Improving Multilingual Medieval Handwriting Recognition through Multimodal Language Modeling

M. Nmachi Igwe, Saurav K. Aryal PhD
AI4PC Lab, Institute for Human Centered AI

INTRODUCTION

Medieval manuscript transcription is challenging because pages combine historical scripts, multilingual content, visual degradation, and irregular page structure. Our project investigates a document-learning pipeline for the **ICDAR CMMHWR multilingual medieval handwriting task**, focusing on full-page recognition rather than isolated word or line prediction. The goal is to build a system that can read manuscript pages while preserving the broader document context needed for realistic historical OCR/HTR use.

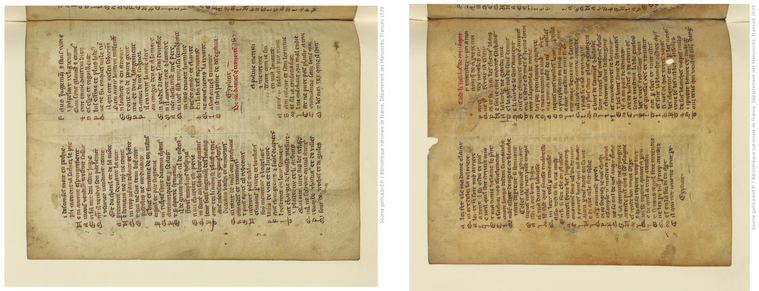

Figure 1: **Samples of Medieval Documents**

METHODS

This project develops a full-page OCR/HTR pipeline for multilingual medieval manuscripts using the CCMHWR26 collection. The system begins by mounting the dataset from Google Drive and parsing manuscript folders organized as mss-###, each containing metadata, page XML annotations, and page images. XML files are processed to extract line-level text, coordinates, and manuscript metadata such as language, century, and script. These line annotations are then reordered and merged into reconstructed **page-level transcriptions**, allowing the dataset to move from archival annotation format into machine-learning training format. Finally, the reconstructed pages are split by manuscript into train and validation sets so that the model is evaluated on unseen manuscripts rather than memorized pages.

Language	Manuscripts	Pages
French	88	833
Latin	125	715
Castillian	52	448
Italian	25	172
Other	10	62

Table 1: **Available languages and page counts in the dataset**

METHODS

The training workflow is implemented in a second notebook, **MMLTr**, which loads the prepared page-level CSV files and converts them into Hugging Face Dataset objects. For the first supervised full-page experiment, the deepseekOcr2 architecture was selected because it provides a stable image-to-text training path within the Hugging Face ecosystem. Each training sample consists of a **full manuscript page image** paired with its reconstructed page_text, enabling the model to learn transcription directly from the page rather than isolated cropped lines. A preprocessing function converts page images to RGB tensors and tokenizes the target manuscript text, while the training loop is managed using Seq2SeqTrainer with gradient accumulation and mixed precision. Validation is tracked during fine-tuning so that CER/WER can later be compared against the earlier inference baseline.

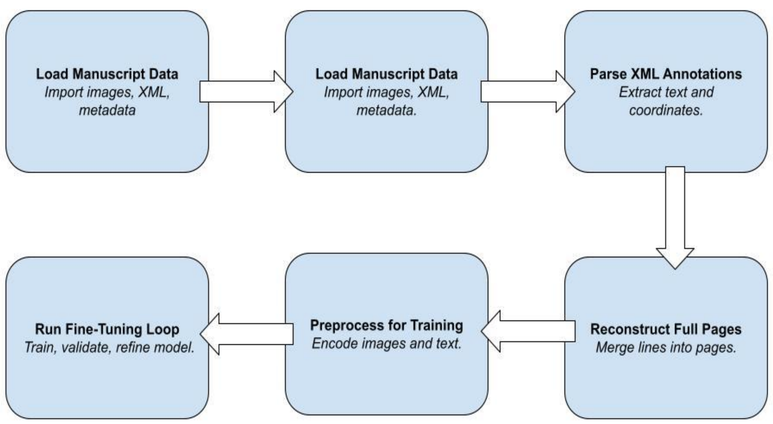

Figure 2: **Overall proposed modeling process**

RESULTS

This work establishes a practical systems pipeline for multilingual medieval manuscript OCR/HTR. By converting XML-aligned manuscript data into reconstructed page-level training targets, the project creates a usable bridge between historical document archives and modern vision-language training workflows. Baseline zero-shot inference showed that prompt-only OCR is not reliable enough for this task, while the new supervised pipeline demonstrates that full-page fine-tuning is feasible. The next phase is to complete larger training runs, improve validation CER/WER, and benchmark the trained system against the baseline.

REFERENCES

[1] Clérice, T., Vlachou-Efstathiou, M., & Chagué, A. (2023). CREMMA Medii Aevi: Literary manuscript text recognition in Latin. Journal of Open Humanities Data, 9, 4.
[2] See SegmOnto's recommendations in Gabay, S., Camps, J.-B., Pinche, A., & Jahan, C. (2021). Segmonto: common vocabulary and practices for analysing the layout of manuscripts (and more). In 16th international conference on document analysis and recognition (ICDAR 2021), et https://segmonto.github.io/.
[3] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.
[4] Kim, G., Hong, T., Yim, M., Park, J., Yim, J., Hwang, W., ... & Park, S. (2021). Donut: Document understanding transformer without ocr. arXiv preprint arXiv:2111.15664, 7(15), 2.
[5] Wei, H., Sun, Y., & Li, Y. (2025). DeepSeek-OCR: Contexts optical compression. arXiv. https://arxiv.org/abs/2510.18234
[6] Wei, H., Sun, Y., & Li, Y. (2026). DeepSeek-OCR 2: Visual causal flow. arXiv. https://arxiv.org/abs/2601.20552

ACKNOWLEDGEMENTS

This research project was supported in part by the Office of Naval Research grant N00014-22-1-2714 and an Amazon Research Award. The work is solely the responsibility of the authors and does not necessarily represent the official view of the sponsors.