Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

LLM Benchmarking for HTR on French Historical Death Records

Gabriel Frossard

gabriel.frossard@utbm.fr

University of Technology of Belfort-Montbéliard

Flavian Theurel

University of Technology of Belfort-Montbéliard

Wissam Al-Kendi

University of Technology of Belfort-Montbéliard

Franck Gechter

University of Technology of Belfort-Montbéliard

Research Article

Keywords: Large Language Model, Handwritten Text Recognition, Historical Demography, Archival Document Analysis

Posted Date: April 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9353651/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at International Journal of Digital Humanities on October 19th, 2026. See the published version at https://doi.org/10.1007/s42803-026-00137-9.

LLM Benchmarking for HTR on French Historical Death Records

Gabriel Frossard[1*], Flavian Theurel[1], Wissam Al-Kendi[1], Franck Gechter[1]

[1]CIAD UR 7533, Université de Technologie de Belfort-Montbéliard, Belfort, France.

*Corresponding author(s). E-mail(s): gabriel.frossard@utbm.fr;
Contributing authors: flavian.theurel@utbm.fr;
wissam.al-kendi@utbm.fr; franck.gechter@utbm.fr;

Abstract

This study evaluates the performance of recent multimodal large language models (LLMs) for the transcription of historical handwritten and printed documents, with a focus on sources commonly used in historical demography. We benchmarked GPT-4O, several QWEN models, and other widely used LLMs against traditional OCR/HTR systems, including Tesseract and Ocelus, on a dataset of archival records containing complex layouts, marginal annotations, and variable handwriting styles. GPT-4O achieved the highest transcription accuracy across both paragraph text and marginal entries, while QWEN emerged as the most effective fully free and unrestricted alternative. Our analysis also highlights the limits of common evaluation metrics: the BLEU score did not consistently correlate with the Character Error Rate (CER), due to BLEU's sensitivity to n-gram overlap rather than character-level fidelity. This distinction is critical for historical sources where spelling preservation and textual exactness are essential. Overall, the results demonstrate the growing potential of LLM-based approaches for large-scale historical data extraction and emphasize the need for continued benchmarking and domain-specific evaluation frameworks as new models rapidly emerge.

Keywords: Large Language Model, Handwritten Text Recognition, Historical Demography, Archival Document Analysis

Statement and Declarations

- **Funding**: This study is part of the Belpop project funded by the Bourgogne Franche Comté region.
- **Conflict of interest/Competing interests**: The authors declare that they have no competing interests.
- **Ethics approval and consent to participate**: Not applicable
- **Consent for publication**: Not applicable
- **Data availability** : All data used are available on the Belfort Archive website: https://archives.belfort.fr
- **Code availability**: The code used for the experiments is available upon request. It will be available in the coming months on this Github repository: https://github.com/gabriel-frossard/Belpop-HTR
- **Author contribution**: All authors contributed to the study conception and design. Data collection, and analysis were performed by Gabriel Frossard. Gabriel Frossard and Flavian Theurel wrote the final manuscript and all authors read and approved it.

1 Introduction

Over the past two decades, massive digitization campaigns have transformed access to archival records essential for historical demography. Parish registers, civil-status documents, census books, tax rolls, and household lists have been scanned and placed online at an unprecedented scale. Yet, despite this progress, a fundamental obstacle persists: the information contained in these images remains largely locked away. Extracting names, dates, occupations, and family relationships from centuries-old handwritten documents continues to require extensive manual transcription, limiting the size of the datasets and the scope of demographic research. Our study faces these challenges because it is part of a project whose goal is to reconstruct population dynamics through the cross-analysis of several archival sources (Heyberger et al. 2024). We aim to extract all the information gathered in civil-state declarations of Belfort, a small industrial town located in eastern France .

Traditional Optical Character Recognition (OCR) and more recent Handwritten Text Recognition (HTR) systems have offered partial solutions, producing high-quality transcriptions for modern handwriting or well-standardized manuscripts. However, historical sources present challenges far beyond the assumptions of these systems. Pages may be irregular, degraded, or ink-stained; scripts vary across regions and time periods; spelling is inconsistent; and abbreviations abound. Even more demanding are complex tabular layouts—such as census grids or multi-column inventories—where visual structure and semantic content intertwine. As a result, OCR/HTR approaches often struggle to achieve usable accuracy on the very documents most critical to historical demography.

In parallel, recent advances in artificial intelligence have given rise to multimodal LLMs capable of processing both images and text. These models promise a fundamentally different approach: rather than learning to transcribe characters, they aim

to interpret entire documents. Vision-enabled LLMs can, in principle, recognize textual content, infer structure, and extract relationships in a single step. Their capacity to "understand" content—despite noise, irregularity, or complexity of layout, complexity—has generated growing interest among researchers seeking to automate the reading of archival materials.

This article explores these questions through a comparative evaluation of LLM-based and OCR/HTR-based approaches applied to historical demographic sources. By analysing their respective strengths, weaknesses, and failure modes, our objective is to assess whether LLMs can complement— or possibly transform—existing workflows for large-scale population reconstruction. More broadly, this study situates emerging AI tools within the methodological landscape of digital historical demography, highlighting both their potential and their current limitations.

2 Related Works

2.1 Key Datasets and Preprocessing challenges

Complex challenges inherent in transcribing medieval texts, particularly those written in Latin, Old French, and Old Spanish, are often addressed (Torres Aguilar and Jolivet 2023). A core challenge tackled by the model introduced is the highly variable nature of medieval handwriting, which includes inconsistent orthography, unique abbreviations, degraded pages, and non-standard punctuation. Additionally, the scarcity of digitized, annotated data for training and evaluation further complicates the development of robust HTR systems. Despite their attempts to mitigate these issues, the model still contends with CERs around 5-10% on out-of-sample datasets, underscoring the difficulty of generalizing across time periods, scribes, and linguistic variations.

In Hodel et al. (2021), authors explore the development of generalizable HTR models capable of transcribing German Kurrent script across various hands and time periods. One of the primary obstacles is the extreme variability in writing style across different scribes, regions, and time periods, which complicates model generalization. The cursive and compact nature of the script further exacerbates difficulties in segmentation and character discrimination, especially when characters are deformed or connected. Moreover, the historic degradation of source materials—such as ink bleed, fading, and physical damage—introduces noise that traditional OCR systems are ill-equipped to handle. Even state-of-the-art recognition engines like HTR+ or PyLaia struggle without robust datasets that capture the full spectrum of script variability.

A few studies recently focused on the state of HTR, with a particular attention to French-language and historical documents (AlKendi et al. 2024). This survey categorizes HTR systems based on employed techniques, utilized datasets, publication years, and recognition levels. It highlights the evolution from traditional methods like Hidden Markov Models (HMMs) to modern deep learning approaches, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Transformer-based models. The challenges posed by historical documents, such as variations in writing styles, overlapping characters, and marginal annotations, are also discussed, using

the Belfort civil archives as a case study. Concerning this historical corpus, image preprocessing and document layout analysis play a crucial role in improving transcription accuracy (AlKendi et al. 2025). This paper highlights that preprocessing is not merely a preliminary step but a determining factor in extracting reliable demographic information from heterogeneous sources.

2.2 Model Architectures

In Prasad et al. (2018), Information Extraction (IE) techniques on 17$^{\text{th}}$ century Catalan marriage records from the Esposalles dataset (Romero et al. 2013) are evaluated. A pipeline approach is implemented, initially applying a Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BLSTM) architecture for HTR, achieving a CER of approximately 5%. Subsequently, various Named Entity Recognition (NER) models, including Conditional Random Fields (CRFs) and BLSTM models, are employed to extract semantic information. These BLSTM models utilizing n-gram features outperform CRF-based models, even when trained on noisy HTR outputs. The best-performing model reaches a 95.5% accuracy on blind test data, surpassing previous state-of-the-art methods.

Other architectures were tested, such as the combination of a Joint Attention Network and Knowledge Distillation (Hamdan et al. 2024). The model, HTR-JAND, utilizes backbone convolutional layers for feature extraction, a Transformer encoder to capture contextual information, and a Connectionist Temporal Classification (CTC). Evaluated on standard datasets such as IAM (Marti and Bunke 2002), RIMES (Grosicki et al. 2009) and Bentham (Causer and Wallace 2012), the model achieves state-of-the-art CER and demonstrates the potential of combining a lightweight architecture with advanced learning techniques.

In Li et al. (2025), a Vision Transformer-based framework specifically designed for HTR is proposed. Unlike traditional CNN-RNN approaches, HTR-VT leverages the self-attention mechanism of transformers to capture both local and global dependencies in handwritten text images. The model employs a 2D positional encoding scheme and an encoder-decoder structure that directly maps image features to transcriptions. Evaluated on standard benchmarking datasets, HTR-VT achieves state-of-the-art results in terms of both CER and Word Error Rate (WER).

2.3 Large Language Models and Multi-modal Large Language Models

LLMs are increasingly being evaluated for HTR tasks. A benchmarking study (Crosilla et al. 2025) assesses the ability of models like GPT-4O-2024-08-06, CLAUDE-3-5-SONNET-20240620 or QWEN2-VL-7B to handle various handwriting styles and languages across different datasets. The study demonstrates that LLMs can achieve competitive results when compared with traditional HTR models, often reducing the reliance on manually annotated training data. The authors highlight the challenges and

potential of using LLMs to perform HTR tasks, emphasizing that LLMs, with proper fine-tuning, can outperform conventional OCR approaches in many HTR scenarios.

In Gutteridge et al. (2025), a systematic evaluation of LLMs for transcribing handwritten historical documents, especially for multi-page documents in a zero-shot setting, is presented. They address the limitations of traditional OCR systems, which often struggle with handwriting and require extensive labelled data for fine-tuning. They evaluate different prompting strategies using GPT-4.O and GPT-4.O-MINI to perform HTR, and propose a novel method called "+FIRST PAGE" where the LLM is provided with the OCR output for the entire document and only the image of the first page. This approach leverages shared formatting and contextual features across pages, improving transcription accuracy without the computational cost of processing all images. Experiments on a multi-page version of the IAM Handwriting Database demonstrate that "+FIRST PAGE" enhances performance, even on out-of-sample text, by extrapolating formatting and OCR error patterns from a single page. This work highlights the potential of LLMs to effectively transcribe multi-page handwritten documents by utilizing contextual information, offering a cost-effective alternative to traditional OCR methods.

Tabular layout historical document transcription have also been investigated in Kim et al. (2025) and Nunes et al. (2025). They compare the performance of models such as —GPT-4O and CLAUDE 3.5 SONNET—with traditional OCR/HTR systems: EASYOCR, KERAS OCR, or TATR + OCR. Findings indicate that LLMs, even in zero-shot and two-shot scenarios, outperform traditional OCR/HTR methods. Specifically, GPT-4O excels in line-by-line transcriptions, while CLAUDE 3.5 SONNET performs best on whole-scan inputs. Notably, LLMs demonstrate superior accuracy but can struggle with precise structural reconstruction.

Recently, several studies benchmarked LLMs on a specific dataset. Sohail et al. (2024) evaluated the ability of GPT-4O to perform an OCR transcription on low-resource scripts. With a zero-shot strict setting and no find-tuning, results showed that visual degradations strongly impact performance but the error rates are lower for latin characters.

Although more studies now assess LLMs on historical HTR tasks, the limited diversity of the evaluations/benchmarks still restricts our understanding of their robustness and generalizability.

3 Methods

Our strategy aims to assess the performance of several LLMs on a custom dataset, in order to evaluate their ability to handle specific tasks outside of standardized corpora. As the funding for proprietary models requiring a subscription or credits to use their API is limited, we used GPT-4O and GEMINI-1.5-PRO but we tried to focus on free-to-use and an open-source LLMs such as QWEN-VL series (see the exhaustive list in Table 1).

Another concern arises from the nature of the dataset itself. The dataset used is not confidential, it can be found on the city archives website, but the next one used in the project will be. As a result, it is not possible for us to use the APIs of proprietary models, which may retain input data for future training. This constraint significantly influenced our choice of models and infrastructure, favouring local, self-hosted solutions that ensure data privacy and compliance with ethical standards. All benchmarks had to be run on two RTX A6000 GPUs, which further constrained our choice of models due to memory and computational limitations. This hardware constraint, combined with the need for local inference to preserve data confidentiality, required us to carefully balance model size, performance, and feasibility throughout our experiments.

Table 1 Benchmarked Models

Model	Launch date
GPT-4.o	May 2024
Gemini-1.5-pro	May 2024
Gemini-2.0-flash	January 2025
Qwen-2.5-VL 70B	January 2025
Qwen-3-VL 30B	April 2025
Llama-3.2 Vision	September 2024
Mistal (le chat)	November 2024
Claude-3.5	June 2024

3.1 Dataset

Our corpus is composed of death records of the city of Belfort spanning from 1807 to 1919. It consists of 9775 scanned double page, representing around 38000 declarations of death in French. This dataset presents a duality, combining OCR and HTR issues because from 1807 to 1884 (see Figure 1 left), the acts are handwritten but from 1885 to 1919, the declarations are hybrid because the booklets were preprinted (Figure 1 right), hence containing printed and handwritten characters. We were provided 370 labelled declarations to assess the LLMs performances. These labels were selected by sampling across the images in order to obtain ground truth that reflects the data as accurately as possible.

3.2 Experimental setup

Our strategy lies with a zero shot prompting of single page images. To create this dataset we first applied a preprocessing to split, resize and convert the images to greyscale. This step was necessary to shorten the processing time because for instance, a raw double page takes several minutes to be processed by a LLM. We used at the beginning a simple prompt but the results were often far from accurate, and with a high degree of hallucination. We opted then for a complex prompt describing the

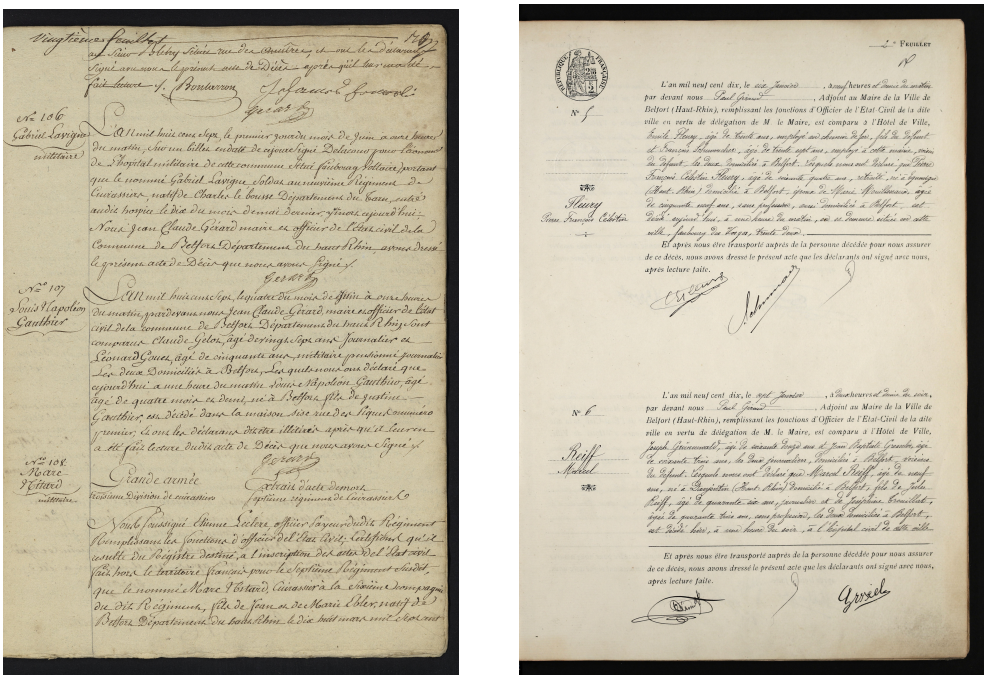


Fig. 1 Pages of death declarations from 1807 (left) and 1910 (right)

complexity of the input images (following listing). Our experiments were constrained by the limited availability of free tokens for Llama, Mistral, and Claude; however, the full dataset was used for the remaining evaluated models. For Qwen, several model sizes were tested beforehand, and we proceeded with those that yielded the highest accuracy. In parallel, we processed the same images using two online OCR systems: Ocelus and Tesseract.

```
You are a model specialized in the analysis of archival
images, civil registry records, and textual segments.
The provided image is a scan of a double-page from a
death register booklet.
Each record consists of the declaration number, the
deceased name and given name(s), optional
annotations, and the main declaration paragraph.
The number, the identity of the deceased, and the
annotations are located in the left margin.
Read and accurately transcribe the text visible in this
area.
Preserve the original spelling (including archaic forms).
Do not complete or guess any parts that are not legible.
Do not mind the signatures
```

3.3 Metrics

To evaluate the performance our strategy, three metrics were used to compare precision across different scopes. Specifically, we define two evaluation scopes: (1) a paragraph scope, and (2) a margin scope, focusing on fields that define individual identity (e.g., number, names, etc.).

3.3.1 Character Error Rate (CER)

The Character Error Rate Levenshtein (1966); Penarrubia et al. (2025); Mayr et al. (2025) quantifies the edit distance between the predicted text and the ground truth at the character level, after normalization and whitespace removal. It's defined as the Levenshtein distance (i.e., the minimum number of insertions, deletions, and substitutions) divided by the length of the ground truth:

$$\text{CER} = \frac{D_c(\text{GT}, \text{PRED})}{\max(1, |\text{GT}|)} \tag{1}$$

Where D_c is the character-level Levenshtein distance and $|GT|$ is the length of the ground truth string. This metric is especially sensitive to small transcription errors and is commonly used in OCR and HTR.

3.3.2 Word Error Rate (WER)

The Word Error Rate Levenshtein (1966); Hunt (1990); Mayr et al. (2025) computes the Levenshtein distance between tokenized ground truth and predicted word sequences. It evaluates the number of word-level substitutions, insertions, and deletions

required to match the prediction to the reference:

$$\text{WER} = \frac{D_w(W_{\text{GT}}, W_{\text{PRED}})}{\max(1, |W_{\text{GT}}|)} \tag{2}$$

Where D_w is the word-level Levenshtein distance, and W_{GT}, W_{PRED} are the tokenized word sequences from the normalized ground truth and prediction. WER provides an interpretable measure of semantic fidelity and is particularly informative when word boundaries are important.

3.3.3 BLEU Score (Bilingual Evaluation Understudy)

The BLEU score Sohail et al. (2024) is an automatic metric used to evaluate the similarity between system-generated text and one or more reference texts. It computes the geometric mean of modified n-gram precisions (typically for $n = 1 \dots 4$), combined with a brevity penalty that discourages overly short hypotheses. Formally, the BLEU score is defined as:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{3}$$

where p_n denotes the modified n-gram precisions and $w_n = \frac{1}{N}$ are uniform weights. The brevity penalty BP is given by:

$$BP = \begin{cases} 1, & \text{if } c > r, \\ \exp\left(1 - \frac{r}{c}\right), & \text{if } c \leq r, \end{cases} \tag{4}$$

with c the length of the candidate text and r the effective reference length. BLEU ranges from 0 to 1 and provides an estimate of how well multi-word sequences in the system output match those in the reference, making it widely used in machine translation, OCR/HTR, and other text-generation tasks. This metric was only used for the paragraph evaluation.

4 Results

The results of the comparison, depicted in Table 2 showcase that LLMs outperform classical OCR techniques as in Kim et al. (2025). Across the set of historical documents evaluated, GPT-4O produced the most accurate transcriptions for both the main text paragraphs and the marginal entries, consistently outperforming all other tested systems. Among the freely available and unrestricted solutions, QWEN emerged as the best-performing model, offering a strong open-access alternative despite slightly lower accuracy than GPT-4O but similar to GEMINI-1.5-PRO. As for the version of Llama, Mistral and Claude tested, they do not seem relevant to be used on our custom dataset.

Interestingly, the BLEU score did not correlate perfectly with the CER metric. This discrepancy arises because BLEU is designed to measure n-gram overlap and is therefore sensitive to phrase-level similarity, whereas CER captures fine-grained

Model	N Images	WER(%)	CER(%)	BLEU
Ocelus	370	48.5	34.3	0.58
Transkribus	370	41.3	29.5	0.57
GPT-4.o	370	**3.0**	**3.5**	**0.82**
Gemini-1.5-Pro	370	6.8	5.7	0.76
Gemini-2.0-Flash	370	9.9	7.8	0.72
Qwen-2.5-VL 70B	370	7.3	6.1	0.72
Qwen-3-VL 30B	370	6.9	6.0	0.75
LLaMA-3.2 Vision	40	9.1	7.4	0.69
Mistral	20	16.2	10.9	0.57
Claude-3.5	20	13.5	9.4	0.62

Table 2 Paragraph Metrics

Model	N Images	WER(%)	CER(%)
Ocelus	370	53.6	28.2
Transkribus	370	49.2	34.4
GPT-4.o	370	**5.7**	**4.9**
GEMINI-1.5-pro	370	8.1	6.8
GEMINI-2.0-flash	370	12.2	9.8
Qwen-2.5-VL 70B	370	8.7	7.1
Qwen-3-VL 30B	370	7.8	7.1
LLAMA-3.2-vision	40	9.6	8.7
Mistral	20	22.7	15.5
Claude-3.5	20	16.3	11.5

Table 3 Margin Metrics

character-level deviations. As a result, a model may achieve a relatively high BLEU score while still introducing small but frequent character errors that significantly affect CER—an important distinction when working with historical documents where fidelity to the original spelling is essential.

Regarding marginal data extraction, Table 3 shows that GPT-4O still outperforms the other LLMs. As well as for the paragraph results, QWEN-3-VL and GEMINI-1.5-PRO give similar outcomes. However, The CER reveals an additional observation. Specifically, the margin-CER is higher than the paragraph-CER because the handwritten numerical identifiers are not well recognized by the LLMs. This is likely due to the substantial variation and calligraphic styling found in the handwritten digits (A), which increases ambiguity and makes them harder for the models to decode accurately.

5 Discussion and future works

The results obtained in this study show that GPT-4O currently achieves the highest accuracy on OCR/HTR tasks applied to historical demographic sources, outperforming all other evaluated models. Concerning free-to-use models, Qwen particularly in its larger vision-enabled configurations, ranks behind GPT-4O and GEMINI-1.5-PRO

but demonstrates competitive performance, especially given its open availability and flexible deployment options. These findings highlight the rapid progress of multimodal LLMs in handling complex handwritten archival material, traditionally considered challenging even for state-of-the-art OCR systems. Nonetheless, the landscape of large vision-language models is evolving quickly. Several more recent models—released after the completion of our experiments—are likely to further shift performance boundaries. Future work should include a systematic evaluation of QWEN-3-VL with multi-shot prompting, as well as a more granular analysis of error types to better understand the strengths and limitations of LLM-based transcription in historical research workflows. A study could also compare LLMs on HTR task with the recently published DANIEL (Constum et al. (2024)) noted for its strong HTR and IE performances on similar corpus.

6 Conclusion

This study demonstrates the substantial progress achieved by recent multimodal LLMs in the transcription of historical handwritten and printed documents. GPT-4O consistently delivered the most accurate results across both paragraph text and marginal entries, confirming its robustness on complex archival material. QWEN and GEMINI-1.5-PRO emerged as strong alternatives—particularly QWEN, which offers unrestricted access while achieving accuracy comparable to some proprietary models. In contrast, the tested versions of Llama, Mistral, and Claude did not yield results that make them suitable for this specific dataset, highlighting the importance of domain-sensitive evaluation.

Our analysis also underscores the need for complementary metrics: the weak correlation between BLEU and CER indicates that character-level fidelity is not fully captured by n-gram similarity, especially in historical contexts where spelling precision is crucial. Moreover, the higher error rates observed in marginal entries reveal a persistent limitation in recognizing handwritten numerical identifiers, likely caused by calligraphic variability.

Overall, these findings confirm that LLM-based approaches are becoming competitive, and in some cases superior, to traditional OCR pipelines for historical transcription. As models continue to evolve rapidly, future work should include testing newly released architectures, refining evaluation protocols, and exploring domain-specific tuning strategies to further reduce errors—particularly in challenging marginal and numerical fields essential for historical demography.

Acknowledgements. We would like to thank the Belfort City Archives for granting us access to the civil registry documents. We are also grateful to the students who contributed to the manual annotation process : Adèle Carter, Elise Pereira da Silva, Gueric Nogret and Achille Vandebrouck.

Appendix A Examples of numbers

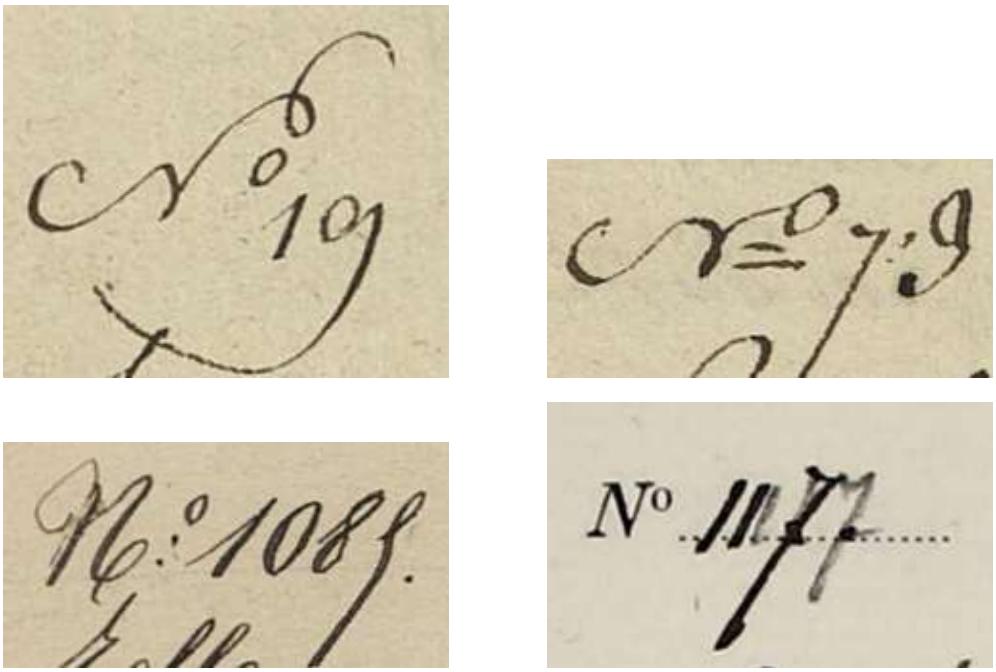


Fig. A1 Number examples from the dataset

References

AlKendi W, Gechter F, Heyberger L, et al (2024) Advancements and challenges in handwritten text recognition: A comprehensive survey. Journal of Imaging 10(1). https://doi.org/10.3390/jimaging10010018, URL https://www.mdpi.com/2313-433X/10/1/18

AlKendi W, Gechter F, Heyberger L, et al (2025) Optimized Data Structuring and Preprocessing Techniques for Belfort Civil Registers of Birth Transcription. SN Computer Science 6(5):550. https://doi.org/10.1007/s42979-025-03987-2, URL https://doi.org/10.1007/s42979-025-03987-2

Causer T, Wallace V (2012) Building a volunteer community: Results and findings from transcribe bentham. Digit Humanit Q 6. URL https://api.semanticscholar.org/CorpusID:8732338

Constum T, Tranouez P, Paquet T (2024) Daniel: A fast document attention network for information extraction and labelling of handwritten documents. URL https://arxiv.org/abs/2407.09103, arXiv:2407.09103

Crosilla G, Klic L, Colavizza G (2025) Benchmarking large language models for handwritten text recognition. URL https://arxiv.org/abs/2503.15195, arXiv:2503.15195

Grosicki E, Carré M, Brodin JM, et al (2009) Results of the rimes evaluation campaign for handwritten mail processing. pp 941–945, https://doi.org/10.1109/ICDAR.2009.224

Gutteridge B, Jackson MT, Kukurin T, et al (2025) Judge a book by its cover: Investigating multi-modal llms for multi-page handwritten document transcription. URL https://arxiv.org/abs/2502.20295, arXiv:2502.20295

Hamdan M, Rahiche A, Cheriet M (2024) Htr-jand: Handwritten text recognition with joint attention network and knowledge distillation. URL https://arxiv.org/abs/2412.18524, arXiv:2412.18524

Heyberger L, Frossard G, Al-Kendi W (2024) Belpop, a history-computer project to study the population of a town during early industrialization. In: Digital History Basel 2024. Zenodo, Bâle (Suisse), Switzerland, https://doi.org/10.5281/zenodo.13904600, URL https://hal.science/hal-05084484

Hodel T, Schoch D, Schneider C, et al (2021) General models for handwritten text recognition: Feasibility and state-of-the art. german kurrent as an example. Journal of Open Humanities Data https://doi.org/10.5334/johd.46

Hunt MJ (1990) Figures of merit for assessing connected-word recognisers. Speech Commun 9:329–336. URL https://api.semanticscholar.org/CorpusID:205065229

Kim S, Baudru J, Ryckbosch W, et al (2025) Early evidence of how llms outperform traditional systems on ocr/htr tasks for historical records. https://doi.org/10.48550/arXiv.2501.11623

Levenshtein VI (1966) Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady 10:707

Li Y, Chen D, Tang T, et al (2025) Htr-vt: Handwritten text recognition with vision transformer. Pattern Recognition 158:110967. https://doi.org/10.1016/j.patcog.2024.110967, URL http://dx.doi.org/10.1016/j.patcog.2024.110967

Marti UV, Bunke H (2002) The iam-database: An english sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition 5:39–46. https://doi.org/10.1007/s100320200071

Mayr M, Neumeier K, Krenz J, et al (2025) Data-efficient handwritten text recognition of diplomatic historical text. Multimedia Tools and Applications pp 1–16. https://doi.org/10.1007/s11042-024-20545-9

Nunes G, Rolla V, Pereira D, et al (2025) Benchmarking table extraction: Multimodal LLMs vs traditional OCR. In: Fei H, Tu K, Zhang Y, et al (eds) Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025). Association for Computational Linguistics, Vienna, Austria, pp 8–15, https://doi.org/10.18653/v1/2025.xllm-1.2, URL https://aclanthology.org/2025.xllm-1.2/

Penarrubia C, Valero-Mas JJ, Calvo-Zaragoza J (2025) Self-supervised learning for text recognition: A critical survey. International Journal of Computer Vision https://doi.org/10.1007/s11263-025-02487-3, URL http://dx.doi.org/10.1007/s11263-025-02487-3

Prasad A, Déjean H, Meunier JL, et al (2018) Bench-marking information extraction in semi-structured historical handwritten records. URL https://arxiv.org/abs/1807.06270, arXiv:1807.06270

Romero V, Fornés A, Serrano N, et al (2013) The esposalles database: An ancient marriage license corpus for off-line handwriting recognition. Pattern Recognition 46. https://doi.org/10.1016/j.patcog.2012.11.024

Sohail MA, Masood S, Iqbal H (2024) Deciphering the underserved: Benchmarking llm ocr for low-resource scripts. URL https://arxiv.org/abs/2412.16119, arXiv:2412.16119

Torres Aguilar S, Jolivet V (2023) Handwritten Text Recognition for Documentary Medieval Manuscripts. Journal of Data Mining and Digital Humanities Historical Documents and automatic text recognition. https://doi.org/10.46298/jdmdh.10484, URL https://hal.science/hal-03892163