Unlocking History: LLM Agents vs. HTR/OCR for Ancient Manuscript Transcriptions

Seorin Kim, Julien Baudru, Hugues Bersini & Vincent Ginis

Abstract: This study explores the ability of Large Language Models (LLMs) to transcribe historical handwritten documents, comparing their performance to traditional OCR/HTR systems. Focusing on Character Error Rate (CER) and layout replication, the research examines whether LLMs can match or exceed conventional methods and how their performance can be optimized, shedding light on their potential and limitations in this complex task.

Keywords: Large Language Models, Historical Documents, Handwritten Text Recognition, Optical Character Recognition & Character Error Rate

1 Introduction

The transcription of historical documents is a crucial yet challenging task, often complicated by the handwritten nature of such texts. These documents, typically produced by hand, are not only difficult for humans to decipher but also require significant time and effort for manual transcription. While traditional Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) methods have been the mainstay for text detection in historical manuscripts, recent advancements in generative AI and multimodal Large Language Models (LLMs) present a compelling alternative.

Despite the promising capabilities of LLMs in various tasks, their performance in transcribing historical handwritten documents remains underexplored. A key aspect often evaluated in LLMs is their ability to handle OCR tasks for typed texts. However, the natural progression from this is to question how well LLMs perform when faced with the more complex challenge of handwritten texts.

This study aims to investigate whether LLMs can perform transcription tasks as well as or better than conventional OCR/HTR methods, particularly in terms of Character Error Rate (CER). Furthermore, it explores strategies to enhance LLMs' transcription performance and addresses the inherent black-box nature of LLMs, which often leads to high variance in results. Achieving stable and consistent outcomes is crucial, particularly when the goal is to ensure that LLMs not only recognize text accurately but also faithfully reproduce the document's layout as seen in the image.

While the objective of most text recognition tasks is to accurately understand and extract information from images, this study emphasizes a different goal: the precise replication of what the LLM sees in the image, including the layout and structure. This distinction is critical because it shifts the focus from fine-tuning models to achieve specific outcomes to assessing the inherent capabilities of LLMs in replicating the exact visual presentation of the document. Through this approach, we aim to provide a deeper understanding of the potential and limitations of LLMs in the transcription of historical handwritten texts.

2 Similar works

In this section, we present a summary of the previous research aiming to combine text detection and error correction with language models (Gpt, Bert, etc.).

In [Fuj23], the authors introduce a new method for text recognition called Decoder-only Transformer for Optical Character Recognition (DTrOCR). DTrOCR uses only a decoder, using a pre-trained generative language model, in contrast to traditional encoder-decoder methods. The authors tested whether a successful natural language processing model could be applied to text recognition in computer vision. Their experiments showed that DTrOCR significantly outperformed current state-of-the-art methods in recognising printed, handwritten and scene

text in both English and Chinese.

In [TGL24], the authors address the challenge of poor OCR quality in digitised historical documents, which is a barrier to humanities research. Traditional post-OCR correction methods use sequence-to-sequence models. Instead, the authors propose the use of generative language models with a prompt-based approach. By tuning Llama 2 with prompts and comparing it to a fine-tuned BART model on 19th century British newspaper articles, they demonstrate significant improvements in OCR error correction. Llama 2 achieves a 54.51% reduction in the character error rate, outperforming BART's 23.30% reduction. This approach shows promise for improving the accessibility of historical texts for researchers.

In [Lö23], the author introduces a method to digitize over 100,000 historic plans from the Swiss Archive for Landscaping Architecture using AI models. The approach employs a three-model architecture: a layout model to identify text, an OCR model to extract words, and a named entity recognition (NER) model to label key information. K-means clustering groups text blocks for OCR processing. Various deep-learning models were evaluated, including German BERT for NER, and retrained on the NVIDIA DGX-2 system. The pipeline achieved an F1 score of 48%, with the NER model scoring 86% and the OCR model correctly extracting 54% of words.

In [BER+24] the authors carried out a comparative study of the ability of 14 LLMs to correct transcriptions produced using OCR, HTR and ASR. They then evaluate these corrections by comparing them with ground truths from each document. They conclude that, although GPT4 appears to be the best model among those tested, all the models degrade rather than improve transcriptions. And that, on the whole, LLMs are better at detecting errors than at correcting them, as they are subject to overcorrection.

In [FSM+24], the authors propose a novel tokenized representation of digital ink for online handwriting recognition, addressing the shortcomings of naive OCR with vision-language models (VLMs). This approach, integrating stroke sequences and images, achieves state-of-the-art quality on three public datasets. Their findings show that VLMs benefit from multimodal inputs, that images are crucial when text representations are too long, and that multiple handwriting tasks can be combined effectively. The method is compatible with both parameter-efficient tuning and fine-tuning, suggesting future exploration of various handwriting task combinations in large VLMs.

Finally, in [TWT+24], the authors show that format constraints like JSON or XML significantly impair LLM's reasoning abilities, with stricter constraints leading to greater performance degradation. This highlights a trade-off between maintaining structured outputs and preserving reasoning performance in state-of-the-art LLMs.

To our knowledge, our study is the first to use LLMs to produce transcriptions of ancient texts based directly on their images. Previous studies have focused on using LLMs to correct text errors produced by OCR/HTR from these images.

3 Approach

3.1 Dataset

3.2 Metrics

In our analysis, we consider human transcriptions to be the ground truth (GT), but given the complex nature of the handwriting in the manuscripts studied, these manual transcriptions are prone to error. To evaluate the performance of different methods and models, we employ the Character Error Rate (CER) as a primary metric.

The CER measures the similarity between the predicted transcription and the GT by calculating the edit distance (Levenshtein distance) between the two strings. The edit distance is the minimum number of single-character edits (insertions, deletions, or substitutions) required to change the predicted transcription into the GT. The equation of the CER is given by 1.

$$\text{CER} = \frac{S + D + I}{N} \tag{1}$$

where:

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions,
- N is the total number of characters in the GT.

A lower CER indicates a better match between the predicted transcription and the ground truth, with a CER of 0% representing a perfect match. This metric is particularly suited for evaluating text recognition models

where even minor character errors can significantly impact the readability and accuracy of the transcription.

4 Framework

5 Results

6 Results

7 Conclusion and Future Work

References

[BER+24] Emanuela Boros, Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, and Frédéric Kaplan. Post-correction of historical text transcripts with large language models: An exploratory study. In Yuri Bizzoni, Stefania Degaetano-Ortlieb, Anna Kazantseva, and Stan Szpakowicz, 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, March 2024. Association for Computational Linguistics.

[FSM+24] Anastasiia Fadeeva, Philippe Schlattner, Andrii Maksai, Mark Collier, Efi Kokiopoulou, Jesse Berent, and Claudiu Musat. Representing online handwriting for recognition in large vision-language models, 2024.

[Fuj23] Masato Fujitake. Dtrocr: Decoder-only transformer for optical character recognition, 2023.

[Lö23] Kevin Löffler. *Digitize Historic Architectural Plans with OCR and NER Transformer Models*. Other thesis, OST Ostschweizer Fachhochschule, May 2023. Thesis advisor: Mitra Purandare.

[TGL24] Alan Thomas, Robert Gaizauskas, and Haiping Lu. Leveraging LLMs for post-OCR correction of historical newspapers. In Rachele Sprugnoli and Marco Passarotti, editors, *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 116–121, Torino, Italia, May 2024. ELRA and ICCL.

[TWT+24] Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung yi Lee, and Yun-Nung Chen. Let me speak freely? a study on the impact of format restrictions on performance of large language models, 2024. Equal contribution, Equal advisorship.

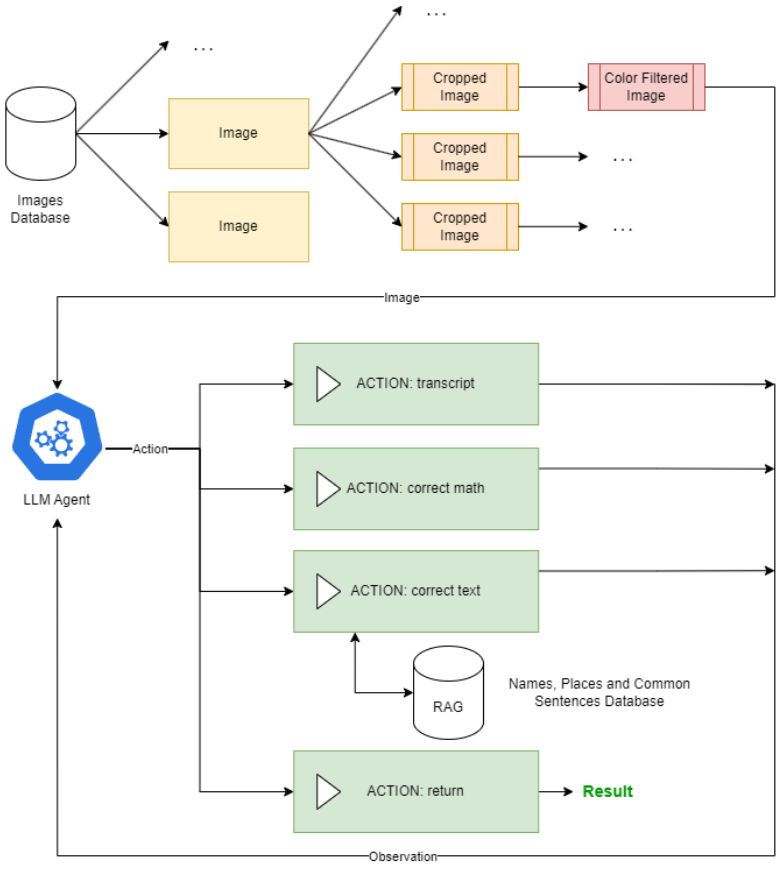

Figure 1: LLM OCR agent