



**HAL**  
open science

# Handwritten Text Recognition for Historical Documents using Visual Language Models and GANs

Sergio Torres Aguilar

► **To cite this version:**

Sergio Torres Aguilar. Handwritten Text Recognition for Historical Documents using Visual Language Models and GANs. 2024. hal-04716654v2

**HAL Id: hal-04716654**

**<https://hal.science/hal-04716654v2>**

Preprint submitted on 17 Oct 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Handwritten Text Recognition for Historical Documents using Visual Language Models and GANs

Sergio Torres Aguilar<sup>1</sup>

<sup>1</sup>Université du Luxembourg, Institut for History, Luxembourg

Corresponding author: Sergio Torres , [sergio.torres@uni.lu](mailto:sergio.torres@uni.lu)

## Abstract

In this study, we focus on Handwriting Text Recognition (HTR) on Medieval and Early Modern documentary manuscripts (10th-16th centuries) using Vision Language models (VLM). We leverage the TrOCR architecture and integrate domain-specific large language models (LLM). This HTR approach show promising results on contemporary documents but the application on historical manuscripts and low-resource languages need domain pre-trained Image models to encode sequential data and adapted LLM's to adequately decode the signals. Furthermore, as training pairs from medieval manuscripts are scarce a synthetic dataset generated using GAN (Generative Adversarial Networks) augmentation techniques will be used during training. For this work, the annotated training dataset comprises more of 2 million tokens and 210,000 graphical text-lines coming from 52 different manuscripts in mostly four ancient languages versions for Latin, French, Spanish and High German. The synthetic GAN dataset comprises 420k graphical lines emulating textual and graphical features from the ground-truth. The results shows relative improvements until 30% in CER, WER and BERT-score compared to CRNN only solutions. This study outlines the following: the training architecture and corpora employed; delves into the encountered challenges during training and validation concerning ancient writing practices, and conducts an analysis of the potential biases and strengths associated with the joint application of vision transformers, GAN's and LLMs for HTR tasks.

## Keywords

HTR for ancient documents, HTR for medieval manuscripts, HTR using visual Language Models, digital diplomatics, medieval charters

## I INTRODUCTION

This study introduces a cutting-edge method for Handwriting Text Recognition (HTR) integrating Large Language Models (LLMs) and Vision Transformer (ViT) for documentary manuscripts spanning different languages and script families from the late-10th to the 16th century. By proposing an encoder-decoder architecture that leverage the capabilities from Image transformers and Language Transformer models in an autoregressive sequence, we depart from traditional HTR methodologies, which are based on CRNN systems, towards a more integrated approach for the accurate transcription and interpretation of historical documentary corpora.

Documentary manuscripts comprises a variety of documents generated by legal, juridical, and administrative practices. These include charters, registers, legal series, reports, and proceedings, among others. They serve as a primary written source for historical studies, particularly from the late Middle Ages when royal, state, and urban administrations began to adopt a more systematic approach to record-keeping. Despite their importance, these sources have often been neglected in digital practices and modeling. Their increased volume, which became massive from the late 14th century, along with their complex page layouts, domain-specific discourse, and cursive

writing, pose a significant challenge for digitization campaigns and digital edition projects. Besides, 19th and 20th-century editions of this kind of documents are scarce and mostly focused on these imitating documentary practices, such as cartularies, or on organic corpora like royal and abbey charters collections, resulting in a lack of aligned corpora for HTR training.

On the other hand, the efficacy of using explicit language models to improve the accuracy of HTR has been demonstrated in previous studies, yielding promising results (Tarride et al. [2024], Boros et al. [2024]). These models capture statistical correlations among tokens in natural language, providing context awareness and proposing more contextually appropriate interpretations. This is particularly crucial for addressing challenges such as complex layouts, ancient or noisy writing, or out-of-domain issues, where a purely graphical interpretation may fall short in providing an accurate transcription, as it lacks the linguistic nuances and contextual understanding that language models bring to the table. In contrast to classical CRNN models, which implicitly learn language features and focus on identifying patterns directly from the data without considering linguistic context, the integration of a language model involves an statistical analyze of how words interrelate in a language and the likelihood of certain sequences co-occurring. This approach results in more precise and contextually informed predictions.

Additionally, language models like RoBERTa (Liu et al. [2019]), acting here as tokenizer and decoder, or mT5 (Xue et al. [2021]) employ strategies such as label smoothing, allowing the model to assign a small but nonzero probability to unseen tokens. By smoothing the probability distribution over the output vocabulary, this strategy penalize over-confident outputs and encourages the model to explore a wider range of possibilities during training, leading to more reliable and realistic predictions, especially when dealing with complex or ambiguous inputs (Gao et al. [2020]). This approach helps the model generalize better and mitigates over-fitting, a critical problem in low-resource scenarios where the model hasn't encountered a comprehensive range of token combinations during training.

Concerning the use of synthetic datasets, recent studies (Rahal et al. [2024], Vidal-Gorène et al. [2023]) have explored the potential of introducing synthetic material to address the scarcity of ancient writing datasets. This includes adding graphical lines based on texts that are only edited or not digitized thereby expanding the available data for training models. The challenge in this case lies in synthesizing handwritten text images by successfully transferring writing family features and encoding individual calligraphic textures into a latent space. This is a natural progression of data augmentation techniques in computer vision, aiming to expose models to a wider range of data scenarios. Synthetics increase diversity with minimal bias increase. Literature shows that, in most cases, synthetic complex content helps models to generalize better in real-data scenarios by introducing new patterns and edge cases not found in the original data (Shorten and Khoshgoftaar [2019]).

Our extensive annotated ground-truth corpus comprises more of 2 million tokens and 210,000 text lines across 52 manuscripts in historical Latin, French, Spanish and High German, with a significant portion annotated for Named Entity Recognition (NER) tasks. Our transformer-based strategy outperforms traditional models in HTR by improving character and word error rates and BERT-score metrics, while efficiently conducting GAN handwriting lines generation and internal post-correction tasks based on LLM replacements. This approach not only reduces computational complexity but also minimizes the dependency on large volumes of ground-truth data for model fine-tuning, making the process more efficient and scalable.

We address several training challenges, including the development of domain-specific language

models; the implications of causal word and sub-word prediction as well as the biases from the Attention mechanism. Furthermore, we tackle the difficulty of transcribing texts with significant graphical and orthographic variation using a formal and fixed vocabulary. Our hybrid training strategy, which combines supervised learning on annotated data with unsupervised learning on larger, unannotated datasets, improves the model’s generalizability in a cross-domain and cross-chronology environment. This method enables effective handling of continuous script and discrete entity recognition, significantly broadening the model’s applicability to a diverse array of historical documents.

This work provides the following contributions:

1. We extensively test transformed-based HTR models which outperforms CRNN solutions and open the way to easily apply diverse LLM-based post-correction and NLP tools.
2. We develop new LLM and GAN models on medieval and modern sources to perform a comparative study on various decoding and data augmentations methods, exploring the potential of synthetic data to improve model performance.
3. We publish in open-source our models in our Huggingface hub and Zenodo:  
<https://huggingface.co/magisttermilitum>  
<https://doi.org/10.5281/zenodo.13862096>

## II RELATED WORKS

In the last decade, Handwriting Text Recognition (HTR) technology for manuscript analysis have transformed how historical texts and digitized collections are automatically processed and shared. The approaches has transitioning from Markov engine-based algorithms (Fischer et al. [2011]) to deep learning methodologies that leverage character-level neural networks. The introduction of Connectionist Temporal Classification (CTC) and bidirectional neural networks has substantially enhanced models’ precision while diminishing the dependency on big ground-truth datasets for produce robust HTR models. Both technologies ensure that models can accurately align and transcribe sequences without requiring pre-segmented input data. Furthermore, the application of transformers based on the Attention Mechanism to computer vision tasks (Dosovitskiy et al. [2020]) has provided a significant leap forward, enabling models to focus on global interactions and relevant features of the text more effectively. These advancements have enabled more efficient and accurate transcription of historical documents, making them accessible for scholarly research and public engagement.

Transfer learning practices are increasingly prevalent in HTR research. This trend has been driven by the development in recent years of both, robust CRNN models, which have been pre-trained on vast line-transcribed datasets; and large image-to-text transformer models which are typically built upon self-supervised Vision Transformer models. These pre-trained foundations are one of the main factor for the adoption of text recognition practices on patrimonial institutions as they improve accuracy and efficiency with minimal setups and ground-truth. By leveraging pre-trained models, projects can avoid the zero-start effect and reduce the technological entry barrier, thereby leading to a more efficient allocation of funds and human resources and facilitating interdisciplinary research and collaboration, opening new avenues for the study of cultural heritage.

As demonstred by (Ströbel et al. [2022], Arias et al. [2023]), the fine-tuning of Transformed-based HTR such as TrOCR using ancient writing sources can lead to models displaying Character Error Rate (CER) accuracy rates reaching 95% in validation, though it’s noted that model

performance, similar to the case of CRNN-based models, can decrease during cross-domain or cross-family script tests. Additionally, the data augmentation techniques based on synthetic material have been explored by TrOCR developers on digitized modern material and in academia using medieval sources with promising results (Rahal et al. [2024]). Furthermore, (Tarride et al. [2023]) demonstrates that NLP tasks as NER (Named entity Recognition) using text-based encoding can be successfully integrated during HTR training. Other studies indicate that while LLM models can enhance the quality of inferences in seq2seq and classification task (de Sousa Neto et al. [2020]), their use in purely causal mode, specially with pre-trained autoregressive models, can degrade performance. (Boros et al. [2024]).

The field of HTR for ancient writing is characterized by a symbiotic blend of historical scholarship and technological innovation. Generating ground-truth data for ancient scripts is a complex process, necessitating deep expertise to decipher ancient languages and scripts, and to develop methodologies that resolve ambiguities and establish precise annotation guidelines. Besides, HTR systems face challenges largely studied by paleography and diplomatics: complex variability in graphical behaviour, document typology, script families, and regional writing practices. In that sense, project initiatives coming from humanistic centers like Himanis (Stutzmann et al. [2017]), Home (Stutzmann et al. [2021b]), Cremma (Clérice et al. [2023]), Catmus (Clérice et al. [2024]) and Bullinger Digital (Ströbel et al. [2022]), backed by Transkribus (Kahle et al. [2017]), Kraken (Kiessling [2019]) and the eScriptorium (Kiessling et al. [2019]) platform, have been instrumental in compiling, analyzing and documenting corpora for ancient texts as well as establishing widely accepted scientific practices and annotation standards (Guéville and Wisley [2024]). This collaborative effort have been essential for refining model performances, through both the rigorous analysis of prediction behaviours and the careful curation of ground-truth data.

### III CORPORA DESCRIPTION

For this work several freely available datasets were used covering a wide range of epochs, scripts families, provenances and typologies.

**Table 1** List of training and testing corpora.

Corpus / Manuscript	Dates	Language	Typologie	Script family	n° lines	n° tokens
<b>TRAIN</b>						
e-NDP project	14-15th	la	Chapter registers	Cursive	33 800	203 565
Alcar-HOME	12-15th	la, fro	Cartulaires (charters)	Textualis, Cursive, Hybrida	103 412	737 635
Himanis	14-15th	la, fro	Royal registers	Cursive; Hybrida	15 600	316 155
Königsfelden Abbey	14-16th	la, gmh	Charters and records	Textualis	12 823	274 615
Bullinger Digital	15-16th	la, deu	Private correspondence	Cursive	10 000	101 493
MLH	13-15th	la, fro, gmh	Cartularies, registers	Textualis, Cursive	22 433	282 897
CODEA	14-16th	esp, la	Charters, legal records	Cursive, Humanistic-Cursive	12 150	128 551
					<b>210 218</b>	<b>2 044 911</b>
<b>TEST</b>						
Cod. Sangallensis, 562	9th	la	Hagiographical book	Carolingial Minuscule	1 410	11 597
Faithfull Transcription	14-15th	la, deu, gmh	Sermons, Hours books	Bastarda, Textualis, Cursive	1 000	6 380
Wien ÖNB Cod. 2160	9th	la	Legal texts, decrets	Carolingian Minuscule	1 949	8 203
Troyes, Méd. Ms 1600	14th	la	Exegetical, dogmatic	Semi-Hybrida	1 622	7 939
Cologne, Bodmer, 168	14th	fro	Litterature	Textualis	1 977	13 016
ANLux, A-X-42-1	14-15th	la, fro	Charters, comptability	Textualis, Cursive	3 147	36 811
					<b>11 105</b>	<b>83 946</b>

## 3.1 Training Corpora

### 3.1.1 *The e-NDP project corpus*

The Notre-Dame-de-Paris registers corpus is a unique resource for HTR of late medieval documentary manuscripts, featuring decisions from weekly cathedral canon meetings from 1326 to 1504 (Claustre and Smith [2022]). The project has released 523 pages transcribed from 26 registers by historians and paleographers<sup>1</sup>. These registers, mainly in Latin with some medieval French, use a cursive script and show an evolving layout typical for administrative documents, including lists, margin notes, and titles. Originally intended for daily use rather than long-term preservation, their design is less meticulous than that of literary manuscripts.

### 3.1.2 *The HOME-Alcar corpus*

The HOME-Alcar corpus (Stutzmann et al. [2021b]), comprises images of medieval manuscripts with line-level aligned scholarly editions and detailed named entity annotations<sup>2</sup>. This bilingual (Latin and French) collection, featuring 17 French cartularies from the 12th to the 14th centuries across four script families, serves as a crucial tool for training Handwritten Text Recognition (HTR) and Named Entity Recognition (NER) models. Cartularies, essential for medieval studies, include documents that were often not preserved in their original form, such as property transfers, wills, land and debt disputes, treaties and indemnities.

### 3.1.3 *The Himanis project*

The Himanis project features documents from the French Royal Chancery, specifically registers JJ35 to JJ211 at the Archives nationales, dating from 1302 to 1483<sup>3</sup>. These registers, containing a variety of charters such as letters of remission, mandates, and ennoblements, were meticulously digitized and matched line by line with Paul Guérin’s semi-diplomatic edition (Guérin [1881]). The training dataset released in 2021 (Stutzmann et al. [2021a]), includes 1,500 images and 30,000 text lines, primarily written in Latin and Old French using the Cursiva script.

### 3.1.4 *The Königsfelden Abbey corpus*

The digital edition ”Charters and Records of Königsfelden 1308–1662”<sup>4</sup>, carried out by the University of Zurich’s Department of History from 2017 to 2020, provides comprehensive access to 1557 charters and records from the 14th to the 17th centuries. The project distinguishes itself by offering a fully annotated corpus in a TEI format (Halter-Pernet et al. [2021]) that includes digital facsimiles, transcriptions, and texts with diplomatic and named entity annotations.

### 3.1.5 *The Bullinger Digital*

The bullinger Digital<sup>5</sup> project aims to digitize and make accessible the correspondence of Heinrich Bullinger (1504-1575), a key figure in the Reformation. The project involves encoding the letters in XML and providing online access to around 10,000 letters received and 2,000 letters sent by Bullinger. These letters, primarily in Latin and Early New High German, offer valuable insights into the political, theological, and social contexts of the Early Modern period. In the last years the bullinger have become a key corpus for German and Neolatin Early-modern HTR models. A multilingual section of this vast corpus (10k lines) was randomly selected for training.

---

<sup>1</sup><https://doi.org/10.5281/zenodo.7575693>

<sup>2</sup><https://doi.org/10.5281/zenodo.5600884>

<sup>3</sup><https://doi.org/10.5281/zenodo.5535306>

<sup>4</sup><https://doi.org/10.5281/zenodo.5179361>

<sup>5</sup><https://github.com/bullinger-digital/bullinger-korpus-tei>

### 3.1.6 *The Monumenta Historica Luxemburgensica*

This is an ongoing project hosted by the University of Luxembourg which comprises several medieval manuscripts produced in Luxembourg institutions between the 12th and the 15th centuries<sup>6</sup>. It entails the transcription and digital edition of cartulaires, feudal books and registers digitized by the National Archives or hosted in The Great Regions repositories. The documents span a three languages register: french, German and Latin.

### 3.1.7 *The CODEA corpus*

The CODEA corpus (*Corpus of Spanish Documents Prior to 1800*), launched in 2012 by the University of Alcalá (Sánchez-Prieto Borja [2012]), offers a wide landscape of the evolution of Spanish from the High Middle Ages<sup>7</sup>. The corpus includes a broad typological range of documents, including both public records—such as privileges, mandates and grants—and private documents like contracts, sales and letters coming from chancelleries, city offices, notaries, and small scriptoria from the 11th to the 16th centuries. Currently, the CODEA corpus comprises 2,500 charters. For research purposes, a subset of 250 documents has been randomly selected and manually aligned to facilitate the study of HTR on Old Spanish documentary sources.

## 3.2 Testing Corpora

In order to offer a large scope of the capabilities of the models we choose six curated manuscripts unseen during training. From this we aim to propose a complete landscape of medieval and early-modern production varying genres, languages and chronologies:

1. The **Saint-Gall** database (late 9th century)<sup>8</sup>, which is a binarized version of the Codex Sangallensis 562, the oldest copy of the vitae of St. Gallus and St. Otmar in the version of Walahfrid Strabo. The Saint-Gall is one of the most used corpus in HTR benchmarks for ancient documents.
2. The **Cremma** medieval dataset<sup>9</sup>. This is a modified version of the original Cremma corpus which use a graphematic transcription, to include the expanded abbreviations of some books (13th-15th) as the Chanson d’Otinél (Cologne, Bodmer, 168, 211ra-222rb).
3. **Faithful** Transcriptions: Published in 2021<sup>10</sup>: The curated version from the Transcribathon (Staatsbibliothek zu Berlin) of 181 pages coming from 12 medieval manuscripts (14th-15th) in German (Middle High variants and Low German), Dutch, and Latin, featuring various scripts such as Textualis, Gothic Cursive, and Bastarda. These manuscripts includes historical Bibles, prayer books, and sermon collections.
4. The **Liber Feudorum** (ANLux, A-X-42-01<sup>11</sup>) which is a feudal book (14th-15th centuries) containing copies of feudal charters, county incomes and list of revenues, included in the Monumenta Luxemburgensica and written in Latin, French and High German.
5. The Wien ÖNB Cod. 2160 (Attwood et al. [2023])<sup>12</sup>. This is the transcribed version of Vienna, Österreichische Nationalbibliothek, 2160 (9th century) containing a copied version of a well-known legal compilation, the *Mosaicarum et Romanarum Legum Collatio*.

---

<sup>6</sup>[www.tridis.me](http://www.tridis.me)

<sup>7</sup><https://corpuscodea.es/>

<sup>8</sup><https://fki.tic.heia-fr.ch/databases/saint-gall-database>

<sup>9</sup><https://doi.org/10.5281/zenodo.7506657>

<sup>10</sup><https://doi.org/10.5281/zenodo.5582483>

<sup>11</sup><https://query.an.etat.lu/Query/detail.aspx?ID=212411>

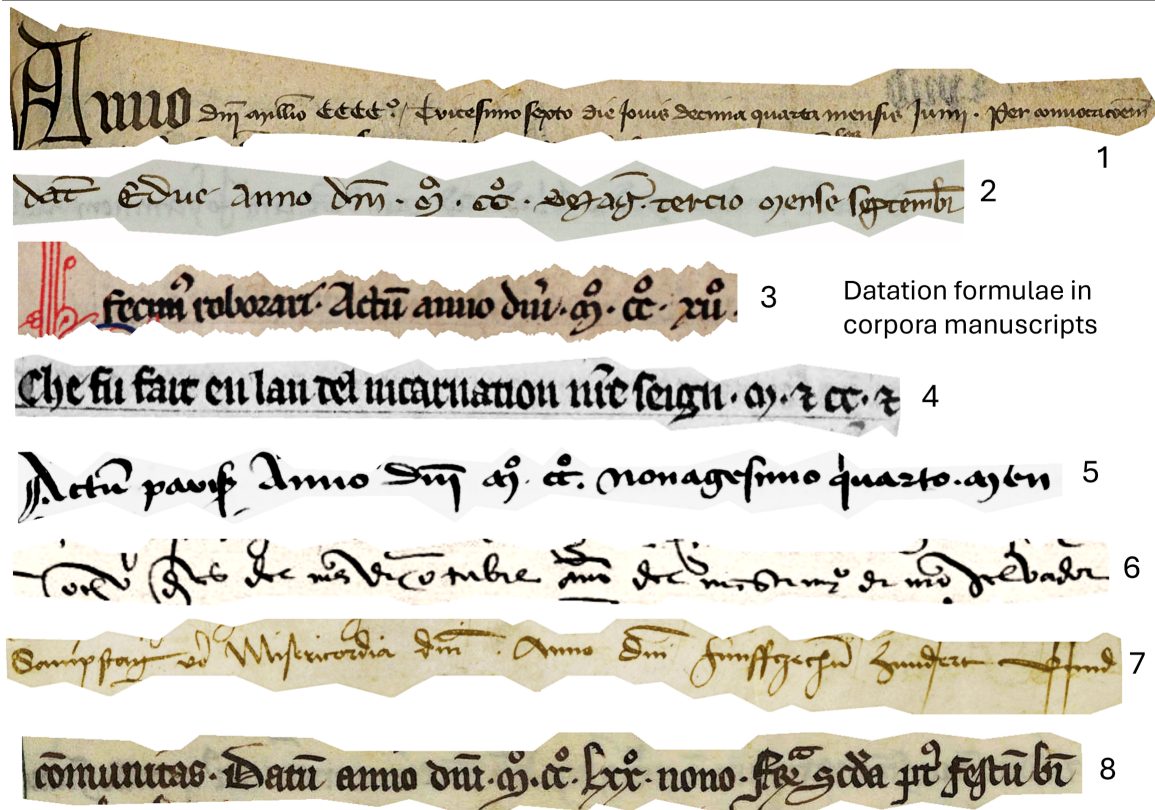
<sup>12</sup><https://doi.org/10.5281/zenodo.7537204>

6. The **Summa of abstinencia** or *Dictionarius pauperum* attributed to Nicolas de Byard (Médiathèque de Troyes Champagne Métropole, Ms 1600<sup>13</sup>). This a highly abbreviated copy (14th century) from a famous collection written at the end of 13th century of preaching material and advice on behaviour intended for help clerics during his work.

**Figure 1** Eight examples of act datation from training corpora manuscripts:

1. e-NDP : Anno domini millesimo CCCC° tricesimo sexto die jovis decima quarta mensis junii.
2. Nesle : datum Edue anno Domini M° CC° sexagesimo tercio mense septembri .
3. Molesmes : fecimus roborari. Actum anno Domini M°. CC°. duodecimo, mense junio.
4. Fervaques : Che fu fait en l’an de l’incarnation nostre seigneur M et C
5. Clairmarais : Actum paris Anno domini M° . CC° . nonagesimo quarto .
6. CODEA : ocho dias del mes de otubre anno del nascimiento de nuestro salvador
7. Königsfelden : sampstag vor misericordia domini , anno domini funfftzechen hundert unnd
8. MLH : communitas. Datum anno Domini M°. CC°. LXX°. nono, feria secunda post festum beati

In Latin, French and German, all eight follow a similar datation pattern based on the year of the incarnation using roman numerals.



### 3.3 Datasets configuration

The training corpus comprises five primary writing families, over a hundred distinct hands (the precise number of interventions remains indeterminable), and four main languages (refer to Table 1). As proved by other works in ancient handwriting (Torres Aguilar and Jolivet [2023]) blending families, hands and languages during training is a recommended strategy. This is not only because it appears to yield robust models by limiting the over-fitting on some hands clusters, but also because ancient manuscripts, especially those from the medieval and early modern periods, can display more than one writing family or include several hand interventions

<sup>13</sup><https://gallica.bnf.fr/ark:/12148/btv1b10510290t>



within a single document or even a single page. Moreover, multilingual approaches, particularly those employing LLMs, seem to provide a more extensive and robust linguistic foundation for models designed to generalize across a wide range of sources spanning several centuries.

**Figure 2** The complete set of 123 characters used to transcribe the training corpus.

Lowercase letters	a b c d e f g h I j k l m n o p q r s t u v w x y z
Capital letters	A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
Numbers	0 1 2 3 4 5 6 7 8 9
Punctuation marks	: . , ; : ! ? ' " " « » - ~   blank
Diacritical marks	à â ä Ä Á æ É È è é ê ë ç Ç ñ ï ö ó œ ù ü ÿ ñ ß
Another glyph	# % \$ * + - § ° □ ○ } / > < † ¶ ...

Both the training and test corpora are multilingual, but with an unbalanced distribution. Latin, French, and Spanish are predominantly represented in an overall ratio of 5:2:1 for both the training and testing sets. Pages written in High German variants can also be found, albeit in a more marginal quantity (less than 5%). Moreover, as the corpora are derived from diplomatic editions, the transcriptions adhere to the semi-diplomatic standard with normalization, but not regularization (manuscript spellings are kept, vg. *echevin*, *eskevin*, *eschevin*). This implies that all abbreviations (by suspension and contraction) and abbreviating symbols have been expanded, punctuation has been standardized, named entities have been capitalized, and variants of a given letter (allographs, for instance, for “s” long or “a” round) have been mostly reduced to the canonical letter, with no distinctions made between them.

In addition, as modern editors have introduced dots and commas (or semicolons) to indicate pauses in the sentences, even though such punctuation marks were not always present in the original manuscript, the accuracy results will be presented (Table 6) considering both a fully edited transcription (hereafter referred to as “raw”) and a transcription that omits punctuation marks and diacritics (hereafter referred to as “cleaned”). This will facilitate the visualization of the impact of modern editorial interventions on the accuracy of the model’s predictions.

The choice of transcription guidelines for any project is influenced by their intended application. In libraries, archives and most online editions, transcription criteria that normalize the text can enhance searchability, content parsing, and processing through NLP and NLU tools for the general reader. However, for researchers in paleography or philology, more specific criteria that follow a transliteration or graphemic / graphetic transcription schema may be preferred (Robinson and Solopova [1993], Guéville and Wrisley [2024]). This approach preserves as much graphical information as possible from the original manuscript, including special letter forms, abbreviations, allographs, and marks through an extended Unicode standard. While this last transcription schema facilitates HTR training by reducing the number of linguistic layers that models must learn, it faces challenges on our approach. Firstly, LLMs and explicit vocabularies which are typically trained on millions of normalized documents could fall short applied on these specialized texts. Secondly, creating diplomatic critical editions can be labor-intensive as they are seldom seen in academia publications for practical reasons. This scarcity may complicate the production of diverse ground-truth. As a countermeasure, recent datasets have emerged that allow for multiple transcription levels linked in TEI collating schemes as well as projects to catalog plural HTR datasets (Chagué and Clérice [2022]).

Our perspective in this article is to propose generalistic models from legacy diplomatic editions of documentary and book manuscripts, mostly from the late medieval and early modern periods (12th-16th centuries) on the base of semi-diplomatic transcriptions. These models are designed to provide a standardized and consistent text directly compatible with modern text processing tools, facilitating semantic annotation and serving as a foundation for developing new documentary editions or incorporating diplomatic layers into critical editions.

## IV HTR ARCHITECTURE

### 4.1 The CRNN model

The CRNN architecture employed as a [baseline](#) in this study has 6.3 million parameters. It consists of: (i) four CNN layers for extracting images local features; (ii) three bidirectional LSTMs layers for processing temporal sequences; and (iii) a CTC algorithm for loss calculation and text rendering as it allows the model to handle sequences of varying lengths and align the predicted characters with the ground-truth. The convolutional block uses layers with different kernel sizes (4x16, 3x8) and 16n filters per layer (32, 32, 64, 64). Each layer is followed by MaxPooling and a 2D dropout with 0.1 probability. The activation function is ReLU. A reshape layer collapses non-1 height dimensions into a single value, accommodating different manuscript line sizes with a fixed image height of 140 pixels. This is an architecture designed to capture both local and temporal features, which makes it highly effective for sequence tasks such as HTR where adaptability to different text styles, layouts and spacing is crucial.

In [Kraken](#), which uses the VGSL network specification, this train architecture can be fully replicated by using : `-s '[1,140,0,1 Cr4,16,32 Do0.1,2 Mp2,2 Cr4,16,32 Do0.1,2 Mp2,2 Cr3,8,64 Do0.1,2 Mp2,2 Cr3,8,64 Do0.1,2 S1(1x0)1,3 Lbx256 Do0.3,2 Lbx256 Do0.3,2 Lbx256 Do0.3]'`

### 4.2 The Transformer model

We adopt the encoder-decoder architecture proposed in the TrOCR paper (Li et al. [2023]), which includes an image Transformer for feature extraction and a text Transformer for language modeling. The decoder is initialized with the Deit model trained on ImageNet (Touvron et al. [2021]), while the encoder utilizes the English RoBERTa model (Liu et al. [2019]). The original model ([microsoft/trocr-large-handwritten](#)) was pre-training over hundred of millions synthetic text lines images, providing a robust foundation for recognizing diverse text patterns.

The Encoder breaks down input images into patches, which are then flattened and projected into D-dimensional vectors. The “[CLS]” token aggregates patch information, and positional embeddings are added. Unlike CNNs that process images holistically, Transformers handle images as sequences of patches. The Decoder generates the wordpiece sequence, considering both the encoder output and its previous iteration. These hidden states are projected to the vocabulary size, with probabilities computed via softmax. The final output generation employs beam search (Von Platen [2020]), selecting the most probable sequence from multiple candidates.

This architecture’s flexibility allows for experimentation with various language models as tokenizers or decoders. For instance, while the TrOCR paper uses an English RoBERTa model as tokenizer, it’s entirely feasible to use other models fine-tuned for specific tasks. In our case, we could conduct an experiment using a RoBERTa model trained on historical texts (RoBERTa-med) and a GPT-2 model (GPT-2 med) fine-tuned on the same material. These experiments aim to verify the impact of domain-specific language models on the system’s performance and guide us in selecting the most suitable model for general and specific tasks.

**Table 2** List of corpora used to train LLM models. As the major corpora were crawled from the internet, some GB of data are overlapping. The cleaned ensemble has 5GB, for about 750 million tokens.

Corpus	Size	Dates	Languages	Typology
Corpus Corporum	3,2Gb	5th BC - 16th	la	Charters and Litterature
CC100	3,0Gb	5th BC - 18th	la	Articles and books
Wiki-latin	1,2Gb	5th BC - 20th	la	Articles and books
CEMA	320Mb	9th - 15th	la, fro	Charters
HOME + e-NDP	38Mb	10th - 15th	la, fro	Charters and Registers
CODEA	13Mb	10th - 17th	la, esp	Charters and Registers
BFM	34Mb	13th - 15th	fro	Litterature
NCA	19Mb	13th - 15th	fro	Litterature
OpenMedFr	5Mb	12th-15th	fro	Litterature

~7,8Gb

750M tokens (5Gb)

### 4.3 The LLM model

The [Roberta-medieval](#) model used during experiments was training from scratch (10 epochs) using a base architecture (768 dimensions, 12 attention heads, 12 hidden layers) on a 8k vocabulary. The training corpora, also used on GPT-2, includes several freely available medieval and classical Latin, Old Frech and Old Spanish datasets (See table 2) spanning from the classical Latin period (5th BC) to the 20th century: The Corpus Corporum<sup>14</sup>, Common-crawl 100<sup>15</sup>, Wikipedia in Latin<sup>16</sup>, CEMA (Cartae Europae Medii Aevi)<sup>17</sup>, HOME<sup>18</sup>, e-NDP<sup>19</sup>, Base de Français Médiéval<sup>20</sup>, Nouveau Corpus d’Amsterdam<sup>21</sup>, Open Medieval French<sup>22</sup>.

In this work the decoding using the LLM model was done using a beam search (value 3), which is an strategy that maintains a set of the most promising sequences at each step of the decoding process. This approach allows the model to explore multiple possible sequences simultaneously, and not just choosing the most probably option, effectively balancing between exploration and exploitation in the search space.

### 4.4 The GAN model

Our experiments utilized synthetic lines generated using the HiGAN+ architecture (Gan et al. [2022]), a state-of-the-art Generative Adversarial Network (GAN) designed for synthesizing realistic handwritten text. Unlike conventional methods, it can mimic calligraphic styles, generate variable-sized images, and handle arbitrary textual contents, including out-of-vocabulary words. It disentangles textual contents and calligraphic styles using a writer-specific auxiliary loss and contextual loss, which enable precise one-shot handwriting style transfer. The architecture comprises five concurrent models:

1. Style-Controlled Generator: Generates variable-length images based on arbitrary textual content.

<sup>14</sup><https://mlat.uzh.ch/>

<sup>15</sup><https://data.statmt.org/cc-100/>

<sup>16</sup><https://dumps.wikimedia.org/backup-index.html>

<sup>17</sup><https://cema.lamop.fr/>

<sup>18</sup><https://doi.org/10.5281/zenodo.5600884>

<sup>19</sup><https://zenodo.org/record/7575693>

<sup>20</sup><https://nakala.fr/collection/10.34847/nkl.1279lie9>

<sup>21</sup><http://srcmf.org/>

<sup>22</sup><https://github.com/OpenMedFr>

2. Global Discriminator: Verifies the fidelity of synthetic images.
3. Patch Discriminator: Refines local texture details of synthetic image by verifying patch fidelity.
4. Style Encoder: Disentangles calligraphic styles from arbitrary handwriting images.
5. Text Recognizer: Guides the generator to produce readable handwriting images conditioned on arbitrary textual content.

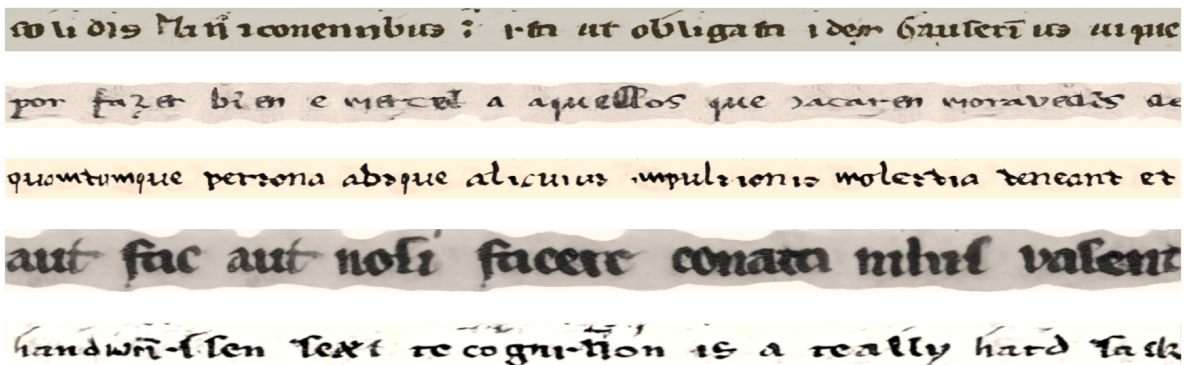
For our experiment, we adapted the system to handle color images, lines up to 30 characters, 4 pixels patches, and generate images with variable width and a fixed height of 64 pixels.

The GAN model was trained on a subset of our HTR training corpus. We select all lines that contained until 12 tokens. The number of writers was defined to 372, and the writer styles interpolation was enabled to generate a variety of new styles. The texts for the new graphical lines were extracted from the charters and literature collections mentioned in Section 4.3, for which we have no annotated graphical lines. We generated a total of 420k lines (2x our training GT), each varying in size from 8 to 12 tokens. Examples of some synthetic lines are illustrated in Figure 3.

---

**Figure 3** Five synthetic generated lines imitating hand-based Cursiva and Textualis styles on real and fictional texts:

1. Latin: solidis Matisconensibus : sita ut obligati idem Gauterius usque
2. Spanish : por fazer bien e merçed a aquellos que sacaron moravedies de
3. Latin : Quomtumque persona absque alicuius impulsione molestia teneant et
4. Neo-Latin : Aut fac aut noli facere conata nihil valent (Yoda dixit)
5. English : handwritten Text recognition is a really hard task



## 4.5 Hyperparameters

### 4.5.1 CRNN (Kraken)

The baseline CRNN model follows a classical 3CNN + 3RNN + CTC architecture. We feature a pad size of {24} pixels, useful in providing more space to the kernel for better coverage of the image. Additionally, we ran a {ReduceOnPlateau} optimizer with patience {3} which can help prevent overshooting the minimum of the loss function on constant LR models.

### 4.5.2 Transformers (TrOCR)

When RoBERTa medieval is used as decoder some modifications were introduced before reuse the TrOCR-large weights, as they were exclusively pre-trained on English data:

1. Re-initialization of the embedding layer and the fully connected layer in the decoder in order to allow these layers to better learn representations specific to ancient languages.

2. Re-starting the positional encoding layer in the decoder. This can be beneficial to better understand the order of tokens in sequences which are variable for each language.

#### 4.5.3 GAN (HiGAN+)

The GAN model follows a custom optimization: The regularization term  $\lambda_{kl}$ , is set to  $1 \times 10^{-4}$ , controlling how closely the encoder’s latent distribution fits the prior. The context loss,  $\lambda_{ctx}$  is set to  $\{5.0\}$ , weighting the importance of preserving textual content in the total loss function. The remaining hyper-parameters,  $\lambda_{ls}$  are dynamically adjusted during training using a gradient balancing strategy.

#### 4.5.4 LLM (RoBERTA-med)

As was observed by (Grobol et al. [2022]), for smaller datasets (under 1 Billion tokens), employing a restricted vocabulary can potentially enhance the performance on the mask tokens task as it helps to mitigate the risk of the model erroneously selecting rare words or uncommon phrases as output. In our case the vocabulary sizes ranges from 10k to 8k tokens, which is optimized to balance the model’s performance and the complexity of the language data.

All the LLM, GAN and HTR training cycles were realized on a Threadripper PRO (24-cores) using a dual A100 (32GB vRAM) coupled to 128GB of RAM. In all cases, data augmentation techniques, such as distortion, blur, rotation and blots were also applied to the images to enhance training performance.

**Table 3** Summary of Model Training Hyper-Parameters and Results. CER (Character Error Rate), WER (Word Error Rate), FID (Fréchet inception distance), KID (Kernel Inception Distance ), PP (Perplexity). Lower is better in all results.

Model	Batch	Learning Rate	Warmup	Epochs	Emb_Dim	Other	Results ↓
CRNN	8	$3 \times 10^{-4}$ (linear)	0	65	768	pad{24} RoP{3}	CER = 0.084 WER = 0.185
Transformers	24	$6 \times 10^{-5}$ (cosine)	0.03	15	1024	FT_epochs{10}	CER = 0.068 WER = 0.142
GAN	12	$1 \times 10^{-4}$ (linear)	0.02	70	256	voc_size{85} n_styles{372}	WID = 0.42 KID = 0.51
RoBERTa-Med	48	$5 \times 10^{-5}$ (linear)	0.04	10	768	voc_size{8k} mlm{0.15}	Loss = 1.25 PP = 4.64

## V EXPERIMENTS

In this work we will explore two training corpus:

1. The full documentary medieval and Early modern corpus (See Table 1).
2. The aforementioned corpus + 420k synthetic medieval data pairs. (TR  $\gamma$ )

The VLM (Visual Language Models) architectures were trained on three flavours:

1. The TrOCR large (1024 dimensions) weights with english RoBERTA as decoder (TR  $\alpha$ )
2. The TrOCR large weights using the RoBERTa medieval as tokenizer. (TR  $\beta$ )
3. The Vision Transformer (ViT) weights using the GPT-2 medieval as tokenizer and decoder on causal mode (TR  $\delta$  & TR  $\gamma$ )

The CRNN models, trained on Kraken v4, used here as our baseline, will be also trained on aforementioned datasets. (CRNN  $\alpha$  and CRNN  $\beta$ )

We generated a total of 6 models (See Table 4) with three primary objectives in mind:

1. Exploring the advantages and limitations of using a Transformers-based model compared to the CRNN+CTC solutions.
2. Determining the impact of integrating an LLM model into the decoder part of the architecture.
3. Measuring the enhancement provided by synthetic material, both during the pre-training and the training phase of the models.

The performance of each model will be evaluated on two scenarios :

1. Validation during training on 10% of the train-dataset (Table 4)
2. Validation against 6 external manuscripts in few-shot (Table 5) and zero-shot mode (tables 5 and 6).

## VI RESULTS

The Results will be presented using widely recognized metrics, including CER (Character Error Ratio), WER (Word Error Ratio) and BERT-score, which computes cosine similarity between candidate texts and reference texts in a BERT-embedded space. These metrics allows us to understand not only how accurately the model is transcribing lines at a word and character level, but also how well it preserves the meaning, and by extension, the readability of the original content. Besides metrics will be computed on both, raw mode (R), without applying normalization, and in a cleaned mode (C), ignoring diacritics and punctuation’s signs.

**Table 4** Evaluation results for the six models during training. CRNN : *Convolutional Recurrent Networks*;  $\phi$  : Type of textual sequence (R : raw; C : cleaned); CER : Character Error Rate; WER : Word Error Rate; BERT : BERT-score)

Model code name / Scores	$\phi$	Model Content	Lines	Validation			Improving		
				CER	WER	BERT	$\Delta$ CER	$\Delta$ WER	$\Delta$ BERT
CRNN $\alpha$	R	Kraken v4 (CRNN + CTC) from scratch (only GT)	210 218	7.1	20.9	92.2			
	C			6.4	19.3	93.0			
TR $\alpha$	R	FT on TrOCR-large (Tridis) Pretrain on IAM (GT)	210 218	10.0	23.8	90.8			
	C			8.5	20.7	92.2	-33%	-7%	-1%
TR $\beta$	R	FT on TrOCR-large + med-RoBERTa as decoder (GT)	210 218	7.6	16.4	93.2			
	C			6.4	14.3	94.2	+0%	+26%	+2%
TR $\gamma$	R	TR $\beta$ config (GT + Synthetic dataset)	630 654	6.4	14.4	94.5			
	C			5.7	12.9	95.3	+11%	+33%	+3%
TR $\delta$	R	FT on Vit-base + + med-GPT2 as decoder (GT)	210 218	11.2	28.4	89.4			
	C			9.7	24.8	90.9	-52%	-28%	-2%
CRNN $\beta$	R	Kraken v4 (CRNN + CTC) from scratch (GT + synthetic)	630 654	6.2	17.8	93.3			
	C			5.8	16.4	93.8	+9%	+15%	+1%

During our experiments, the CRNN models, serving as our baseline, demonstrated notable improvements when synthetic data was incorporated into the training process. Specifically, the CRNN  $\beta$  model, which was trained with both ground-truth and synthetic data, achieved a CER of 0.058 and a WER of 0.164. This represents a relative improvement of 9% and 15% in CER and WER validation over the CRNN  $\alpha$  model, which was trained solely with ground-truth data. These results underscore the significant impact of synthetic material, particularly given that the CRNN architecture does not leverage an explicit LLM vocabulary.

**Table 5** Evaluation results on out-of-domain manuscripts using models on Zero-shot and Few-Shot (150 lines) modes. CRNN (Models based on Convolutions + Recurrent networks. TR: Models based on encoder-decoder transformers.(See table 4).  $\phi$  : Type of textual sequence (Z: Zero-Shot, F: Few-Shot).  $\Delta$ : improvement percentage comparing to the baseline.

Model / Scores	$\phi$	Saint-Gall			$\Delta$		Cremma exp.			$\Delta$		Faithful			$\Delta$	
		CER	WER	BERT	CER	WER	CER	WER	BERT	CER	WER	CER	WER	BERT	CER	WER
CRNN $\alpha$	Z	12.5	41.8	85.7			11.0	50.4	87.9			13.2	43.3	83.2		
	F	10.1	35.6	87.4			9.5	46.7	88.9			11.7	39.6	84.4		
TR $\alpha$	Z	13.1	36.4	85.5	-5	+13	10.9	44.6	88.1	+1	+12	11.9	39.8	86.3	+10	+8
	F	9.8	26.1	89.0	+3	+27	9.2	41.5	89.0	+3	+11	9.4	30.2	87.6	+20	+24
TR $\beta$	Z	12.0	33.8	86.5	+4	+19	10.5	42.0	88.6	+5	+17	11.6	38.0	87.1	+12	+12
	F	9.0	25.4	89.8	+11	+29	8.7	39.2	89.8	+8	+16	9.1	29.2	87.7	+22	+27
TR $\gamma$	Z	10.9	30.4	88.1	+13	+27	9.8	41.2	89.3	+11	+18	11.2	37.8	87.3	+15	+13
	F	7.6	24.0	91.5	+25	+33	8.1	37.3	90.6	+15	+20	8.5	27.9	88.5	+27	+29
TR $\delta$	Z	14.2	40.2	83.7	-14	+4	12.2	48.4	86.2	-11	+4	13.6	42.6	84.4	-3	+1
	F	11.7	32.2	85.8	-16	+9	11.0	46.6	87.6	-16	+0	12.2	38.6	85.1	-4	+2
CRNN $\beta$	Z	12.1	40.7	86.2	+3	+3	10.4	47.8	88.5	+6	+5	12.2	41.3	86.5	+8	+5
	F	9.5	34.2	87.8	+6	+4	9.1	44.8	89.2	+4	+4	9.0	31.8	87.4	+23	+20

**Table 6** Evaluation results on out-of-domain manuscripts. CRNN (Models based on Convolutions + Recurrent networks. TR: Models based on encoder-decoder transformers.(See table 4).  $\phi$  : Type of textual sequence (R: raw, C: cleaned).  $\Delta$ : improvement percentage comparing to the baseline.

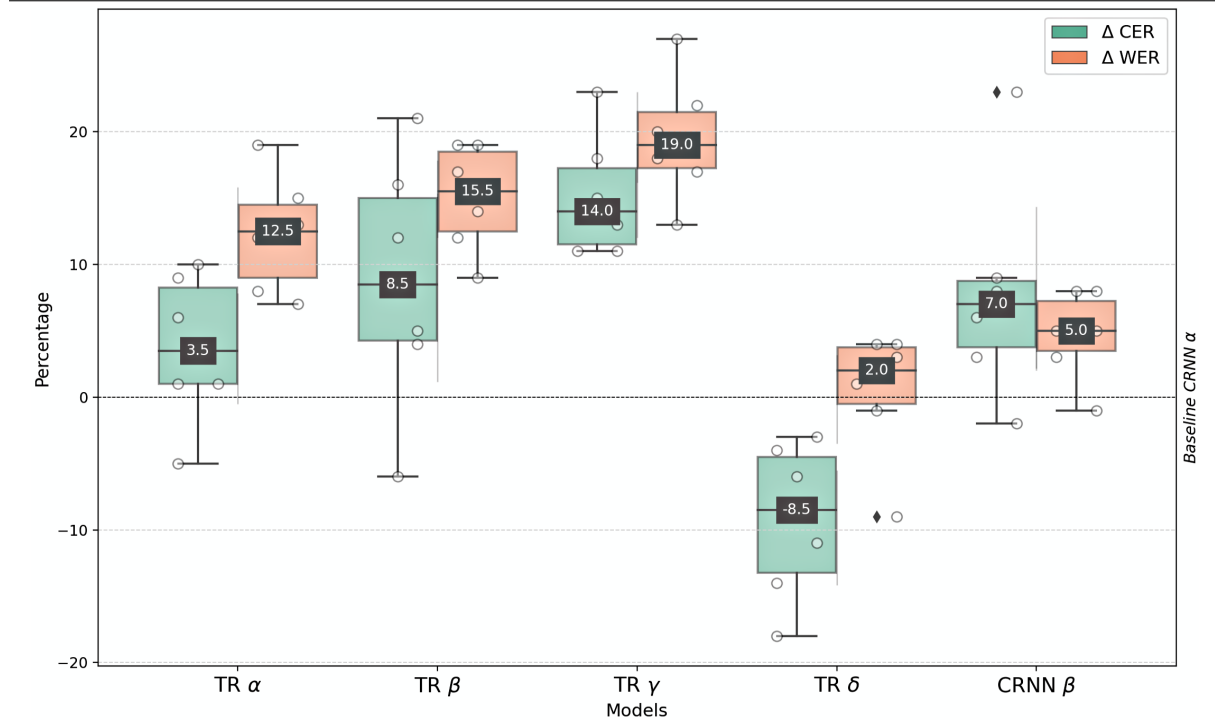
Model / Scores	$\phi$	Collatio			$\Delta$		S. Abstinentia			$\Delta$		Liber Feudorum			$\Delta$	
		CER	WER	BERT	CER	WER	CER	WER	BERT	CER	WER	CER	WER	BERT	CER	WER
CRNN $\alpha$	R	29.5	86.1	73.1			18.7	56.9	81.8			15.6	40.8	83.4		
	C	28.0	84.2	74.0			16.5	48.1	83.6			12.6	33.6	85.5		
TR $\alpha$	R	29.1	80.4	77.7	+1	+7	17.6	46.3	85.8	+6	+19	14.2	34.9	86.5	+9	+15
	C	24.0	67.5	79.0	+14	+20	14.4	37.8	86.8	+13	+21	11.0	26.9	88.5	+13	+20
TR $\beta$	R	23.4	73.8	79.5	+21	+14	19.8	51.5	82.2	-6	+9	13.1	32.9	87.3	+16	+19
	C	19.5	64.2	80.0	+30	+24	17.5	46.6	84.0	-6	+3	10.0	24.8	89.3	+21	+26
TR $\gamma$	R	22.7	71.6	80.3	+23	+17	16.7	44.5	86.5	+11	+22	12.8	32.7	87.5	+18	+20
	C	18.3	59.5	81.8	+34	+29	13.7	36.5	87.5	+17	+24	9.6	24.1	89.6	+24	+28
TR $\delta$	R	31.2	87.2	74.0	-6	-1	22.1	61.9	78.8	-18	-9	16.3	39.4	84.8	-4	+3
	C	27.7	81.2	75.8	+1	+4	19.4	52.0	81.3	-18	-8	12.9	32.0	87.8	-2	+5
CRNN $\beta$	R	22.8	79.3	76.1	+23	+8	17.1	52.1	83.5	+9	+8	16.0	41.4	83.1	-2	-1
	C	21.1	76.6	77.1	+24	+9	14.9	41.9	85.2	+10	+13	12.8	34.0	85.3	-1	-1

Interestingly, the TR  $\alpha$  model, trained on the English weights of TrOCR, and the TR  $\delta$  model, which employs a specific LLM (GPT-2) and ViT as encoder, produced subpar classifiers (See Figure 4). These experiments were designed to investigate the performance of standard pre-trained weights (TrOCR + RoBERTa English) and auto-regressive modes on HTR. Unlike TR  $\beta$  and TR  $\gamma$ , which leverage historical-specific tokenizers and synthetic data, TR  $\delta$  relies on generic pre-trained weights (ViT) on causal mode (GPT-2). On the other hand, the TR  $\beta$  model’s exhibits an important improvement (+26% on WER metrics and +2% in the BERT-score) by simply replacing the decoder with a historical-specific. This suggests that while general pre-training provides a solid foundation, aligning the tokenizer and decoder with the target domain, even when other components of the architecture remain unchanged, can become crucial for achieving optimal performance in specialized tasks such as HTR on historical manuscripts as it allows the model to better adheres to the linguistic and stylistic conventions of these documents.

Moreover, the addition of synthetic data in the training of transformer models led to further enhancements. The TR  $\gamma$  model, our best-performing model overall, trained with both ground-



**Figure 4** % of relative progression in CER and WER across CRNN and TRansformers models in zero-shot scenarios for the six external test corpus (Tables 5 and 6). Values in the box plots are the medians.



truth and synthetic data, achieved a CER of 0.064 and a WER of 0.143 during training validation. This represents a relative improvement of 11% in CER and a remarkable 33% in WER compared to the baseline (See Table 4). This indicates that the inclusion of such data can bridge the gap between general pre-training and domain-specific requirements. Synthetic data not only enhances the model’s ability to generalize but also helps in capturing new patterns of medieval scripts, which heterogeneity can not be fully represented in standard datasets.

On out-of-domain external manuscripts test (Tables 5 and 6), the models exhibited varying levels of performance. However, the trend of improved performance with the use of Medieval RoBERTa as a tokenizer and the inclusion of synthetic data in training remained evident (See Figure 4). For instance, our best TR  $\gamma$  model surpassed the baseline in CER, WER, and BERT score, with a median improvement of 14% in CER and 19% in WER in ‘raw mode’. The performance further improved in ‘cleaned mode’, achieving 26% in CER and 27% in WER. This suggests that a significant portion of the transformers’ errors are concentrated on punctuation, diacritics, and spaces (See Section VII). Therefore, future work should focus on enhancing the model’s handling of these elements. This could involve developing more sophisticated tokenization strategies or adding additional pre-training on datasets that address these aspects.

In few-shot scenarios (150 lines, ca.3 pages), there was an overall progression of 2-3 points in CER and 5-8 points in WER for all models compared to zero-shot. On homogeneous corpora such as *Saint-Gall* and *Cremma*, the best models (TR  $\gamma$ , TR  $\beta$ ) achieved SOTA performance in zero-shot mode (circa 0.10 CER), consistently outperforming the baseline in WER. In the case of *Cremma* and *Summa of Abstinencia*, the high WER must be attributed to the nature of the editors transcription, which prefers to follow the original writing spacing (*scripta continua*) and does not introduce full modernization by splitting words as HTR models do. Moreover, when applied to linguistically diverse corpora such as *Faithful*, which includes languages that are either novel (Low German, Dutch) or have limited training data for the model (German variants),



transformer-based solutions face significant challenges due to their inability to leverage specific linguistic knowledge. Their performance are robust but only marginally superior to that of best CRNN models in all metrics. (cf. TR  $\gamma$  vs CRNN  $\beta$ ).

It is noteworthy that even subpar models such as TR  $\alpha$  delivered useful inferences in out-of-domain scenarios (Table 6). This is not the case for TR  $\delta$ , which, despite being pre-trained with specific ancient texts, delivers degraded inferences in all cases. The model’s reliance on a strict causal mode with GPT-2 limited its ability to utilize bidirectional context, leading to inaccurate predictions. This highlights the importance of considering both the training data and the mode of operation when designing models for historical text recognition. Interestingly, TR  $\alpha$ , despite its poor validation performance, provided better inferences in CER and WER in real-world cases. This underscores the effectiveness of transformers in generalization demands, which is particularly advantageous in low-resource settings such as historical document transcription where annotated data is scarce. The performance of specific-data models like TR  $\beta$  and TR  $\gamma$  in zero-shot tests, which largely exceeded their validation results, further supports this observation.

On the other hand, the CRNN  $\beta$  model, although improved with synthetic data compared to the baseline, still lags behind transformer-based models (+7% and +5% in CER and WER in median improvement against 14% and 19% respectively from TR  $\gamma$ ). As we will see in section VII, this discrepancy is not only due to the absence of an explicit language model but also stems from the inherent advantages (v.g. The self-attention mechanism and the Maximum Likelihood Estimation) of the transformer architecture in capturing implicit language patterns and dependencies in historical manuscripts.

## VII ERROR TAXONOMY

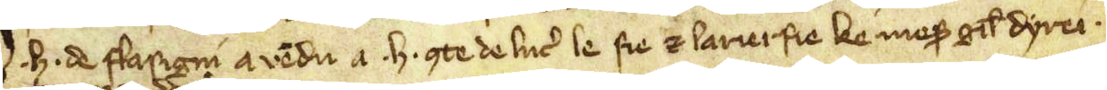



*Word or Sub-word Mis-predictions.* Most errors from Transformer models are linked to mis-predictions at the word or sub-word level, paradoxically increasing the CER more than the WER. Transformers, trained on large datasets, predict words based on observed statistical patterns. During inference, they leverage their understanding of broader linguistic patterns to complete difficult readings with statistically plausible vocabulary items. This strategy is highly effective, as evidenced by the significant difference in WER between Transformer and CRNN models. CRNN models often propose seemingly *verbatim* transcriptions due to falling into local optima during training, leading them to “imitate” the input, especially when faced with unfamiliar words or phrases, rather than understanding the underlying linguistic patterns. Consequently, many of their outputs ignore the lexical relationships within the text, resulting in transcriptions that may be technically accurate at the character level but correspond to non-existent or nonsensical words.

*Statistical Pattern Misleading.* However, as observed in the examples displayed in Figure 5, Transformers can also be misled by the statistical patterns learned during training, leading them to propose incorrect readings in situations where first inferences are not in the vocabulary, were poorly observed during training or conversely, were highly observed, causing a bias problem. This is a common scenario in ancient documents, where stereotyped (formulaic) discourse co-exists with unique occurrences, specific vocabulary, and dialectal variants, a challenge exacerbated by the fragmentary nature of surviving collections. Named entities, for instance, are a well-known example of this as they do not belong to the language dictionary and could appear as rigid designators. Experience indicates that in many cases (v.g place names, locative names, surnames, periphrastic names) they can benefit more from imitative transcriptions than from vocabulary-based completion.

It's well-known that in medieval times, the pool of personal names was quite restricted, leading to frequent abbreviations. Developing these abbreviations correctly is impossible unless to rely on external knowledge bases, a requirement that poses a significant challenge for HTR pipelines. For instance, in Example 1, two instances of "H." ("Henrions" and "Henri") were completed as "Hugues" (a common baptismal name) in the first instance and verbatim transcribed in the second ("H."). Such false positives led by beam-based decoding can highly penalize TR accuracy at CER level as they introduce many individual character mistakes.

**Figure 5** Four transcribed lines from external datasets. For each one three model erroneous transcriptions are displayed. GT (ground-truth), TR (Transformers), CRNN (Convolutional Recurrent models).

1. ANLux, A-X-42-1. fol. 36r (Liber Feudorum)
2. Wien ONB Cod. 2160. fol. 184r (Collatio)
3. Médiathèque de Troyes, Ms 1600. fol. 36r (Summa of Abstinencia)
4. Leipzig, UB, Ms 758. fol. 24v (Faithfull Transcriptions)

1		
GT	H[enrions] de Flasiigni a vendu a H[enri] conte de Lucel[burch] le fié et l'arierfié ke messires Giles d'Yrei	
TR $\gamma$	Hugues de Flasiigni a vendu a H. conte de Lucembour le fié et l'arier fié ke mesigneur Giles Dyrei	
TR $\beta$	Hugues de Flasiigni a vendu a H. conte de Luce le fié et larier fié ke messire Giles d'Yrei.	
CRNN $\beta$	H. de Flasiigni a Vendu a H. ote de Lutre le fié et larier fié kee Mesigre il Dyrei	
2		
GT	inceps qui in locum defuncti paren-	tis qui ex eodem nati sunt succes
TR $\gamma$	inceps quun locum defuncti patri	tis, qui ex eodem natisunt succes
TR $\beta$	inceptis quin locum defuncti Petri	tis qui ex eodem natisunt succes
CRNN $\beta$	inceps quim locum defunctiparei	tisqui ex fodem natisunt succer
3		
GT	omnium peccatorum quia ex quo dyabolus	quando esset infusum, stultus esset
TR $\gamma$	omnium peccatorum quia ex quo dyaconus	quando esset in futurum. stultus esset
TR $\beta$	omni peccatorum quod ex quo dyaus	que esset in fustum scultus esset
CRNN $\beta$	onium pocorum quia ex quo dyas	que essent infusus.stultus essent
4		
GT	omni populo et cetera. Quod figuratum est hester viii° ludeis noua lux oriri	
TR $\gamma$	omni populo etcetera. Quod sigillatum est Hester Villo, Indeis noua lux oriri	
TR $\beta$	omni populo et cetera. Quod signatum est hester : VIII°, indeis noua lux oriri	
CRNN $\beta$	omni propulo et tr Rd figuatum est hester VIII judeis noua lux orui	

**Abbreviation Expansion.** In addition to this, contracted and suspended abbreviations as well as token co-occurrences that are underrepresented in the training data, can be a significant source of errors for similar reasons. This is seen in Examples 1, 2 and 3 where at least four abbreviations have been incorrectly expanded at different levels. In many cases the completion depends on declension or language state variations ("messires" completed as "mesigneur"; "quando" as "que"), which are easier to predict for TR models. However, the models struggle with abbreviations that are rooted in regional or historical contexts. For example, "Lucelburch"

completed as "*Luce*"; "*d'Yrei*" as "*Dyrei*"; "*stultus esset*" as "*scultus esset*") are challenging because they are not widely known and require specific local knowledge that goes beyond general language patterns. Despite these challenges, the contextual and vocabulary-based replacements, even when incorrect, are in general lexically closer to the ground-truth (higher BART-score) and consequently easier to process and read than the transcriptions proposed by the CRNN systems, which in many cases propose hapax or false-lemma outputs.

*Line-level Transcription Limitations.* Besides, since we are working on line-level transcriptions the models may have limited their ability to capture long-range dependencies. This limitation becomes particularly problematic when dealing with short lines, such as those found in double-column manuscripts, or with words split across lines or pages. When confronted with broken words, even if they are well-read, TR models face significant challenges because the reading of a broken word will correspond to an out-of-vocabulary word, a scenario that can be difficult to handle for this kind of models. This issue is clearly illustrated in Example 2 of Figure 5, where all the models struggle to accurately transcribe the end of the line. In this instance, the occurrence "*defuncti paren // tis*" is replaced by "*defuncti patri*" and "*defuncti Petri*" by the Transformer models as their attempt to find a likelihood replacement, is hindered by its inability to visualize the content of the next line.

*Editorial Normalization.* Examples 2 to 4 aims to demonstrate that Transformer models can effectively learn the editorial normalization practices. When working with semi-diplomatic transcriptions, HTR models learns to develop abbreviations, split words, and modernizing punctuation. This is a reflection of their training, which involves large amounts of normalized text. However, the way these models handle word splitting and graphical pauses can vary significantly depending on the manuscript. Word splitting is not always evident in ancient texts, and this can lead to different outcomes for the two types of models. On the same test CRNN models can infer non-existing words ("*defunctiparei*", "*tisqui*", "*natisunt*", "*infusus.stultus*", *figruatum*) due to the model's attempt to transcribe the text in a verbatim mode. On the other hand, TR models apply the separation correctly to correspond to two or more vocabulary tokens. This suggests that TR models are better at recognizing and handling word boundaries which is in general positive for semi-diplomatic transcriptions paradigms, but can hardly affect WER performances on diplomatic ones. This ability likely stems from the model's attention mechanism, which allows it to consider the entire input sequence when making predictions.

*Contextual Syntax Replacements.* Moreover, TR models are more adept at avoiding wrong syntax replacements by exploiting the context, as seen in the examples 2 and 3 ("*ex fodem*" vs "*ex eodem*", "*quando esset*" vs "*que essent*", "*et tr Rd*" vs "*et cetera. Quod*"). This ability to consider the broader context is one of the key strengths of TR models and is a major factor in their superior performance not only on HTR but on most NLP processing tasks. However, it's important to note that this strength can also lead to over-correction errors when the model's assumptions about the context (the statistical best path) do not fit the actual text as we can observe in examples 3 and 4 ("*dyabolus*" replaced by "*dyaconus*", "*esset infusum*" by "*esset in futurum*", "*figuratum*" by "*sigillatum*", "*Iudeis*" by "*Indeis*").

## VIII DISCUSSION

These results demonstrate the potential of transformer-based models for HTR tasks, especially when combined with language-specific decoders and synthetic data. Transformers are particularly effective for tasks that involve capturing contextual and temporal sequences, such as HTR. However, they require large quantities of data (and computational resources) for pre-training

and training phases to be on par with CRNN solutions trained from scratch. This could explain the decrease in performance of our models using an standard Vision Transformer (TR  $\delta$ ) and unspecific encoders which generates mediocre inferences (TR  $\delta$ , TR  $\alpha$ ).

The choice of decoder can also play a significant role on the quality of the output. The use of a Medieval RoBERTa as a decoder (TR  $\beta$ ) leads to initial improvements in performance, especially on token and sub-token levels. The effect is unclear during validation, but when applied to real-world data, the power of a specific LLM model trained on ancient texts clearly surpasses the English RoBERTa used by the original TrOCR. However when the model encounters languages it was not specifically trained on (like those in the Faithful corpus), the advantages of the specialized decoder diminish. The tokenization becomes generic, and the performance may not surpass that of robust CRNN-based models. This proves that the customization of the decoder and tokenizer is not merely a technical detail, but a key consideration in the implementation of transformer-based models for HTR tasks.

Employing synthetic lines in training could provide additional variability and novel data scenarios boosting the generalization capacities of the model face to unseen data. This is evident from the broad improvement in performance of the CRNN  $\beta$  and TR  $\gamma$  models, which are trained with both ground truth and synthetic data. From the results, the risk of over-fitting seems mitigated, which suggest that synthetic data is representative of the real data and the increase in bias do not lead the model to prematurely converge to local minima during the training process.

In an overall scope, these VLM (Visual Language Models) solutions based on transformers outperform classical CRNN models in generating reliable and realistic transcriptions in a one-step training process. This is achieved through the use of explicit LLM decoders combined with implicit language patterns. Transformers, with their parallelizable processing and self-attention mechanisms, excel in handling long-range dependencies and processing multiple data representations simultaneously rather than step-by-step. This allows them to consider the entire input sequence at once, leading to more contextually aware transcriptions. The LLM post-correction process, typically applied on CRNN inferences, is integrated into the training of transformers, eliminating the need for a separate process. Additionally, smoothing the probability distribution over the output vocabulary during training helps the model to generalize more effectively in low-resource scenarios. These combined techniques make transformer-based VLMs an efficient and scalable solution for HTR tasks.

Many of the local errors in TR transcriptions (See Section VII) originate from the inherent transcription level of HTR documentary ground-truth. Firstly, models tend to statistically replicate normalization practices and diplomatics conventions, leading to transcriptions that often interpret content not necessarily present graphically in the manuscript (a common issue in CRNN systems as well). This is particularly evident in the case of abbreviations, signs, and glyphs. Their development depends not only on conventional writing practices, but also on discourse context, local references, and syntactic coherence. Accurately transcribing these elements using semi-diplomatic transcriptions becomes significantly more complex as it goes beyond simple character recognition and involves a deeper understanding of the manuscript's content. Restoring the correct interpretation, especially for heavily abbreviated manuscripts, remains an open challenge and is a primary goal of most post-correction systems. On the other hand, as we have seen in Section VIII, the transformer decoder, when facing a hard lecture, can also propose full and sub-word replacements that are close but do not necessarily match the full graphical writing. This mechanism, based on beam search and vocabulary leverage, while highly effective on WER, can become penalizing on CER or diminishing his boost effect on unseen languages.

Moving to other point, the line-based training approach limits the effectiveness of transformers on large-range sequences. Contexts such as short lines, broken lines, tables, and marginalia pose more difficulties than full lines due to their fragmented or incomplete nature. This limitation reduces the number of output paths and could leads to inconsistencies in the transcription, as the model struggles to maintain the overall structure and flow of the text resulting in errors that are not as prevalent in more continuous text sequences.

These challenges highlight the inherent limitations of current models architectures and suggest that future models could benefit from mechanisms that better control interpreted transcriptions. For instance, integrating external knowledge bases like Retrieval-Augmented Generation (RAG) on named entities or lexical thesaurus embeddings could provide additional specific context. Additionally, more effective handling of short-context scenarios could be achieved by introducing paragraph or full-page level systems and incorporating more sophisticated layout analysis techniques which would allow the model to face complex structures.

## IX CONCLUSION

In this study, we extensively evaluated Visual Language models (VLM) for Handwriting Text Recognition (HTR), specifically tailored for documentary manuscripts from the 10th to the 16th centuries. Our experiments demonstrated that transformer-based models outperform conventional CRNN+CTC models across all metrics. Notably, they improve the classical CER:WER ratio from 3.5:1 to a more favorable 2.5:1. Furthermore, they provide more reliable and realistic transcriptions, thereby enhancing the BERT score, a robust indicator of output legibility. Additionally, our tests on unseen data reveal that these state-of-the-art models can achieve an overall BERT score of 85% in zero-shot scenarios and surpass 90% in few-shot situations (150 lines) across a wide range of manuscript typologies.

Moreover, our experiments have proved that extending the capabilities of VLM systems through domain-specific LLMs and synthetic ground-truth data can further enhance robustness. This combination introduces both explicit and implicit linguistic patterns, which assists the model to decide on more varied and contextually accurate interpretations and reduces the need to apply vocabulary-based corrections in a separate step.

## X MODEL REPOSITORIES

The weights and models supporting this study are available under open source licences:

Transformers models:

[https://huggingface.co/magisttermilitum/tridis\\_HTR](https://huggingface.co/magisttermilitum/tridis_HTR)

[https://huggingface.co/magisttermilitum/tridis\\_v2\\_HTR\\_historical\\_manuscripts](https://huggingface.co/magisttermilitum/tridis_v2_HTR_historical_manuscripts)

Kraken-based CRNN model:

baseline: <https://doi.org/10.5281/zenodo.10788591>

Enhanced: <https://doi.org/10.5281/zenodo.13862096>

GAN ancient-writing model:

<https://github.com/magisttermilitum/HiGANplus>

RoBERTA-med model:

[https://huggingface.co/magisttermilitum/ROBERTa\\_medieval](https://huggingface.co/magisttermilitum/ROBERTa_medieval)

## References

- Esteban Garces Arias, Vallari Pai, Matthias Schöffel, Christian Heumann, and Matthias Aßenmacher. Automatic transcription of handwritten old occitan language. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
- Attwood, Sweeney, Stitts, Audebrand, D’Amico, Geelhaar, Hofmann, and Gnasso. Wien ÖNB Cod. 2160 f. 164-184 Ground Truth from HTR Winter School 2022, January 2023. URL <https://doi.org/10.5281/zenodo.7537204>.
- 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. *LaTeCH-CLfL 2024*, pages 133–159, 2024.
- Alix Chagué and Thibault Clérico. Sharing htr datasets with standardized metadata: the htr-united initiative. In *Documents anciens et reconnaissance automatique des écritures manuscrites*, 2022.
- Julie Claustre and Darwin Smith. e-ndp notre-dame de paris et son cloître (1326-1504). *Revue Mabillon*, 33: 218–226, 2022.
- Thibault Clérico, Malamatenia Vlachou-Efstathiou, and Alix Chagué. Cremma medii aevi: Literary manuscript text recognition in latin. *Journal of Open Humanities Data*, 9:4, 2023.
- Thibault Clérico, Ariane Pinche, Malamatenia Vlachou-Efstathiou, Alix Chagué, Jean-Baptiste Camps, Matthias Gille Levenson, Olivier Brisville-Fertin, Federico Boschetti, Franz Fischer, Michael Gervers, et al. Catmus medieval: A multilingual large-scale cross-century dataset in latin script for handwritten text recognition and beyond. In *International Conference on Document Analysis and Recognition*, pages 174–194. Springer, 2024.
- Arthur Flor de Sousa Neto, Byron Leite Dantas Bezerra, Alejandro Héctor Toselli, and Estanislau Baptista Lima. Htr-flor: A deep learning system for offline handwritten text recognition. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 54–61. IEEE, 2020.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- Andreas Fischer, Volkmar Frinken, Alicia Fornés, and Horst Bunke. Transcription alignment of latin manuscripts using hidden markov models. In *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, pages 29–36, 2011.
- Ji Gan, Weiqiang Wang, Jiayu Leng, and Xinbo Gao. Higan+: Handwriting imitation gan with disentangled representations. *ACM Trans. Graph.*, 42(1), 2022. doi: 10.1145/3550070. URL <https://doi.org/10.1145/3550070>.
- Yingbo Gao, Weiyue Wang, Christian Herold, Zijian Yang, and Hermann Ney. Towards a better understanding of label smoothing in neural machine translation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 212–223, 2020.
- Loïc Grobol, Mathilde Regnault, Pedro Ortiz Suarez, Benoît Sagot, Laurent Romary, and Benoît Crabbé. Bertrade: Using contextual embeddings to parse old french. In *13th Language Resources and Evaluation Conference*, 2022.
- Paul Guérin. *Recueil de documents concernant le Poitou contenus dans les registres de la Chancellerie de France*, volume 11. Société des archives historiques de Poitou, 1881.
- Estelle Guéville and David Joseph Wrisley. Transcribing medieval manuscripts for machine learning. *Journal of Data Mining & Digital Humanities*, 2024.
- Colette Halter-Pernet, Simon Teuscher, Tobias Hodel, Lukas Barwitzki, Salome Egloff, Fabian Henggeler, Michael Nadig, Anina Steinmann, Sabine Stettler, and Ismail Prada Ziegler. Charters and Records of Königsfelden Abbey and Bailiwick (1308-1662), October 2021. URL <https://doi.org/10.5281/zenodo.5179361>.
- Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 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 4, pages 19–24. IEEE, 2017.
- Benjamin Kiessling. Kraken—an universal text recognizer for the humanities. In *ADHO, Éd., Actes de Digital Humanities Conference*, 2019.
- Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. escriptorium: an open source platform for historical document analysis. In *2019 international conference on document analysis and recognition workshops (icdarw)*, volume 2, pages 19–19. IEEE, 2019.
- Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu



- Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13094–13102, 2023.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL <https://arxiv.org/abs/1907.11692>.
- Najoua Rahal, Lars Vögtlin, and Rolf Ingold. Approximate ground truth generation for semantic labeling of historical documents with minimal human effort. *International Journal on Document Analysis and Recognition (IJ DAR)*, pages 1–13, 2024.
- Peter Robinson and Elizabeth Solopova. Guidelines for transcription of the manuscripts of the wife of bath’s prologue. *The Canterbury Tales Project Occasional Papers*, 1:19–52, 1993.
- Pedro Sánchez-Prieto Borja. Desarrollo y explotación del” corpus de documentos españoles anteriores a 1700”(codea). *Scriptum Digital*, (1):0005–35, 2012.
- Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.
- Phillip Benjamin Ströbel, Simon Clematide, Martin Volk, and Tobias Hodel. Transformer-based htr for historical documents. *arXiv preprint arXiv:2203.11008*, 2022.
- Dominique Stutzmann, Jean-François Moufflet, and Sébastien Hamel. La recherche en plein texte dans les sources manuscrites médiévales: enjeux et perspectives du projet himanis pour l’édition électronique. *Médiévales. Langues, Textes, Histoire*, 73(73):67–96, 2017.
- Dominique Stutzmann, Sébastien Hamel, Iseut de Kernier, Günter Mühlberger, and Günter Hackl. Himanis guérin, September 2021a. URL <https://doi.org/10.5281/zenodo.5535306>.
- Dominique Stutzmann, Sergio Torres Aguilar, and Paul Chaffenet. HOME-Alcar: Aligned and Annotated Cartularies, 2021b. URL <https://zenodo.org/record/5600884>. Zenodo: <https://doi.org/10.5281/zenodo.5600884>.
- Solène Tarride, Mélodie Boillet, and Christopher Kermorvant. Key-value information extraction from full handwritten pages. In *International Conference on Document Analysis and Recognition*, pages 185–204. Springer, 2023.
- Solène Tarride, Yoann Schneider, Marie Generali-Lince, Mélodie Boillet, Bastien Abadie, and Christopher Kermorvant. Improving automatic text recognition with language models in the pylaia open-source library. *arXiv preprint arXiv:2404.18722*, 2024.
- Sergio Torres Aguilar and Vincent Jolivet. Handwritten text recognition for documentary medieval manuscripts. *Journal of Data Mining and Digital Humanities*, 2023.
- Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.
- Chahan Vidal-Gorène, Jean-Baptiste Camps, and Thibault Clérice. Synthetic lines from historical manuscripts: an experiment using gan and style transfer. In *International Conference on Image Analysis and Processing*, pages 477–488. Springer, 2023.
- Patrick Von Platen. How to generate text: using different decoding methods for language generation with transformers. *Access mode: https://huggingface.co/blog/how-to-generate*, 2020.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, 2021. URL <https://aclanthology.org/2021.naacl-main.41>.