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

Seorin Kim, Julien Baudru, Hugues Bersini & SEORIN SUPERVISOR

## Abstract

Your abstract.

## 1 Introduction

Introduce why it's important to transcript old document ??

The aim of this paper is to see whether generative AI methods, and multimodal LLMs in particular, can be used effectively to detect text in pictures of old administrative documents. All the documents studied here are produced in handwriting that is difficult for humans to decipher, and the manual transcription process tends to take a considerable amount of time. We compare traditional OCR/HTR text detection methods with the results obtained by a multimodal LLM and those obtained by manual annotation.

## 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In [TGL24], the authors addresse the challenge of poor OCR quality in digitised historical documents, which is a barrier to humanities research.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.

Finally, 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.

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.

# 4 Framework

# 5 Results

- Compare LLM vs OCR vs GT (Human)
- Ask LLM to correct extracted text

# 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.

[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.

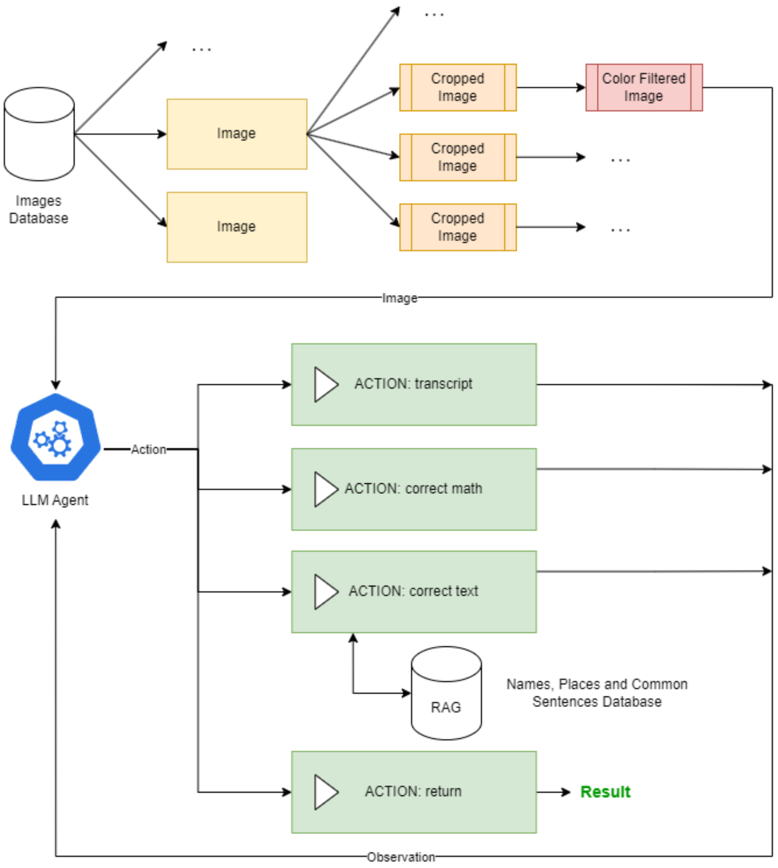

Figure 1: LLM OCR agent