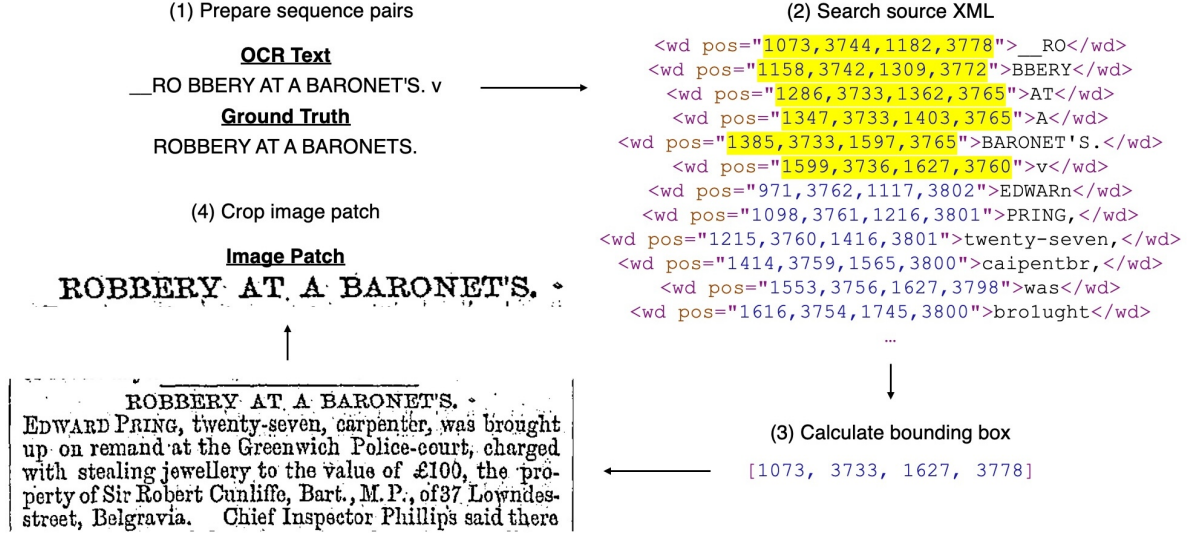


Figure 1: Data preparation pipeline for creating paired data of OCR text, ground truth and image patch

produce metadata for layout information and detected characters or words, which we can leverage for this task. While BLN600 offers only plain text OCR and individual article images, Gale’s archives provide structured OCR data along with full-page source images. The structured OCR adheres to the ALTO XML schema, encoding each word with its spatial coordinates on the original image (Beals and Bell, 2020). We obtain access to Gale’s text mining drives through our institution. Our data preparation pipeline is described below and illustrated in Figure 1.

Firstly, we prepare sequence pairs of OCR text and ground truth. Since BLN600 is not aligned, we segment the ground truth into sentences and use a search algorithm to match each sentence with the corresponding OCR text. Some manual effort was required to ensure proper alignment. When the OCR engine merges words across sentences, we adjust by combining the corresponding ground truth sentences. Next, we locate the OCR text sequence within the source XML file and extract all associated word coordinates. These coordinates are then used to compute a bounding box that encompasses the entire OCR text sequence. Finally, we use this bounding box to crop the full page image, generating an image patch that corresponds to a sequence pair of OCR text and ground truth ⁶.

⁶Note that this method of computing a bounding box leads to excess text from preceding and succeeding sequences being included in the bounding box. We explored a masking strategy which trims these excess additions, but the impact on performance was negative. Details are reported in Appendix D

We prepare a dataset of 12,166 sequence pairs, which is then split into training (70%), validation (10%) and testing (20%) sets. Table 1 presents a breakdown of the sets. Character error rate (CER) and word error rate (WER) are metrics used to evaluate the performance of text recognition systems. CER measures the proportion of characters that are incorrect, while WER calculates errors at the word level. Further details are provided in Section 3.4.

	# sample	# sequence	μ CER	μ WER
Total	600	12,166	0.0736	0.2404
Train	420	8,507	0.0709	0.2357
Val	60	1,280	0.0758	0.2397
Test	120	2,379	0.0822	0.2577

Table 1: BLN600 splits with mean CER/WER

This method for preparing data can be applied to most corpora that follow digitisation standards and provide spatial information for OCR output. Word coordinates are commonly used for text highlighting, document navigation and other features in digital archives. By leveraging existing metadata, a multimodal approach to post-OCR correction can generate corrections that are directly traceable to corresponding regions in the source image, improving usability and facilitating potential integration with existing tools in digital archives.

For our experiments, model evaluation requires aligned OCR text, ground truth, images, and text coordinates, but limited public datasets provide all four. The ICDAR post-OCR corpus (Rigaud et al., 2019) includes OCR text and ground truth

but lacks images or coordinates, which our method requires. In the NZZ black-letter dataset (Ströbel and Clematide, 2019), the ground truth is derived from a new OCR run and is, therefore, not aligned with the original OCR text. A Dutch historical newspaper dataset (Wilms et al., 2020), with 2,000 images and aligned original and manually corrected OCR in ALTO format, most closely matches our requirements, but access requires a signed agreement that we did not have in place at the time of writing. Therefore, we use BLN600 in this paper and leave validation on other datasets to future work.

3.3 Model Selection

For our experiments, we use recent models from the Llama 3 (Dubey et al., 2024) and Qwen 2.5 (Yang et al., 2024) families, chosen for their open-access availability and sufficiently long context windows. These model families offer both VLMs and LLMs, enabling direct comparison of text-only and multimodal post-OCR correction. The VLMs (Llama 3.2V and Qwen 2.5 VL) extend their LLM counterparts (Llama 3.1 and Qwen 2.5) by incorporating a pre-trained vision encoder for image processing. As both VLMs are trained on general OCR-related data, they may offer advantages for document transcription tasks. We focus on medium-sized models to balance computational cost and practical deployment, specifically Llama 3.1 8B, Llama 3.2V 11B, Qwen 2.5 7B, and Qwen 2.5 VL 7B.

For text-only experiments, we use instruction-tuned LLMs as baselines, as prior work (Thomas et al., 2024) reports that these LLMs outperform seq2seq models on the BLN600 dataset. As we aim to evaluate the contribution of visual input rather than to compare text-only architectures, we do not include additional seq2seq baselines, such as BART or ByT5, in this work.

3.4 Evaluation Metrics

To assess the effectiveness of our approach, we focus on the percentage reduction in character error rate (CER) and word error rate (WER), which quantifies the overall reduction of errors at both the character and word levels. These metrics reflect the improvement in OCR quality after correction.

CER and WER are computed at the corpus level using Levenshtein distance, which counts the number of edits required to match OCR text to ground truth through substitutions (replacing), deletions (removing) and insertions (adding). We define a general formula for the error rate as

$$ER = \frac{\sum_{i=1}^n (S_i + D_i + I_i)}{\sum_{i=1}^n N_i}, \quad (1)$$

where S_i is the number of substitutions, D_i is the number of deletions, and I_i is the number of insertions at either character or word level in sequence i . N_i is the total number of characters or words in the reference text and n is the total number of samples.

To measure relative improvement of error rates, we define error rate reduction as

$$ER_{reduction} = \left(1 - \frac{ER_{corrected}}{ER_{original}}\right) \times 100, \quad (2)$$

where $ER_{original}$ is the error rate of the original OCR texts, and $ER_{corrected}$ is the error rate of the corrected texts. This reflects the percentage reduction in error rate, where a value of 0% indicates no improvement, 100% represents a perfect correction, and negative values indicate an increase in errors.

4 Experiments & Results

In this section, we detail our experimental setup and findings, exploring the performance of VLMs and LLMs in both prompting and supervised fine-tuning settings.

4.1 Prompting

We use instruction-tuned models for prompting, as they are trained to produce assistant-like responses. Following prior work (Boros et al., 2024), we design a prompt that provides OCR text and instructs the model to return the corrected transcription rather than explanations or commentary. For VLMs, the prompt additionally indicates that the corresponding image is provided as input. To improve consistency, we use a one-shot prompt that includes an example OCR text and its corrected version. The full prompt templates are provided in Appendix A. We apply the prompt template to the test samples and collect the generated corrections. Results are reported in Table 2, averaged over five runs to account for variability.

Model	Type	CER _r (%)	WER _r (%)
Llama 3.1 8B	LLM	14.71	40.21
Llama 3.2V 11B	VLM	42.90	67.53
Qwen 2.5 7B	LLM	9.88	33.40
Qwen 2.5 VL 7B	VLM	45.34	64.58

Table 2: Model performance with prompting

We note a significant reduction in WER across the board, primarily due to the removal of excess OCR artefacts, such as random symbols and punctuation, rather than actual words. Therefore, reduction in CER better reflects the correction of mis-recognized words. VLMs outperform LLMs, suggesting that incorporating images aids in resolving OCR errors. While LLMs sometimes produce reasonable corrections, they deviate from the ground truth by altering capitalisation, introducing spaces and rephrasing text. VLMs perform better as the image helps inform the correction. However, OCR errors persist in their outputs, whereas LLMs always attempt a correction. VLMs are likely trained on clean OCR data and require fine-tuning to better adapt to specific image features and error patterns.

4.2 Fine-tuning

Using the sequence pairs in our train set, we create a dataset for instruction-tuning LLMs and VLMs. We format the samples such that the system prompt informs the model of the task, the user provides OCR text and the model generates the ground truth. For VLMs, special tokens are added to indicate the presence of image input. We perform parameter-efficient fine-tuning using QLoRA (Dettmers et al., 2023), which reduces memory usage through quantization and by updating a low-rank decomposition of the weight matrix instead of all model weights. For LLMs, we focus on fine-tuning all linear layers. For VLMs, we fine-tune the attention layers in both the vision encoder and language model, along with the MLP layers in the language model. Since Llama 3.2 uses cross-attention, we also fine-tune these layers to improve alignment between modalities. We access the pre-trained versions of models through Hugging Face’s transformers library (Wolf et al., 2020). All models are trained on a single NVIDIA A100 GPU with 80GB memory. Full training details are available in Appendix B. After fine-tuning, we generate corrections for our test set with sampling disabled for deterministic outputs. We aggregate results over five training and evaluation runs and report performance in Table 3.

We observe a significant improvement in performance over out-of-the-box models, with VLMs outperforming LLMs. After fine-tuning, the models produce more accurate and stable outputs by not only handling common misspellings but also learning to recognize recurring error patterns in the dataset. Furthermore, they can more effectively handle period- and genre-specific linguistics

Model	Type	Input		CER _r (%)	WER _r (%)
		Image	Text		
Llama 3.1 8B	LLM	✗	✓	41.63	66.84
Llama 3.2V 11B	VLM	✓	✓	92.03	91.40
		✓	✗	84.27	90.05
		✗	✓	43.42	61.79
Qwen 2.5 7B	LLM	✗	✓	43.78	65.79
Qwen 2.5 VL 7B	VLM	✓	✓	89.27	90.28
		✓	✗	83.49	89.52
		✗	✓	-	-

Table 3: Model performance after fine-tuning. Text-only results for Qwen 2.5 VL were not obtainable.

tic features. However, LLMs still struggle with highly incoherent or illegible text, often hallucinating and fabricating a plausible but incorrect correction. This highlights their lack of a strong grounding mechanism to reliably handle noisy and unclear inputs. Conversely, VLMs seem capable of localising OCR text within the image and adapting to visual features for improved text correction. We also compare our approach with VLM-based re-OCR using PaddleOCR-VL (Cui et al., 2025) in Appendix C, where our larger reductions in CER and WER show that leveraging both OCR text and visual evidence is more effective than re-OCRing images alone.

4.2.1 Fine-tuning strategies

We explore the effects of minimal fine-tuning strategies with VLMs. Our original setup fine-tunes all attention and MLP layers. To reduce the number of trainable parameters, we apply the QLoRA adapter selectively, fine-tuning only attention layers or both attention and MLP layers. We also experiment with freezing the vision encoder. The results in Table 4 show that minimal fine-tuning still performs well. Ignoring MLP layers does not impact performance much but freezing the vision encoder results in a slight performance drop. Fine-tuning the vision encoder improves the model’s ability to extract and align visual features with text, enabling it to better handle historical document images.

4.2.2 Varying training data

Large generative models can effectively adapt to downstream tasks with a limited amount of high-quality fine-tuning data, as most knowledge is acquired during pre-training (Zhou et al., 2023). To investigate this, we partition the original training

Model	Strategy	Freeze Vision	# Params	CER _r (%)	WER _r (%)
Llama 3.2V 11B	attention + MLP	✗	29.5M	92.03	91.40
		✓	26.2M	79.35	84.18
	attention	✗	11.8M	90.06	90.71
		✓	8.5M	73.28	80.77
Qwen 2.5 VL 7B	attention + MLP	✗	25.7M	89.27	90.28
		✓	20.2M	80.54	85.02
	attention	✗	7M	86.57	87.31
		✓	5M	75.09	80.92

Table 4: Comparison of QLoRA fine-tuning strategies

Model	Window Size	Avg Seq Length	Avg Image Size	CER _r (%)	WER _r (%)
Llama 3.2V 11B	1	115 tokens	723 x 121px	92.03	91.40
	3	241 tokens	749 x 285px	90.59	91.08
	6	409 tokens	751 x 502px	87.89	88.60
	9	553 tokens	751 x 688px	86.74	88.07
Llama 3.1 8B	1	115 tokens	—	41.63	66.84
	3	241 tokens	—	28.66	53.70
	6	409 tokens	—	27.89	51.84
	9	553 tokens	—	23.74	45.30

Table 5: Model performance with varying input size

set in Table 1 into six subsets of 70 samples each and incrementally re-train our models, incorporating sequences from additional subsets. As shown in Figure 2, the performance gains are minimal and our approach remains effective with limited data.

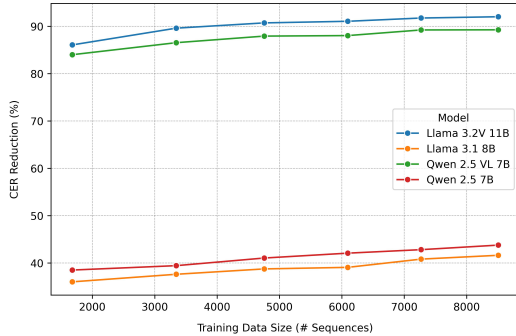


Figure 2: CER reduction versus training data size

Since our models perform well with minimal training data, we explore the effects of increasing input size without expanding the dataset. While most post-OCR correction approaches operate at the line level, LLMs and VLMs support large context windows, enabling evaluation on larger inputs. To increase the input context, we merge sequences into fixed-size groups of 3, 6, and 9. A window size of 1 corresponds to the original dataset. For each merged sequence, we generate a larger image patch using our data preparation pipeline. Due to the col-

umn layout of newspapers, merging sequences increases the height of image patches rather than their width. We fine-tune Llama 3.1 8B and Llama 3.2V 11B on these expanded inputs. As shown in Table 5, LLM performance declines as sequence length increases, whereas VLMs remain more robust. This finding shows text-only models are prone to compounding errors, where early mistakes propagate through the sequence. In contrast, VLMs produce corrections that remain aligned with the source image and localize and correct errors more reliably.

5 Discussion

To better understand the performance gap between LLMs and VLMs in post-OCR correction, we analyse the corrections produced by Llama 3.1 8B and Llama 3.2V 11B. For each test sample, we compute the CER reduction and plot it against the initial CER in Figure 3, where a higher initial CER indicates more challenging corrections. The results show that LLMs often further degrade OCR outputs, even for “easy” samples, and when improvements occur, they are typically minor. In contrast, VLMs rarely degrade outputs and consistently handle “harder” samples more effectively, with the additional visual input enabling corrections that are substantially closer to the ground truth.

Beyond overall performance, we examine how well our models correct different types of OCR

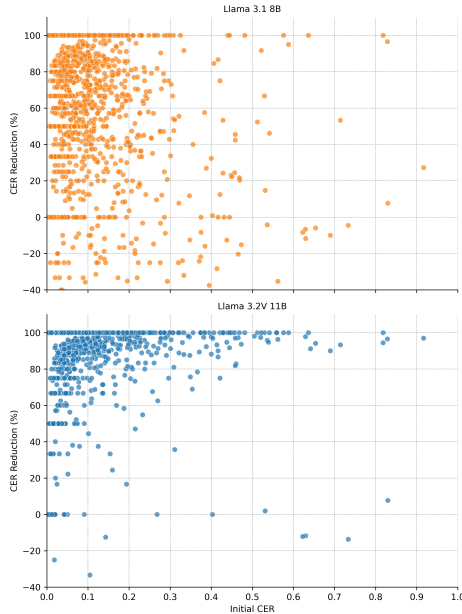


Figure 3: CER reduction versus initial CER

errors, categorized using Levenshtein distance operations as substitutions, deletions and insertions. To measure this, we count the total occurrences of each type at the character level and determine the percentage that were correctly fixed by Llama 3.1 8B and Llama 3.2V 11B. The results in Figure 4 reveal distinct patterns. Deletions are the easiest errors to correct, likely because OCR engines often produce excess artifacts such as punctuation errors or random symbols that both models can readily remove. LLMs perform reasonably well on substitutions, but struggle significantly with insertions, where missing characters or words need to be restored. This typically arises when the OCR engine fails to transcribe text due to severe fading or poor image quality. In such cases, LLMs often insert a synonym or semantically similar word, preserving meaning, but failing to match the ground truth. VLMs exhibit greater robustness across all error types, but insertions remain the most challenging category, even when the source image is highly degraded, as visual information can be ambiguous.

In Figure 5, we present selected examples of corrections made by Llama 3.1 8B and Llama 3.2V 11B that illustrate common occurrences. Example 1 highlights a scenario where the source image is noisy, leading to poor OCR output. In this case, the LLM hallucinates an inaccurate correction based on the portions it can recognize, whereas the VLM accurately restores the text. Example 2 demonstrates an OCR engine error where text from

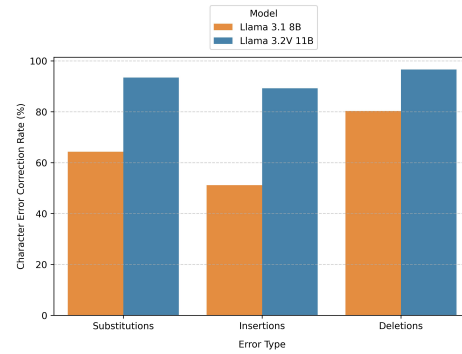


Figure 4: Correction rate on different error types

a neighbouring article has mistakenly been transcribed, introducing unintended words. While the VLM can localize the OCR text and remove the additional word, the LLM is unable to identify and correct this error. Example 3 showcases a limitation of VLMs, where a folded page causes a visual distortion in the source image, resulting in a mistake during the correction process. These examples showcase the benefits of VLMs for post-OCR correction while also indicating that there is still room for further improvement.

6 Conclusion

In this work, we demonstrate the potential of vision-language models (VLMs) to improve the process of post-OCR correction by integrating the source image alongside OCR text. By comparing Llama 3.2 Vision and Qwen 2.5 VL to their LLM counterparts, we show that a multimodal approach is superior, achieving an approximately 90% reduction in character and word error rates. Optimal results are obtained when the VLMs are given both patches of the original document image and corresponding noisy OCR text as inputs, demonstrating substantive contributions from both modalities. Our results further indicate that VLMs perform well with limited training data and surpass LLMs on larger OCR text sequences. Our pipeline for data preparation can work effectively within existing digital archive infrastructures, re-using existing word-image bounding box information for search and display purposes, enabling continuous enhancements in OCR quality over time. These findings suggest that VLM-based post-OCR correction is a robust and effective solution for improving the usability of digital archives, while preserving their existing metadata.

Example 1 - LLM hallucinates correction
1 contained in the warrant produced. When arrested 2 the accused said, "I have been out of jurisdiction. I OCR Text: -Wh'eilarreaed r the aoeiisid sah, 't i have'been out of jucildiotiotti.; LLM: While I was in the accused's house, I have been out of my mind. VLM: When arrested the accused said, "I have been out of jurisdiction. Ground Truth: When arrested the accused said, "I have been out of jurisdiction.
Example 2 - VLM removes extraneous text
"Really, gentlemen, the court is being turned into a liver chamber of horrors." The jury eventually found a been OCR Text: "Really, gentlemen, the court is being turned into a liver chamber of horrors." LLM: "Really, gentlemen, the court is being turned into a liver chamber of horrors." VLM: "Really, gentlemen, the court is being turned into a chamber of horrors." Ground Truth: "Really, gentleman, the court is being turned into a chamber of horrors."
Example 3 - VLM misreads distorted text
you well, you are Marsilius!" This apostrophe produced a profound effect, both upon the prisoner and the audi- ence. The Jury acquitted Grig, but fov Marsilius Ni- OCR Text: This apostrophe „rodued a plofound effect, bothi upon the prisoner arc;t the audi- ence. LLM: This apostrophe produced a profound effect, both upon the prisoner and the audience. VLM: This apostrophe produced a profound effect, both upon the prisoner arid the audience. Ground Truth: This apostrophe produced a profound effect, both upon the prisoner and the audience.

Figure 5: Examples of model corrections

Limitations

Our experiments are limited to the BLN600 dataset, which contains printed newspaper articles written in English from the 19th century. Since large generative models are predominantly trained in English and newspaper text follows a relatively structured format, adapting our approach to low-resource languages, different time periods or more complex documents may require further adjustments. The creation of paired training data depends on the accuracy of word coordinates generated by OCR engines, which is sufficient in our case but may be unreliable for highly degraded texts, necessitating alternative alignment strategies. In addition, our

approach assumes the availability of aligned image-text data, which may not be present in some legacy corpora where only plain text has been preserved. We focus on medium-sized models to ensure practicality for institutions with limited computational resources but larger models could further enhance performance. Finally, we do not provide a quantitative cost comparison between our solution and full re-OCR, because a fair comparison would need to account for several factors, including computational requirements, integration effort, and the cost of adapting existing systems to a new format. These factors cannot be reliably quantified in an initial study whose motivation is to establish the technical viability and benefits of a multimodal post-OCR correction approach. Further work is needed to carry out a quantitative cost-benefit comparison between full re-OCR of an existing digital archive, with attendant costs and performance issues arising from recomputing word-level image grounding data, and our lighter-weight approach of multimodal post-OCR correction, which reuses existing word-level grounding metadata in correcting the OCR text, allowing current archive infrastructure to remain largely unchanged, though at the cost of corrected text being grounded at the image patch level, rather than word level.

Reproducibility

BLN600 is made publicly available by Booth et al. (2024) under the CC BY-NC-ND 4.0 license. The source code is available at https://github.com/Shef-AIRE/vlms_post-ocr_correction. The image patch dataset and fine-tuned models will be released on Hugging Face at <https://huggingface.co/pykale>.

Acknowledgements

The authors would like to thank Michael Pidd, Robert Shoemaker and Callum Booth of the School of History, Philosophy and Digital Humanities at the University of Sheffield for stimulating our interest in the problem of post-OCR text correction. We would also like to thank Gale, and specifically Chris Houghton, for allowing us to release the BLN600 dataset and for helpful discussions on the technical and business case issues surrounding reuse of word level grounding metadata while correcting word level transcription. X.L. and H.L. acknowledge funding from EPSRC (Grant UKRI396).

References

- Chantal Amrhein and Simon Clematide. 2018. [Supervised OCR Error Detection and Correction Using Statistical and Neural Machine Translation Methods](#). *Journal for Language Technology and Computational Linguistics*, 33(1):49–76.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.
- Melodee Beals and Emily Bell. 2020. [The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges](#). Loughborough University.
- Angel Beshirov, Milena Dobрева, Dimitar Dimitrov, Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2025. Post-ocr text correction for bulgarian historical documents. *International Journal on Digital Libraries*, 26(1):4.
- Callum William Booth, Alan Thomas, and Robert Gaizauskas. 2024. [BLN600: A parallel corpus of machine/human transcribed nineteenth century newspaper texts](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2440–2446, Torino, Italia. ELRA and ICCL.
- Emanuela Boros, Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, and Frédéric Kaplan. 2024. [Post-correction of historical text transcripts with large language models: An exploratory study](#). In *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.
- Jacob Carlson, Tom Bryan, and Melissa Dell. 2024. [Efficient OCR for building a diverse digital history](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8105–8115, Bangkok, Thailand. Association for Computational Linguistics.
- Yung-Hsin Chen and Phillip B. Ströbel. 2024. [TrOCR Meets Language Models: An End-to-End Post-correction Approach](#). In *Document Analysis and Recognition – ICDAR 2024 Workshops*, pages 12–26. Springer Nature Switzerland.
- Yung-Hsin Chen and Yuli Zhou. 2023. [Enhancing OCR Performance through Post-OCR Models: Adopting Glyph Embedding for Improved Correction](#). *arXiv preprint arXiv:2308.15262*.
- Guillaume Chiron, Antoine Doucet, Mickael Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017a. [Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information](#). In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–4.
- Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux. 2017b. [ICDAR2017 Competition on Post-OCR Text Correction](#). In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pages 1423–1428.
- Giorgia Crosilla, Lukas Klic, and Giovanni Colavizza. 2025. Benchmarking large language models for handwritten text recognition. *Journal of Documentation*, 81(7):334–354.
- Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, Yue Zhang, Yubo Zhang, Handong Zheng, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. PaddleOCR-VL: Boosting multilingual document parsing via a 0.9B ultra-compact vision-language model. *arXiv preprint arXiv:2510.14528*.
- Florian Debaene, Aaron Maladry, Els Lefever, and Veronique Hoste. 2025. [Evaluating transformers for OCR post-correction in early Modern Dutch theatre](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10367–10374, Abu Dhabi, UAE. Association for Computational Linguistics.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient Fine-tuning of Quantized LLMs](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.
- Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. 2020. [Pp-ocr: A practical ultra lightweight ocr system](#). *arXiv preprint arXiv:2009.09941*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. [The Llama 3 Herd of Models](#). Technical report, Meta.
- Gale. 2025. [British Library Newspapers](#).
- Gavin Greif, Niclas Griesshaber, and Robin Greif. 2025. Multimodal llms for ocr, ocr post-correction, and named entity recognition in historical documents. *arXiv preprint arXiv:2504.00414*.
- Shuhao Guan and Derek Greene. 2024. Advancing post-ocr correction: A comparative study of synthetic data. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6036–6047.

- Shuhao Guan, Cheng Xu, Moule Lin, and Derek Greene. 2024. [Effective synthetic data and test-time adaptation for OCR correction](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15412–15425, Miami, Florida, USA. Association for Computational Linguistics.
- Rose Holley. 2009. [How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs](#). *D-Lib Magazine*, 15(3/4).
- Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. [Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents](#). In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 04, pages 19–24.
- Seorin Kim, Julien Baudru, Wouter Ryckbosch, Hugues Bersini, and Vincent Ginis. 2025. [Early evidence of how llms outperform traditional systems on ocr/htr tasks for historical records](#). *arXiv preprint arXiv:2501.11623*.
- Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. [TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13094–13102.
- Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024. [OCR-Bench: on the hidden mystery of OCR in large multi-modal models](#). *Science China Information Sciences*, 67(220102).
- Viktoria Löfgren and Dana Dannélls. 2024. [Post-OCR correction of digitized Swedish newspapers with ByT5](#). In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLjL 2024)*, pages 237–242, St. Julians, Malta. Association for Computational Linguistics.
- Ayush Maheshwari, Nikhil Singh, Amrith Krishna, and Ganesh Ramakrishnan. 2022. [A benchmark and dataset for post-OCR text correction in Sanskrit](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6258–6265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Laura Manrique-Gomez, Tony Montes, Arturo Rodriguez Herrera, and Ruben Manrique. 2024. [Historical ink: 19th century Latin American Spanish newspaper corpus with LLM OCR correction](#). In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 132–139, Miami, USA. Association for Computational Linguistics.
- Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. [Survey of Post-OCR Processing Approaches](#). *ACM Computing Surveys*, 54(6):1–37.
- Juan Antonio Ramirez-Orta, Eduardo Xamena, Ana Maguitman, Evangelos Milios, and Axel J. Soto. 2022. [Post-OCR Document Correction with Large Ensembles of Character Sequence-to-Sequence Models](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11192–11199.
- Christophe Rigaud, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux. 2019. [ICDAR 2019 Competition on Post-OCR Text Correction](#). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1588–1593.
- Sina Semnani, Han Zhang, Xinyan He, Merve Tekgürler, and Monica Lam. 2025. [Churro: Making history readable with an open-weight large vision-language model for high-accuracy, low-cost historical text recognition](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34765–34812.
- Ray Smith. 2007. [An Overview of the Tesseract OCR Engine](#). In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 629–633. INSTICC, IEEE Computer Society.
- Elizabeth Soper, Stanley Fujimoto, and Yen-Yun Yu. 2021. [BART for post-correction of OCR newspaper text](#). In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 284–290, Online. Association for Computational Linguistics.
- Phillip Ströbel and Simon Clematide. 2019. [Improving ocr of black letter in historical newspapers: the unreasonable effectiveness of htr models on low-resolution images](#). *Digital Humanities*.
- Alan Thomas, Robert Gaizauskas, and Haiping Lu. 2024. [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*, pages 116–121, Torino, Italia. ELRA and ICCL.
- Daniel van Strien, Kaspar Beelen, Mariona Coll Ardanuy, Ksra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. [Assessing the Impact of OCR Quality on Downstream NLP Tasks](#). In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1: AR-TIDIGH*, pages 484–496. INSTICC, SciTePress.
- L Wilms, R Nijssen, and T Koster. 2020. [Historical newspaper ocr ground-truth data set](#). *KB Lab: The Hague*.

- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Nianheng Wu. 2023. [Multimodal OCR post-correction on German historical documents](#). Master’s thesis, University of Stuttgart.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. [Qwen2.5 Technical Report](#). Technical report, Qwen.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. [A Survey on Multimodal Large Language Models](#). *National Science Review*, 11(12).
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinu Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. [LIMA: Less Is More for Alignment](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.

A Prompt Templates

User: I will provide you with OCR output. Please fix the errors in the provided text. Return only the corrected text. Do not include any additional commentary or explanations. If you understand, please reply with OK.

Assistant: OK

User: __RO BBERY AT A BARONET'S. V

Assistant: ROBBERY AT A BARONETS.

User: {OCR}

Figure 6: One-shot prompt for LLMs

User: I will provide you with OCR output and source image. Please fix the errors in the provided text. Return only the corrected text. Do not include any additional commentary or explanations. If you understand, please reply with OK.

Assistant: OK

User: __RO BBERY AT A BARONET'S. V

Assistant: ROBBERY AT A BARONETS.

User: {IMG} + {OCR}

Figure 7: One-shot prompt for VLMs

B Training Details

```
<|im_start|>system
You are a helpful assistant, designed to
correct errors in OCR text. Users will
provide OCR text, you will generate a
correction.<|im_end|>

<|im_start|>user
__RO BBERY AT A BARONET'S. v<|im_end|>

<|im_start|>assistant
ROBBERY AT A BARONETS.<|im_end|>
```

Figure 8: Fine-tuning data for LLMs

```
<|im_start|>system
You are a helpful assistant, designed to
correct errors in OCR text. Users will
provide OCR text and source image, you will
generate a correction.<|im_end|>

<|im_start|>user
<|vision_start|><|image_pad|><|vision_end|>
__RO BBERY AT A BARONET'S. v<|im_end|>

<|im_start|>assistant
ROBBERY AT A BARONETS.<|im_end|>
```

Figure 9: Fine-tuning data for VLMs

We structure our dataset samples as conversations, where the user provides OCR text and the model generates the ground truth. As different models separate messages with distinct tokens, we make use of Hugging Face’s chat templating feature. Figures 8 and 9 present the same sample, formatted for Qwen models, which use the popular ChatML format, where `<|im_start|>role` marks the beginning of a message and `<|im_end|>` denotes the end. Note that the VLM template includes additional tokens, `<|vision_start|>`, `<|image_pad|>` and `<|vision_end|>`, to indicate the presence of image input to the model. Table 6 summarizes the key hyperparameters for fine-tuning LLMs and VLMs.

LLM	
LoRA Alpha	16
LoRA Rank	16
LoRA Dropout	0.05
Training Epochs	1
Batch Size	8
Learning Rate	2e-4
Max Sequence Length	1024

VLM	
LoRA Alpha	16
LoRA Rank	8
LoRA Dropout	0.05
Training Epochs	1
Batch Size	32
Learning Rate	2e-4
Max Sequence Length	1024

Table 6: Training hyperparameters

C Comparison with VLM-based OCR method

Table 7 compares different uses of VLMs to highlight the advantage of multimodal post-OCR correction. As a baseline, we include PaddleOCRVL (Cui et al., 2025), a recent VLM-based OCR system that transcribes images directly using its official pre-trained model. Our models achieve substantially larger reductions in CER and WER, demonstrating that combining visual information with existing OCR text is an effective strategy for improving transcription quality in historical documents.

D Masking image patches

Due to the column layout of newspapers, the image patch for a given sequence often contains ex-

Model	CER _r (%)	WER _r (%)
Llama 3.2V 11B	92.03	91.40
Qwen 2.5 VL 7B	89.27	90.28
PaddleOCR-VL	52.32	71.97

Table 7: Model performance comparison with VLM-based OCR method

cess text from preceding and succeeding sequences. This additional text may affect the model’s ability to accurately localize the relevant OCR text, potentially impacting performance. To explore this, we apply a masking strategy to remove the excess text from image patches. As we have the coordinates for each word in the OCR text sequence, we simply filter out content outside of these bounding boxes, effectively masking regions that do not correspond with words in a given sequence, as shown in Figure 10. This ensures that the model only processes image information directly associated with the given OCR text. However, masking causes a slight performance decrease, as shown in Table 8. The model receives an OCR sentence and an image patch including that sentence and perhaps surrounding words, but outputs only the corrected OCR sentence. Removing this surrounding visual information reduces the performance, indicating that local context, such as neighbouring words and document structure, benefits post-OCR correction. For highly noisy OCR with inaccurate word coordinates, masking may also remove the actual text.

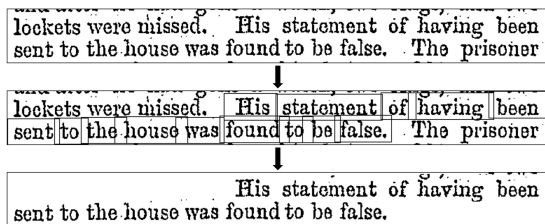


Figure 10: Image before (top) & after (bottom) masking

Model	Image	CER _r (%)	WER _r (%)
Llama 3.2V 11B	original	92.03	91.40
	masked	87.19	89.05
Qwen 2.5 VL 7B	original	89.27	90.28
	masked	86.16	87.67

Table 8: Model performance after masking images