

CHURRO: Making History Readable with an Open-Weight Large Vision-Language Model for High-Accuracy, Low-Cost Historical Text Recognition

Sina J. Semnani Han Zhang Xinyan He Merve Tekgürler Monica S. Lam
Stanford University, Stanford, CA
{sinaj, parisz, xinyanh, lam}@cs.stanford.edu, mtekgurl@stanford.edu

Abstract

Accurate text recognition for historical documents can greatly advance the study and preservation of cultural heritage. Existing vision-language models (VLMs), however, are designed for modern, standardized texts and are not equipped to read the diverse languages and scripts, irregular layouts, and frequent degradation found in historical materials.

This paper presents  CHURRO, a 3B-parameter open-weight VLM specialized for historical text recognition. The model is trained on CHURRO-DS¹, the largest historical text recognition dataset to date. CHURRO-DS unifies 155 historical corpora comprising 99,491 pages, spanning 22 centuries of textual heritage across 46 language clusters, including historical variants and dead languages.

We evaluate several open-weight and closed VLMs and optical character recognition (OCR) systems on CHURRO-DS and find that CHURRO outperforms all other VLMs. On the CHURRO-DS test set, CHURRO achieves 82.3% (printed) and 70.1% (handwritten) normalized Levenshtein similarity, surpassing the second-best model, Gemini 2.5 Pro, by 1.4% and 6.5%, respectively, while being 15.5 times more cost-effective.

By releasing the model and dataset, we aim to enable community-driven research to improve the readability of historical texts and accelerate scholarship.²

1 Introduction

Historical documents—including letters, diaries, government records, and newspapers—are primary sources that offer invaluable insights into the thoughts, experiences, and contexts of individuals and communities in the past. The vast diversity of

¹CHURRO stands for **C**ollection of **H**istorical **U**nified **R**esources for **R**obust **O**CR.

²Dataset, model, and code are available at <https://github.com/stanford-oval/Churro>.

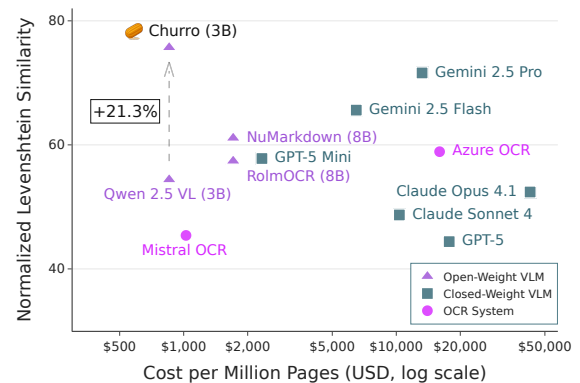


Figure 1: Summary of the performance-cost trade-off across different VLMs on historical documents, measured on the full CHURRO-DS test set. CHURRO outperforms the 3B-parameter Qwen 2.5 VL by 21.3%. It also outperforms the second-best model (Gemini 2.5 Pro) by 4.1% while being 15.5 times more cost-effective. Costs for closed models are measured with the 50% batching discount offered by providers. The costs of open-weight models are estimated following Poznanski et al. (2025).

languages, including historical variants of modern languages and dead languages, alongside evolving writing conventions, handwriting techniques, spelling, and typefaces, requires specialized expertise to study these materials.

Moreover, document degradation caused by environmental factors over time further complicates text extraction (Sulaiman et al., 2019). Historical documents frequently contain features such as marginalia (Cheng et al., 2024), rubrication (Whetter, 2017), and illumination (Wyatt and Tymms, 1861), which are rarely encountered in modern texts. In addition, the wide variety of fonts used in printed documents and the lack of consistent orthographic standards (e.g., multiple spellings of the same word within a single document) pose further challenges (Drobac et al., 2017).

Handwritten text recognition (HTR) (AIKendi

et al., 2024) is even more challenging due to the vast diversity of handwriting styles and calligraphic traditions (Blair, 2020), even within the same language. Historical scribes frequently abbreviated words to expedite writing, resulting in thousands of unique abbreviations (Candido and Aluísio, 2009), especially in pre-modern documents (Guéville and Wrisley, 2022; Rogos-Hebda, 2025).

To improve accessibility, many efforts have been undertaken over the past two decades to digitize historical materials by scanning and converting them into machine-readable text (Oberbichler et al., 2022). The current state-of-the-art approach involves training collection-specific models using fine-grained (paragraph-level or line-level) annotated data (Reul et al., 2019). Various academic competitions (e.g., RASM (Clausner et al., 2018), REID (Clausner et al., 2017, 2019a)) have been organized to advance these efforts. Several platforms such as Transkribus (Kahle et al., 2017) and eScriptorium (Kiessling et al., 2019), provide tools for annotation and model training. However, reliance on annotated training data remains a significant bottleneck.

The recent advent of vision-language models (VLMs) (Liu et al., 2023a; Zhang et al., 2023; Bai et al., 2023; Gemini Team Google, 2023; OpenAI et al., 2023) with strong zero-shot vision capabilities holds promise for addressing these challenges. OCR-related benchmarks such as DocVQA (Mathew et al., 2021), CharXiv (Wang et al., 2024b), CC-OCR (Yang et al., 2024b), OmniDocBench (Ouyang et al., 2024), and OCR-Bench (Liu et al., 2024b; Fu et al., 2025) have become standard among VLM developers. However, these benchmarks primarily target modern documents.

We introduce CHURRO-DS, a large-scale page-level dataset for OCR and HTR on historical corpora. CHURRO-DS is the most diverse historical dataset to date, created by unifying and cleaning over 150 existing datasets. It contains 99,491 documents from 46 *language clusters*, 29 of which appear in the validation and test sets: Arabic, Bangla, Bulgarian, Catalan, Chinese, Czech, Dutch, English, Finnish, French, German, Greek, Hebrew, Hindi, Italian, Japanese, Khmer, Latin, Norwegian, Persian, Polish, Portuguese, Romanian, Sanskrit, Slovenian, Spanish, Swedish, Turkish, and Vietnamese. The dataset also encompasses 14 *scripts* from European, East

Asian, Southeast Asian, Middle Eastern, and Indic script families. Documents include newspapers, books, handwritten diaries, government records, and more, spanning from the 3rd century BC to the 20th century and covering three writing directions. Each page includes an image paired with a full-text annotation, making it suitable for both benchmarking and training VLMs. Figure 2 shows the breakdown of language clusters and several example pages from CHURRO-DS.

We introduce 🍌 CHURRO, a low-cost open-weight VLM that outperforms larger models by fine-tuning a relatively small 3-billion-parameter VLM on CHURRO-DS. The CHURRO VLM achieves 82.3% and 70.1% on the printed and handwritten subsets of CHURRO-DS, respectively, outperforming all open and commercial VLMs by at least 1.4% and 6.5%. This result highlights the effectiveness of specialized training on carefully curated historical documents.

We benchmark a wide range of VLMs and OCR systems. To the best of our knowledge, this is the first study assessing how well VLMs process images of historical documents. Our findings reveal substantial room for improvement among state-of-the-art VLMs in the zero-shot setting. The best-performing commercial model achieves 80.9% and 63.6% normalized Levenshtein similarity on printed and handwritten documents, respectively, averaged over all languages in CHURRO-DS. The best open-weight VLMs we evaluated lag behind, achieving 72.7% and 54.5%, respectively.

2 Related Work

2.1 Document Understanding Benchmarks

Most existing document benchmarks focus on modern documents and layouts. PubLayNet (Zhong et al., 2019) and DocBank (Li et al., 2020) automatically annotate PDF articles from PubMed Central and arXiv, respectively, for layout segmentation. The RVL-CDIP (Harley et al., 2015) and IIT-CDIP (Lewis et al., 2006) datasets are derived from industry documents for document classification and retrieval. CCpdf (Turski et al., 2023) and WordScape (Weber et al., 2023) leverage Common Crawl to extract and annotate multilingual, visually rich documents.

Building on this foundation, recent benchmarks such as CC-OCR (Yang et al., 2024b) and OCR-Bench (Liu et al., 2024b; Fu et al., 2025) com-



Figure 2: Overview and examples from the two subsets of CHURRO-DS. Donut charts show the distribution of *language clusters*. For example, the German cluster includes Modern German, Old High German, Middle High German, and Middle Low German.

pile datasets for text recognition, visual question answering, and key information extraction across many domains into unified benchmarks, primarily to evaluate VLMs. However, these efforts typically remain focused on short outputs, modern documents, and text in English or Chinese (Ying et al., 2024; Liu et al., 2023b).

Concurrently, research trends are shifting toward direct, page-level OCR rather than pipeline methods that rely on bounding-box detection. For instance, recent works introduce models capable of page-level OCR but report only limited evaluation either on noisy, LLM-annotated data (Poznanski et al., 2025) or only in English and Chinese (Wang et al., 2024a; Yang et al., 2024a). Our proposed CHURRO-DS advances this direction by providing curated, human-annotated page-level transcriptions across many languages, with explicit page-level reading order.

2.2 Vision-Language Models

CLIP (Radford et al., 2021) and its successors, such as SigLIP (Zhai et al., 2023), laid the foundation for contrastive pretraining between vision and language. Instruction-tuned open-weight VLMs like LLaVA (Liu et al., 2023a), LLaVA-Next (Liu et al., 2024a), InternVL (Chen et al., 2024), and Qwen-VL (Bai et al., 2023) have demonstrated strong performance in reasoning and fine-grained

text understanding. Lee et al. (2023); Wang et al. (2024a) extended these capabilities to variable-resolution input images, and Wang et al. (2024a); Team et al. (2025) incorporated localization and bounding-box detection. These efforts bridge the gap between general-purpose VLMs and specialized models for document structure and layout understanding (Blecher et al., 2024). Proprietary LLMs have also incorporated vision capabilities, such as GPT-4V (OpenAI et al., 2023) and Claude 3 (Anthropic, 2024). Gemini 1.5 (Gemini Team Google, 2023) further improved text localization. Recently, Huang et al. (2025) and Guo et al. (2025) extended the explicit reasoning capabilities of LLMs to vision.

3 The CHURRO-DS Dataset

CHURRO-DS is a comprehensive benchmark and training resource curated for recognizing printed and handwritten text in historical documents. We standardize diverse annotation formats into a single text string per page in proper reading order. This transformation reduces the problem to pure image-to-text, enabling straightforward evaluation of vision-language models.

Through an extensive survey of NLP, computer vision, and humanities literature from the past two decades, we identified 153 historical corpora containing document page images and their cor-

responding transcriptions. We also release the first publicly available historical OCR datasets for Persian and Turkish. CHURRO-DS reflects the practical challenges of processing historical scans: images often contain irregularities such as page creases, scanning artifacts, and extra margins (e.g., parts of scanning tables). Although all images are correctly oriented, slight misalignments are common.

3.1 Dataset Selection

CHURRO-DS comprises two subsets: one for printed documents and one for handwritten documents. Figure 2 provides an overview of the languages included in each subset, along with illustrative examples. Appendix E lists all included datasets.

The selection criteria were: 1) Documents must originate prior to the mid-20th century. 2) Datasets must have licenses permitting research use. 3) Datasets must include human-annotated, gold-standard text covering entire pages.

We exclude distantly supervised datasets, such as VieBookRead (Do et al., 2024). Many historical datasets provide annotations at the character, word, or line level (Belay et al., 2023; Romanello et al., 2021); we include these if their annotations can be combined to reconstruct complete page-level text.

Many of these datasets were originally transcribed by humanities researchers and have not previously been used for OCR, because they lack bounding-box annotations typically required by OCR models. By adopting a page-level task definition, we can leverage these previously untapped resources. Additionally, we create and include the first publicly available HTR dataset for Ottoman Turkish, transcribed by the authors.

3.2 Dataset Curation

Each example in CHURRO-DS consists of (1) an image of a document page, (2) page-level metadata indicating language and script, and (3) a text string representing the page content. The metadata enables fine-grained analysis of evaluated models.

Because the source datasets come from many different origins, they are transcribed in various formats and follow different conventions. To facilitate effective training, we carefully curate the datasets with the goal of producing an output text string that is (1) in the **proper reading order** of the page, (2) as **accurate** as possible, and (3) as **faithful** as possible to the image, a practice known

as *diplomatic transcription* (Romero et al., 2019). We achieve this through a combination of automated (LLM-based) transformations and manual verification, as described below.

Page-Level Metadata. Each image is annotated with metadata indicating language and script. Such information is not always available from the original datasets. Some datasets lack language annotations entirely, while others (e.g., Clausner et al. (2015); Papadopoulos et al. (2013); Romein et al. (2020); Maheshwari et al. (2022a); Katuák et al. (2024)) span multiple languages without indicating which page corresponds to which language.

We leverage ISO 639-3 (Agency, 2007), which contains 7,916 languages and historical variants, grouping them into *language clusters* based on macrolanguage classifications. For languages without macrolanguage categories (e.g., English or German), we use ISO reference codes for contemporary forms as umbrella categories for historical variants. We prompt a VLM with descriptions and examples from each cluster to detect the language of each page. When available, we use dataset documentation to constrain possible language clusters. We then manually review subsets for accuracy.

We classify scripts similarly, using ISO 15924 (International Organization for Standardization, 2022) and Glottolog (Hammarström et al., 2024) script families, yielding 14 distinct scripts from 5 families. Tables 3 and 4 list all language clusters, scripts, and script families in CHURRO-DS.

Reading Order. We standardize various annotation formats (e.g., ALTO XML (Library of Congress, 2025), PAGE XML (Pletschacher and Antonacopoulos, 2010), JSON, plain text) into a single text string per page in correct reading order. While ALTO and PAGE XML support explicit reading-order annotations, many datasets lack them. For datasets with consistent layouts (e.g., single- or double-column), we develop dataset-specific heuristics. For more complex layouts, we employ a VLM to determine reading order and manually spot-check outputs. Manual checks indicate that more than 98% of reading orders are accurate.

Accuracy. We identified annotation issues in sixteen datasets. For instance, Dolfing et al. (2020) omits text in top margins, and datasets from Halter-Pernet et al. (2021) and Najam and Faizullah (2024) contain frequent transcription errors. To

address these problems, we use a VLM (Gemini 2.5 Pro (Kavukcuoglu, 2025)) to suggest corrections based on image bounding boxes. Corrections are subsequently reviewed and validated by the authors.

Faithfulness. Our dataset includes scholarly editions from diverse institutions, each following distinct transcription standards (David Glück, 2018). Some editions employ *modernized transcriptions* with standardized spelling, punctuation, and capitalization, while others adopt *diplomatic transcriptions* (Romero et al., 2019) to faithfully represent original documents.

For CHURRO-DS, we standardize on diplomatic transcriptions, ensuring evaluated systems need only *transcribe what they see*, independent of dataset-specific guidelines. In datasets such as Quirós et al. (2018), which include both original and expanded forms of abbreviations and hyphens, we retain original transcriptions. We exclude significantly modernized datasets where faithful transcription is unrecoverable.

To restore spacing, line breaks, and diacritics from images, we use a VLM guided by image bounding boxes and transcribed text to perform minimal edits. Character-level edit distances between original and edited transcriptions remain below 5%. We manually verify samples from each dataset for accuracy.

Finally, we normalize typographic elements such as fractions (e.g., converting $\frac{1}{4}$, Unicode U+00BC, to 1/4) and ligatures (e.g., Unicode U+FB06 to “st”).

3.3 Dataset Preparation

Data Filtering. Some pages contain minimal text, such as chapter dividers and title pages. We remove pages with fewer than 30 tokens (tokenized by Qwen 2.5 VL (Bai et al., 2025)). Duplicate pages arising from dataset errors or reuse across datasets are identified and removed using Min-Hash (Broder, 1997), eliminating approximately 15,000 near-duplicates. Large images are resized to fit within a 2500 × 2500 pixel box, preserving aspect ratios. Appendix A.1 provides text and image size statistics.

Dataset Splits. To facilitate efficient evaluation across diverse language clusters, we construct balanced development and test sets by randomly sampling 60 pages per language cluster per document type (printed and handwritten). These samples are evenly divided into validation and test sets, with

the remaining documents forming the training set. Due to limited examples for some languages, only 29 of the 46 language clusters appear in the validation and test sets; the remaining 17 appear only in the training set. The resulting dataset contains 97,151 training samples and 1,170 samples each for validation and test sets. Note that training sizes vary substantially, with as few as 6 and 12 samples for printed Chinese and handwritten Portuguese, respectively. Table 3 summarizes the dataset distribution across languages and splits.

4 Experiments and Results

We first evaluate a broad set of commercial and open-weight VLMs on the CHURRO-DS test set in a zero-shot setting. We then investigate the impact of fine-tuning using the CHURRO-DS training set.

4.1 Experimental Setup

 **CHURRO.** We fine-tune a VLM on CHURRO-DS to obtain CHURRO. We selected the 3-billion-parameter Qwen 2.5 VL for its compact size, strong zero-shot OCR performance, and ability to handle high-resolution, variable-sized images—all critical for processing diverse layouts. We fine-tuned this model on the training portion of CHURRO-DS.

VLMs. We select the top proprietary VLMs from the LM Arena Vision Leaderboard (Chiang et al., 2024), excluding earlier versions of some models that were superseded by updated releases. The selected models are GPT-5, GPT-5 Mini, and GPT-5 Nano (OpenAI, 2025a); GPT-4.1, GPT-4.1 Mini, and GPT-4.1 Nano (OpenAI et al., 2025); GPT-4o and GPT-4o Mini (OpenAI et al., 2023); O1 (Jaech et al., 2024), O3, and O4 Mini (OpenAI, 2025b); Claude Sonnet 3.7 (Anthropic, 2024) and 4 (Anthropic, 2025b); Claude Opus 4.1 (Anthropic, 2025a); and Gemini 2.5 Flash and Pro (Kavukcuoglu, 2025).

We also evaluate several strong open-weight, instruction-tuned VLMs of varying sizes: Qwen 2.5 VL (3B and 72B)(Bai et al., 2025); Gemma 3 27B(Gemma Team et al., 2024); MiMo VL (8B)(Xiaomi, 2025); Nemotron Nano VL (8B)(NVIDIA, 2025); InternVL 3.5 (30B)(Wang et al., 2025); R (4B)(Yang et al., 2025); and Phi 4 Multimodal (5B) (Microsoft et al., 2025); and VLMs specifically tuned for page-level OCR: NuMarkdown (8B)(Numind, 2025), olmOCR (8B)(Poznanski et al., 2025), RolmOCR (8B)(AI,

2025), and Nanonets OCR (3B)(Mandal et al., 2025). We exclude models such as DeepSeek-VL2 (Wu et al., 2024), whose supported context length is insufficient for the long texts in CHURRO-DS. All models are evaluated using the same prompt in Table 6.

OCR Systems. We also benchmark several specialized OCR systems: Azure OCR (Microsoft, 2024), which supports handwritten text recognition in five languages and printed text recognition in over 100 languages; and Mistral OCR (Mistral AI, 2025), a multilingual OCR engine optimized for long-form, dense documents.

Unlike VLMs, which directly generate complete text from images, Azure OCR is a *pipeline model*: it first detects text bounding boxes, performs OCR within these boxes, and then determines the correct reading order.

Recognizing the difficulty VLMs face with long inputs, we also evaluate a *hybrid* system that combines Azure OCR with a VLM. In this approach, Azure OCR identifies bounding boxes and their reading order, after which a VLM processes each bounding box individually. This division simplifies the VLMs task by providing smaller image inputs and requiring shorter textual outputs.

We further report the performance of an oracle model that selects, for each example, the highest score among all models. This establishes an upper bound for ensembling methods that combine outputs from multiple systems.

4.2 Evaluation Metrics

To enable direct comparison across fundamentally different approaches, we follow Blecher et al. (2024) and evaluate each systems final textual output against the gold annotations. Specifically, we compute the character-level Levenshtein distance (Levenshtein, 1965), which measures the number of insertions, deletions, and substitutions required to transform one string into another. We normalize this distance by the length of the longer string and convert it to similarity by subtracting it from 1, yielding a score in $[0, 1]$. This metric closely aligns with the Character Recognition Rate (1 minus Character Error Rate) commonly used in OCR studies (Neudecker et al., 2021; Karpinski et al., 2018; Liebl and Burghardt, 2020), but is bounded within $[0, 1]$, enabling comparison across varying text lengths. We refer to this metric as *normalized Levenshtein similarity*.

We compute the metric for each example and

average the scores separately for printed and handwritten document types. For fairness, we apply the same normalization procedures (described in Section 3.2) to all model predictions. For texts in Arabic script, we perform additional normalization of diacritics and hamza using the PyArabic toolkit (Zerrouki, 2010).

4.3 Evaluation Results on Printed Documents

Table 1 presents results for the printed subset of the CHURRO-DS test set. 🍌 **CHURRO attains the highest average performance among all systems (82.3%)**. Fine-tuning improves Qwen 2.5 VL (3B) across all languages by an average of 14.5%, with the largest gains for Japanese (37.0%), Finnish (32.9%), and Swedish (22.9%). Compared to Gemini 2.5 Pro, the much smaller CHURRO performs better on 12 of 18 languages (with the largest gaps in Romanian, Japanese, and Finnish) and matches it on Czech. Low performance on Chinese print is attributable to the CHURRO-DS training set containing only 6 examples, underscoring the importance of training data size.

Among proprietary VLMs, Gemini 2.5 Pro achieves the highest zero-shot performance (80.9%), with other proprietary models performing between 7.2% and 42.2% worse.

Among open-weight VLMs, OCR-specialized models perform best: NuMark-down, olmOCR, and Nanonets OCR reach 72.7%, 69.8%, and 69.7%, respectively.

Among OCR systems, Azure OCR achieves the best average performance (71.9%), falling 10.4% short of CHURRO. The hybrid model, combining Azure OCR with Gemini 2.5 Pro, averages 52.6% and trails at least one of its components in every language. Relative to Azure OCR alone, the hybrid system improves notably for Bengali and Romanian.

We do not observe a clear correlation between model size and performance within model families. For example, the 3B-parameter Qwen 2.5 VL outperforms the 72B-parameter variant from the same family, mirroring trends with Mini models in the GPT-4.1 and GPT-5 families.

Overall, historical Chinese presents the greatest challenge; even the oracle ensemble achieves only 14.4%. This subset includes many two-page newspaper images with vertical text, non-rectangular article boundaries, and interspersed Latin characters.

System	ben	bul	ces	deu	eng	fin	fra	hin	jpn	lat	nld	pol	ron	san	slv	spa	swe	zho	Avg
VLM Fine-Tuned on CHURRO-DS																			
 CHURRO (3B)	70.9	96.1	95.6	82.3	91.0	63.2	89.6	94.6	74.1	92.0	95.7	81.6	79.6	93.1	97.6	91.4	87.2	6.2	82.3
Δ with Qwen 2.5 VL (3B)	+0.9	+12.8	+2.0	+17.7	+9.1	+32.9	+9.7	+16.7	+37.0	+11.6	+20.0	+14.0	+20.5	+20.6	+2.8	+4.2	+22.9	+5.7	+14.5
Closed VLMs (Zero-Shot)																			
Gemini 2.5 Pro	94.5	93.4	95.6	76.9	93.3	55.3	93.5	93.2	61.7	87.5	97.9	80.3	66.3	89.7	97.4	91.3	81.1	6.6	80.9
Gemini 2.5 Flash	91.2	91.0	94.8	69.3	84.6	32.8	80.9	78.6	47.1	85.0	87.1	73.7	60.2	82.1	97.0	94.4	68.9	8.3	73.7
GPT-4.1 Mini	77.3	90.3	93.8	75.3	88.3	40.1	76.6	82.2	55.3	82.3	82.7	73.1	56.1	71.9	96.3	89.9	79.0	5.3	73.1
Claude Sonnet 3.7	85.2	86.4	94.8	64.7	80.1	25.8	76.1	89.8	29.8	85.0	76.1	71.3	64.1	82.1	95.0	84.8	70.2	1.8	70.2
GPT-5 Mini	78.6	83.8	92.6	67.8	86.4	39.3	83.4	78.2	44.3	81.5	80.5	70.4	43.5	65.5	92.9	90.6	72.7	4.6	69.8
Claude Opus 4.1	82.3	82.9	91.5	58.7	78.3	16.8	77.7	87.5	17.3	84.1	70.1	65.3	57.7	79.5	96.3	85.7	65.2	1.8	66.6
O4 Mini	75.8	81.9	89.2	63.2	82.3	21.2	80.5	73.3	38.8	79.5	74.2	64.0	39.4	59.6	95.1	89.0	69.6	2.4	65.5
GPT-4.1	71.4	78.3	89.5	63.7	81.8	34.5	75.0	76.4	19.7	76.9	76.6	64.8	40.9	60.6	90.6	88.0	66.5	3.9	64.4
O3	76.5	77.8	88.0	58.4	81.6	19.7	80.1	76.3	12.0	79.6	73.1	65.5	33.6	59.8	94.8	89.4	60.0	1.5	62.7
O1	74.7	76.0	87.5	58.9	82.1	24.3	72.2	78.1	15.6	77.5	74.0	64.2	35.3	60.2	90.4	89.0	62.6	2.0	62.5
Claude Sonnet 4	73.3	77.6	86.6	56.1	76.2	15.8	69.9	76.5	17.5	83.1	70.7	63.4	51.0	69.1	93.8	78.9	58.8	2.0	62.2
GPT-5	73.6	77.1	86.9	59.2	81.1	19.9	79.8	61.5	14.2	78.2	71.8	64.2	33.1	53.0	90.3	89.6	58.6	1.3	60.7
GPT-4o	66.7	63.0	82.9	48.8	78.3	15.6	70.2	70.6	11.9	74.2	62.8	54.8	40.1	48.1	89.5	86.4	47.1	2.1	56.3
GPT-4o Mini	50.9	58.7	76.9	48.9	77.1	17.0	66.6	57.8	11.9	69.4	56.9	52.1	43.8	42.1	85.2	85.4	47.9	2.4	52.8
GPT-4.1 Nano	36.2	61.0	82.5	54.4	81.3	20.4	67.8	34.1	14.0	71.9	58.2	58.4	41.4	26.1	86.3	82.3	53.6	3.6	51.9
GPT-5 Nano	28.2	27.9	68.2	41.1	73.0	5.3	65.3	28.5	6.1	56.9	32.2	38.7	27.3	13.0	78.0	69.5	37.2	0.5	38.7
Open-Weight VLMs (Zero-Shot)																			
NuMarkdown (8B)	84.7	87.1	95.7	71.9	83.0	39.0	87.1	75.5	55.2	80.6	84.7	77.4	52.9	73.3	95.6	86.9	77.3	1.1	72.7
olmOCR (8B)	79.2	84.8	90.0	66.4	79.7	28.5	83.7	85.0	42.5	82.5	82.9	72.4	51.4	66.4	96.2	90.7	73.2	0.2	69.8
Nanonets OCR (3B)	75.6	85.5	94.7	69.4	85.7	24.8	82.9	83.1	32.9	79.7	78.1	74.2	57.3	71.4	95.7	91.2	72.2	0.4	69.7
Qwen 2.5 VL (3B)	70.0	83.3	93.6	64.6	82.0	30.3	79.9	77.9	37.1	80.4	75.7	67.7	59.1	72.6	94.8	87.2	64.3	0.5	67.8
RolmOCR (8B)	82.7	79.4	87.5	63.9	88.4	17.0	78.4	85.1	35.1	79.2	73.7	65.5	57.9	67.7	86.1	90.7	70.8	1.2	67.2
Qwen 2.5 VL (72B)	87.9	89.6	83.6	53.3	75.7	10.6	80.2	89.0	47.3	85.3	65.9	57.1	57.4	70.2	95.9	89.5	50.9	3.7	66.3
Gemma 3 (27B)	61.6	62.0	78.3	49.7	70.4	21.0	66.8	74.3	18.3	69.1	50.4	52.7	48.7	61.4	89.4	72.5	44.0	5.4	55.3
MiMo VL (8B)	63.8	72.7	82.4	51.0	71.5	7.2	72.6	51.4	26.8	82.6	50.4	49.6	40.0	44.7	92.4	75.7	50.5	2.1	54.8
Skywork R1V3 (38B)	20.7	35.8	70.6	42.3	73.4	17.6	68.7	18.8	12.2	67.3	47.2	46.8	31.5	17.1	81.6	72.3	37.0	2.7	42.4
InternVL 3.5 (30B)	16.4	10.2	56.5	35.3	76.3	13.1	64.6	16.0	13.2	62.8	42.7	31.5	18.8	11.6	77.3	67.5	27.8	2.8	35.8
R (4B)	19.0	17.2	43.9	28.4	67.7	13.9	62.0	25.2	6.3	48.8	32.0	27.3	24.6	16.9	62.5	62.8	27.0	2.9	32.7
Nemotron Nano VL (8B)	7.7	14.2	35.8	34.3	57.1	6.6	60.0	11.7	0.4	55.9	32.7	33.1	14.8	9.4	59.0	63.4	30.9	1.1	29.3
Phi 4 Multimodal (5B)	2.2	1.7	4.0	5.3	42.9	5.2	11.6	3.7	1.8	11.4	7.0	6.3	5.6	3.5	25.6	16.7	7.4	0.7	9.0
OCR Systems																			
Azure OCR	15.3	91.4	91.7	77.9	86.1	65.0	84.5	72.9	53.6	82.6	91.6	81.1	50.7	80.8	93.9	85.5	79.6	9.4	71.9
Mistral OCR	59.4	88.1	86.5	57.1	79.3	28.6	74.7	82.2	3.3	79.8	78.9	60.9	48.6	82.7	93.8	89.3	58.1	1.3	64.0
Azure OCR + Gemini 2.5 Pro	66.9	57.3	74.0	42.4	63.9	10.8	65.9	56.7	28.4	73.2	59.1	48.8	51.0	59.2	55.3	81.0	50.7	2.4	52.6
Ensemble																			
Oracle	95.4	97.8	99.1	87.0	95.4	76.2	94.1	97.5	82.8	95.4	98.9	92.4	87.1	96.7	98.6	98.3	92.8	14.4	88.9

Table 1: Normalized Levenshtein similarity scores of models evaluated on the **printed** subset of the CHURRO-DS test set (higher is better). Columns represent language clusters arranged alphabetically from left to right: Bangla, Bulgarian, Czech, German, English, Finnish, French, Hindi, Japanese, Latin, Dutch, Polish, Romanian, Sanskrit, Slovenian, Spanish, Swedish, Chinese, and the average across all language clusters. Within each system category, models are sorted by average similarity (highest to lowest).

4.4 Evaluation Results on Handwritten Texts

Table 2 presents evaluation results on the handwritten subset of the CHURRO-DS test set.  **CHURRO achieves the highest average performance among all models (70.1%)**, improving upon the original Qwen 2.5 VL (3B) by 27.2%. The largest gains from fine-tuning are observed for Greek (62.6%), Japanese (54.7%), Turkish (42.3%), Hebrew (42.1%), and Persian (36.8%). Compared to Gemini 2.5 Pro, the fine-tuned model performs better on 18 of 21 languages, trailing only on Arabic, Persian, and Norwegian.

Among proprietary models, Gemini 2.5 Pro achieves the highest average similarity (63.6%). Notably, the performance gap between Gemini models and other proprietary VLMs is considerably larger for handwritten documents than for printed ones.

Among open-weight VLMs, Qwen 2.5 VL (72B) leads with an average performance of 54.5%. Smaller, specialized OCR VLMs NuMarkdown, RolmOCR, and Nanonets OCR follow with 51.2%, 49.0%, and 43.2%, respectively.

As in the printed setting, the hybrid system combining Azure OCR with Gemini 2.5 Pro does not outperform its individual components.

All models struggle particularly with Sanskrit, Khmer, and Hebrew; even the oracle ensemble achieves only 26.1%, 32.9%, and 59.1% on these languages, respectively. Greek and Turkish are also challenging for all zero-shot models we evaluated.

Across both handwritten and printed subsets, multiple VLMs outperform OCR systems, underscoring the promise of VLMs for page-level OCR over pipeline methods.

System	ara	cat	deu	ell	eng	fas	fra	heb	ita	jpn	khm	lat	nld	nor	por	san	spa	swe	tur	vie	zho	Avg
VLM Fine-Tuned on CHURRO-DS																						
CHURRO (3B)	77.5	90.2	83.1	67.4	84.0	78.0	81.1	42.3	88.4	59.6	25.7	70.9	83.8	79.7	52.9	21.5	78.0	85.4	69.1	75.1	78.2	70.1
Δ with Qwen 2.5 VL (3B)	+22.5	+11.1	+31.9	+62.6	+16.5	+36.8	+21.9	+42.1	+14.8	+54.7	+23.4	+19.2	+19.8	+12.8	+16.8	+21.4	+26.5	+32.9	+42.3	+16.6	+25.4	+27.2
Closed VLMs (Zero-Shot)																						
Gemini 2.5 Pro	87.0	87.9	73.9	54.6	80.0	81.6	74.5	30.4	88.0	18.0	24.6	66.5	78.6	87.7	45.9	15.8	75.5	76.1	45.7	69.3	74.6	63.6
Gemini 2.5 Flash	80.7	80.4	69.4	47.1	77.2	76.6	71.6	29.0	84.3	12.9	27.2	57.4	73.9	81.2	38.1	11.4	66.3	68.7	42.8	66.3	71.0	58.7
GPT-4.1 Mini	60.6	78.0	55.7	38.3	73.9	50.7	64.8	28.0	79.7	14.7	12.7	57.1	66.0	77.9	36.4	10.2	61.0	57.9	31.2	38.3	62.1	50.2
GPT-5 Mini	54.2	78.4	56.3	34.0	76.8	45.8	66.0	25.9	75.5	12.2	14.8	51.8	60.1	78.6	37.4	9.2	59.7	49.6	30.5	32.5	48.5	47.5
Claude Sonnet 3.7	61.2	83.2	58.6	50.4	75.8	42.8	63.9	29.6	74.8	2.4	8.2	58.0	64.4	75.2	41.8	6.5	60.7	49.7	31.6	11.1	28.3	46.6
O4 Mini	48.6	73.1	48.9	26.8	74.3	38.0	59.7	9.8	73.3	6.3	8.1	47.6	55.2	72.3	30.8	6.7	53.3	39.2	26.3	28.7	44.0	41.5
GPT-4.1	50.7	66.9	48.9	28.3	70.6	42.6	57.2	21.7	68.8	11.1	10.6	48.0	51.7	71.4	34.4	9.6	49.8	38.8	31.6	25.4	31.5	41.4
Claude Opus 4.1	48.1	76.4	52.3	30.7	71.6	39.2	59.0	23.9	69.8	7.3	2.2	50.5	57.2	66.6	33.5	9.2	55.4	39.1	25.6	7.5	18.6	40.2
Claude Sonnet 4	45.3	70.0	48.6	30.4	63.0	33.0	49.6	29.3	64.3	9.6	8.0	46.5	46.4	54.5	34.7	8.6	52.1	30.8	27.1	6.0	20.5	37.1
O1	43.1	63.9	44.2	21.3	67.7	37.7	49.0	6.0	64.7	3.4	0.5	39.1	48.0	67.3	24.0	2.9	46.8	38.3	25.8	18.5	24.7	35.1
GPT-4o	45.1	56.8	39.1	24.1	65.5	30.3	48.9	18.4	55.2	9.2	7.1	37.9	42.8	60.9	27.5	8.8	41.2	31.7	26.2	17.9	23.6	34.2
O3	32.3	64.5	41.5	19.8	71.7	12.2	43.2	8.6	62.3	3.3	0.1	31.7	36.2	71.3	13.8	0.2	47.0	22.5	24.6	13.5	28.7	30.9
GPT-5	38.4	64.3	35.4	19.3	70.6	16.7	44.7	8.1	63.6	3.8	1.1	29.6	36.8	67.3	18.4	0.8	41.2	26.3	19.1	5.4	27.8	30.4
GPT-4o Mini	32.0	49.1	32.4	22.3	64.0	28.0	42.4	14.6	51.3	9.7	3.1	33.2	37.2	53.8	24.7	8.6	38.4	29.7	24.7	10.6	15.4	29.8
GPT-4.1 Nano	26.2	50.4	33.9	22.1	59.9	23.2	42.9	12.6	53.8	8.8	9.2	33.0	35.3	54.6	20.4	4.3	34.1	25.5	18.2	12.4	14.5	28.3
GPT-5 Nano	8.7	24.9	18.8	4.4	50.5	7.2	28.1	1.5	30.6	2.7	1.2	16.1	14.4	41.3	6.4	0.9	16.8	5.6	4.4	1.3	11.6	14.2
Open-Weight VLMs (Zero-Shot)																						
Qwen 2.5 VL (72B)	67.8	85.7	67.0	38.7	77.0	55.9	75.1	16.3	80.4	11.9	0.5	56.1	74.1	74.8	43.8	13.3	67.8	74.8	36.8	68.1	58.0	54.5
NuMarkdown (8B)	55.9	81.0	64.7	24.6	70.2	48.1	67.4	17.7	85.2	11.0	11.9	61.4	72.6	68.5	41.2	10.3	60.6	69.2	16.4	71.5	66.3	51.2
RolmOCR (8B)	51.8	81.9	63.9	14.9	74.1	50.9	67.0	2.8	84.3	7.9	0.6	58.6	69.1	72.3	37.9	9.2	60.7	69.2	21.1	66.6	64.5	49.0
Nanonets OCR (3B)	52.0	77.8	56.0	20.4	62.3	17.3	61.6	0.8	77.5	4.3	0.3	56.4	64.9	67.9	36.2	2.0	57.3	63.2	13.9	57.6	56.8	43.2
Qwen 2.5 VL (3B)	55.0	79.1	51.2	4.8	67.5	41.2	59.2	0.2	73.6	5.0	2.2	51.7	64.1	66.9	36.0	0.2	51.6	52.5	26.7	58.6	52.8	42.9
olmOCR (8B)	44.0	80.6	53.4	11.7	68.0	23.0	60.5	4.3	79.4	3.9	0.4	52.2	64.6	67.2	21.5	0.5	48.0	60.6	13.1	60.0	55.0	41.5
MiMo VL (8B RL)	16.6	76.1	43.9	16.3	63.0	30.7	44.8	1.4	63.1	4.1	8.2	33.8	57.6	57.2	29.2	8.1	46.6	33.3	16.9	44.5	31.6	34.6
Gemma 3 (27B)	35.6	53.4	37.9	29.8	62.1	33.4	46.9	22.7	53.4	9.4	9.1	40.8	42.4	56.0	33.2	8.8	41.7	28.6	24.2	20.4	26.5	34.1
InternVL 3.5 (30B)	8.3	44.4	23.1	17.2	52.4	9.6	32.7	6.3	41.2	5.6	6.9	31.2	27.5	40.5	20.2	1.8	24.9	21.4	11.6	64.4	62.5	26.4
Skywork R1V3 (38B)	14.5	45.2	26.0	16.1	50.6	16.6	36.9	6.0	37.8	5.3	7.0	33.4	30.7	36.4	25.1	6.7	31.9	20.5	14.7	35.4	41.0	25.6
R (4B)	17.7	30.0	22.7	11.9	41.5	18.8	30.0	8.9	30.5	5.2	3.6	28.6	24.8	30.5	21.2	5.1	23.6	21.1	13.0	33.0	34.9	21.7
Nemotron Nano VL (8B)	6.2	28.5	21.1	4.2	38.3	7.7	26.0	3.5	34.0	0.9	2.9	23.5	22.7	37.0	17.1	2.0	23.0	15.7	4.3	1.0	0.8	15.3
Phi 4 Multimodal (5B)	1.1	9.1	8.6	2.6	20.8	1.0	6.5	0.9	12.5	0.7	1.5	8.1	5.2	13.3	5.9	0.6	4.5	5.4	0.9	1.5	1.3	5.3
OCR Systems																						
Azure OCR	74.3	72.7	49.4	20.5	73.5	45.4	60.6	38.6	70.9	11.4	3.7	53.0	58.0	64.1	37.7	14.6	62.2	46.5	39.8	55.1	50.6	47.7
Azure OCR + Gemini 2.5 Pro	59.2	64.9	47.1	32.4	53.9	49.9	55.3	18.0	68.4	5.8	20.4	41.6	32.0	56.5	29.4	11.9	44.5	57.2	31.6	48.8	20.1	40.4
Mistral OCR	60.3	56.3	29.0	5.8	58.6	46.2	38.7	2.7	62.8	3.8	3.6	35.6	33.0	52.4	14.0	1.7	37.4	32.1	37.5	2.4	4.4	29.4
Ensemble																						
Oracle	88.8	92.6	89.4	74.9	89.7	83.5	88.0	59.1	93.1	65.4	32.9	76.2	87.8	89.3	57.8	26.1	82.6	88.9	74.9	78.5	82.9	76.3

Table 2: Normalized Levenshtein similarity scores of models evaluated on the **handwritten** subset of the CHURRO-DS test set (higher is better). Columns represent language clusters arranged alphabetically from left to right: Arabic, Catalan, German, Greek, English, Persian, French, Hebrew, Italian, Japanese, Khmer, Latin, Dutch, Norwegian, Portuguese, Sanskrit, Spanish, Swedish, Turkish, Vietnamese, Chinese, and the average across all language clusters. Within each system category, models are sorted by average similarity (highest to lowest).

4.5 Error Analysis

Reading Order Errors. In a sample of 50 predictions, 42% from zero-shot Qwen 2.5 VL (3B) contain major reading-order errors, compared to 16% for CHURRO.

The most common issue involves column ordering: many CHURRO-DS pages have two or more columns, which some models mishandle. Reading-order errors are also more frequent in East Asian scripts, where the zero-shot model often fails to recognize top-to-bottom writing direction, as illustrated in Figure 6. CHURRO performs substantially better in this regard.

Major Hallucinations. Thirty-six percent of the 50 predictions from the zero-shot Qwen 2.5 VL (3B) exhibit major hallucinations. These arise when a model cannot reliably recognize the text and instead generates content that seems contextually plausible. Figure 14 shows an 18th-century Dutch letter where the model correctly identifies the document as a letter but, unable to transcribe its content, produces text such as “Dit is een brief”

(“This is a letter.”). None of the predictions from CHURRO exhibit hallucinations.

Errors in Gold Annotations. We found that 2 of the 50 examples contain gold-standard text that omits parts of the page. These omissions stem from annotation errors in the original datasets that persisted through our cleaning process. For example, the transcription in Figure 13 begins halfway through the left column.

Other Minor Errors. We observe recurring minor errors across languages and periods. Models often confuse visually similar characters, such as transcribing ‘o’ for ‘ø’. These errors frequently occur in proper names. In Figure 17, the model renders “Suso” as “Süss” and “Birgittiner” as “Bisitiner”. Scan quality can also impact accuracy, as in Figure 14, where a letter bound in a volume leads to many character errors near the spine due to page curvature.

Appendix D provides additional examples and error analyses.

5 Conclusion

In this paper, we introduced CHURRO, a compact vision-language model fine-tuned from Qwen 2.5 VL for historical text recognition. To support its training and evaluation, we curated CHURRO-DS, the largest and most diverse human-annotated text recognition dataset historical text recognition assembled to date. It unifies 155 corpora and nearly 100,000 pages spanning 22 centuries across 46 language clusters. Its diversity in languages, scripts, layouts, and formats makes it a powerful resource for benchmarking and improving VLMs on historical texts.

CHURRO-DS reveals the difficulty of the task: Gemini 2.5 Pro, the best-performing commercial VLM, achieves only 80.9% (printed) and 63.6% (handwritten) normalized Levenshtein similarity. Fine-tuned on CHURRO-DS, CHURRO achieves 82.3% and 70.1%, surpassing Gemini 2.5 Pro while being 15.5 times more cost-effective. These results represent improvements of 14.5% and 27.2% over the base model, highlighting the benefits of training on a diverse historical corpus.

Even imperfect model outputs can aid in digitization of documents as valuable first drafts, substantially reducing the effort required for scholarly correction. Such transcriptions make historical documents accessible to a broader audience, including readers without specialized training in historical scripts.

Limitations

Although CHURRO-DS is by far the most diverse resource for historical text recognition to date, it still underrepresents some languages. This imbalance reflects the scarcity of annotated historical datasets for many lower-resource languages. Notably, CHURRO-DS currently does not include any languages native to the African continent.

Moreover, we only experimented with standard supervised fine-tuning. Future work should explore more advanced training methods in this setting.

Ethical Considerations

No crowdsourcing was conducted in this research. All included datasets have licenses permitting research use (see Appendix E for detailed license information). The CHURRO-DS dataset is released under the *Creative Commons Attribution Share Alike 4.0* license to comply with these licenses. We

do not anticipate any potential harm arising from the use of this dataset.

Acknowledgement

We thank Professor Trevor Getz for valuable discussions.

This work is supported in part by the Verdant Foundation, the Alfred P. Sloan Foundation, Google, Stanford Human-Centered Artificial Intelligence (HAI) Institute, David and Helen Gurley Brown Institute for Media Innovation, Microsoft Azure AI credits, and the NAIRR Pilot program.

References

- 2023. Aldicam-2023: Diachronic and interactive linguistic atlas of the community of madrid. <https://aldicam.corpuscodea.es/consultas.php>. Accessed: 2025-05-18.
- 2025. Osman agha dataset: Images and transcription from british library ms. or. 3213 and kreutel’s 1980 edition. Based on London, British Library, MS. Or. 3213 and Richard F. Kreutel’s 1980 transcription: *Die Autobiographie des Dolmetschers Osman Aga aus Temeschwar*, The E. J. W. Gibb Memorial Trust, Cambridge.
- Nathalie Abadie, Stéphane Bacciochi, Edwin Carlinet, Joseph Chazalon, Pascal Cristofoli, Bertrand Duménieu, and Julien Perret. 2022. [A Dataset of French Trade Directories from the 19th Century \(FTD\)](#).
- ISO 639-3: Language Coding Agency. 2007. Iso 639-3. <https://iso639-3.sil.org/>. Accessed: 2025-03-25.
- Giorgia Agostini. 2024. [giorgiaagostini/lidi1.0-project: Gt upload](#).
- Reducto AI. 2025. Rolmocr: A faster, lighter open source ocr model.
- Tuija Ainonen, Suse Andresen, Lois Bakker, Amy Boylan, Silvia Della Manna, Wiktor Dziemski, C. E. M. Henderson, and 1 others. 2022. [Ground truth for Önb, cod. 3891](#).
- Rachele Alba, Giorgia Rubin, Federico Boschetti, Franz Fischer, Thibault Clérice, and Alix Chagué. 2023. [Htromance: Medieval italian corpus of ground-truth for handwritten text recognition and layout segmentation](#). If you use this dataset, please cite it using the metadata from this file.
- Matos Alexandre, Neves Rui, Monteiro Gonçalo, Coelho Catarina, and Bastos Pedro. 2025. [iforal-dataset](#).

- Wissam AlKendi, Franck Gechter, Laurent Heyberger, and Christophe Guyeux. 2024. [Advancements and challenges in handwritten text recognition: A comprehensive survey](#). *Journal of Imaging*, 10(1).
- Anthropic. 2024. [Introducing the next generation of claude](#).
- Anthropic. 2025a. [Claude opus 4.1](#).
- Anthropic. 2025b. [Introducing claude 4](#).
- Matthias Arnold. 2022. [Early chinese periodicals online \(ecpo\) \[metadata\]](#).
- Davide Aruta, Martina Lenzi, Armelle Le Huërou, Marylène Possamaï, and Ariane Pinche. 2023. [Liber](#). HTR datasets from 14th15th century manuscripts, including Pierre Bersuire’s Old French translation of Titus Livius and Nicolas Trevet’s commentaries. Licensed under CC BY 4.0. Released on 2023-04-13.
- Attwood, Sweeney, Stitts, Audebrand, D’Amico, Geelhaar, Hofmann, and Gnasso. 2022. [Wien Önb cod. 2160 f. 164-184 ground truth from htr winter school 2022](#).
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. [Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond](#). *Preprint*, arXiv:2308.12966.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-vl technical report](#).
- Pedro Barcha. 2025. Old books dataset. <https://github.com/PedroBarcha/old-books-dataset>. Accessed: 2025-05-18.
- Stefan Beckert. 2024. [Ground truth set for handwritten text recognition \(htr/ocr\): Dresdner hofdiarium 1665 \(mscr.dresd.k.80\) 17th century kurrent manuscript](#).
- Birhanu Hailu Belay, Isabelle Guyon, Tadele Mengiste, Bezawork Tilahun, Marcus Liwicki, Tesfa Tegegne, and Romain Egele. 2023. Hhd-ethiopic a historical handwritten dataset for ethiopic ocr with baseline models and human-level performance.
- CrossLang (Belgium). 2024a. [Manually validated pagexml files for images in "dagboek ernst clarysse"](#).
- CrossLang (Belgium). 2024b. [Manually validated pagexml files for images in "diario del soldato bruno celestino"](#).
- CrossLang (Belgium). 2024c. [Manually validated pagexml files for images in monography "mémoire sur st domingue par h ? m. michel"](#).
- Michael Berger, Henrike Bolte, Veronika Führer, Felix Hausleitner, Sarah Hutterer, Tim Lüthi, Michaela Nancu, Erica Passoni, Katalin Pataki, Sophie Schröcksnadel, Giovanni Verri, Dennis Wegener, and Sandra Hofert. 2022. [Klosterneuburg, stiftsbibl., cod. 48 - ground truth: Initial release](#).
- Yngvil Beyer and Per Erik Solberg. 2023. [Norhand v3 / dataset for handwritten text recognition in norwegian](#).
- Sébastien Biay, Victor Bobby, Kristina Konstantinova, and Zoé Cappe. 2022. [tnah-2021-decameronfr](#).
- Bibliothèque Interuniversitaire de la Sorbonne. 2024. [Nubis-ocr: Ground truth dataset for printed books from the nubis digital library](#). Ground truth dataset of 57 pages from 19 French and Latin books (16021989), transcribed using eScriptorium with the CATMuS Print [Large] model and manually corrected. Includes ALTO files, images, OCR model, and transcribed texts.
- S.S. Blair. 2020. [Islamic Calligraphy](#). Edinburgh University Press.
- Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2024. [Nougat: Neural optical understanding for academic documents](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Matthias Boenig. 2024. [gt_structure_text](#).
- Callum William Booth, Alan Thomas, and Robert Gaizauskas. 2024. [BLN600: A parallel corpus of machine/human transcribed nineteenth century newspaper texts](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2440–2446, Torino, Italia. ELRA and ICCL.
- Julie Bordier, Matthias Gille Levenson, Olivier Brisville-Fertin, Thibault Clérice, and Alix Chagué. 2023. [Htromance, medieval spain corpus of ground-truth for handwritten text recognition and layout segmentation](#). If you use this dataset, please cite it using the metadata from this file.
- Andrei Z Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.
- Arnaldo Junior Candido and Sandra Maria Aluísio. 2009. [Building a corpus-based historical Portuguese dictionary : Challenges and opportunities](#). In *Traitement Automatique des Langues, Volume 50, Numéro 2 : Langues anciennes [Ancient Languages]*, pages 73–102, France. ATALA (Association pour le Traitement Automatique des Langues).
- Constance Carta, Pauline Jacsont, and Élina Leblanc. 2022. [Fondue spanish chapbooks 19th c. dataset](#).

- Silvia Cascianelli, Marcella Cornia, Lorenzo Baraldi, Maria Ludovica Piazzzi, Rosiana Schiuma, and Rita Cucchiara. 2021. Learning to read linfito: handwritten text recognition with synthetic training data. In *Computer Analysis of Images and Patterns: 19th International Conference, CAIP 2021, Virtual Event, September 28–30, 2021, Proceedings, Part II 19*, pages 340–350. Springer.
- Lien Ceard, Fanny Lebreton, and Cécile Sajdak. 2022. [Projet correspondance berlioz](#). Created as part of a transcription project of the active correspondence from Hector Berlioz to his sister Nanci Berlioz, by students of the TNAH master’s program at the École nationale des chartes.
- Center for Open Data in the Humanities. 2019. [Japanese classics dataset \(collection of the national museum of japanese literature, etc.\)](#).
- Center for Open Data in the Humanities (CODH) and National Institute for Japanese Language and Linguistics (NINJAL). 2025. [Modern magazine ocr training dataset](#).
- Alix Chagué, Kévin Champougny, Nina Meissel, Jean-Damien Genero, Eden Skilbeck-Gaborit, Laurie Vanneau, Laura Bey, Victoria Le Fournier, Anaïs Albert, Charles Riondet, and Manuela Martini. 2023a. Time Us Corpus. Available at <https://github.com/HTR-United/timeuscorpus>.
- Alix Chagué, Julie Cissé, and Radia Kichou. [Gt celestine doniau-danest](#).
- Alix Chagué, Thibault Clérice, Anaïs Mazoue, and Elsa Van Kote. 2023b. [Cremma-an-testamentdepoilus](#). Available at <https://github.com/HTR-United/CREMMA-AN-T-estamentDePoilus>.
- Alix Chagué. 2021. [Tapuscorpus](#). If you use this software, please cite it as above.
- Alix Chagué and Gilles Pérez. 2023. [Peraire ground truth](#).
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. [Intern VL: scaling up vision foundation models and aligning for generic visual-linguistic tasks](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16–22, 2024*, pages 24185–24198. IEEE.
- Liang Cheng, Jonas Frankemölle, Adam Axelsson, and Ekta Vats. 2024. [Uncovering the handwritten text in the margins: End-to-end handwritten text detection and recognition](#). In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 111–120, St. Julians, Malta. Association for Computational Linguistics.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. [Chatbot arena: An open platform for evaluating llms by human preference](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21–27, 2024*. OpenReview.net.
- Floriane Chiffolleau. 2021. [dahncorpus](#).
- Kelly Christensen, Baudoin Davoury, Anahi Haedo, Paul Kervegan, and Esteban Sanchez-Oeconomo. 2022. [Projet exposition universelle de 1878](#). Created as part of a transcription project on the 1878 International Congress of Ethnographic Sciences held during the Exposition Universelle in Paris, by students of the TNAH master’s program at the École nationale des chartes.
- Victor Claass, Justine Gain, and Suzanne Martin-Vigier. 2021. [Les papiers barye transcriptions brutes en format alto](#). Transcriptions created with Transkribus in 2020/2021 from the Antoine-Louis Barye archival collection (1795/1875), held at the Bibliothèque de l’INHA, Jacques Doucet collections. The collection is catalogued in Calames.
- Christian Clausner, Apostolos Antonacopoulos, Tom Derrick, and Stefan Pletschacher. 2017. [Icdar2017 competition on recognition of early indian printed documents - reid2017](#). In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 1411–1416.
- Christian Clausner, Apostolos Antonacopoulos, Tom Derrick, and Stefan Pletschacher. 2019a. [Icdar2019 competition on recognition of early indian printed documents-reid2019](#). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1527–1532. IEEE.
- Christian Clausner, Apostolos Antonacopoulos, Nora McGregor, and Daniel Wilson-Nunn. 2018. [Icfhr 2018 competition on recognition of historical arabic scientific manuscripts rasm2018](#). *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 471–476.
- Christian Clausner, Stefan Pletschacher, and Apostolos Antonacopoulos. 2015. [Prima europeana newspaper dataset](#). <https://www.primaresearch.org/datasets>. Accessed: 2025-03-25.
- Christos Clausner, Stefan Pletschacher, and Apostolos Antonacopoulos. 2019b. [RASM2019 dataset: Historical arabic scientific manuscripts](#). <https://www.primaresearch.org/RASM2019/resources>. Accessed: 2025-03-24.
- Julie Claustre, Darwin Smith, Sergio Torres Aguilar, Isabelle Brethauer, Pierre Brochard, Olivier Can-teaut, Emilie Cottureau, Fabrice Delivré, Mathilde Denglos, Vincent Jolivet, Véronique Julerot, Thierry Kouamé, Elisabeth Lusset, Anne Massoni, Sebastien

- Nadiras, Nicolas Perreaux, Hugo Regazzi, and Mathilde Treglia. 2023. *The e-ndp project : collaborative digital edition of the chapter registers of notre-dame of paris (1326-1504). ground-truth for handwriting text recognition (htr) on late medieval manuscripts*.
- Thibault Cl rice, Alix Chagu , Margaux Faure, Jade Norindr, Ana s Mazoue, and Baudoin Davoury. 2022a. *Cremma manuscrits du 17e*. Dataset published by HTR United. Edited by Alix Chagu  and Thibault Cl rice. <https://github.com/HTR-United/CREMMA-MSS-17>.
- Thibault Cl rice, Alix Chagu , and Malamatenia Vlachou-Efstathiou. 2022b. *Cremma medii aevi*. <https://github.com/HTR-United/CREMMA-Medieval-LAT>. DOI: 10.5281/zenodo.7013436.
- Thibault Cl rice. 2021. *Cremma early modern books*.
- Thibault Cl rice and Alix Chagu . 2023. *Cremma manuscrits du 20e*.
- Thibault Cl rice, Alix Chagu , Baudouin Davoury, Soline Doat, Margaux Faure, and Maxime Humeau. 2023. *Cremma manuscrits du 19e*.
- Thomas Constum, Nicolas Kempf, Thierry Paquet, Pierrick Tranouez, Cl ment Chatelain, Sandra Bree, and Fran ois Merveille. 2022. *Popp datasets: datasets for handwriting recognition from french population census*.
- Valentin De Craene, rayondemi l, Thibault Cl rice, and Virgile Reignier. 2022. *psl-chartes-htr-students/tnah-2021-argusdesbrevets: 1.0*.
- Pascale Cugy, Caroline Fieschi, Alix Peyrard, Lucie Prohin, Marie-Anne Sarda, Institut National de l’histoire de l’art (INHA), and Biblioth que nationale de France. 2022. *La correspondance jacques doucet - ren  jean*. <https://gitlab.inha.fr/snr/LaCorrespondanceDoucetReneJean>. Project website: <https://skylab.inha.fr/PENSE/LettresDeJacquesDoucetAReneJean1908-1929/>. Letters and documents (1908–1929) conserved at the Biblioth que nationale de France, D partement des manuscrits, NAF 13124.
- Hoang-Quan Dang, Duy-Anh Nguyen, Phu-Phuoc Pham, Ngoc-Th nh Nguyen, Tan Chau, Duc-Vu Ngo, Trung-Hieu Nguyen, Chau-Thang Phan, The-Hien Trinh, Minh-Tri Nguyen, and 1 others. 2022. *Nomnaocr: The first dataset for optical character recognition on han-nom script*. In *2022 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 476–481. IEEE.
- Andreas Wagner David Gl ck, Cindy Rico Carmona. 2018. *Edition guidelines*. <https://www.salamancaschool/guidelines.html>.
- Melissa Dell, Jacob Carlson, Tom Bryan, Emily Silcock, Abhishek Arora, Zejiang Shen, Luca D’Amico-Wong, Quan Le, Pablo Querubin, and Leander Heldring. 2023. *American stories: A large-scale structured text dataset of historical U.S. newspapers*. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Deutsches Textarchiv. *Grundlage f r ein referenzkorpus der neuhochdeutschen sprache*.
- Thao Do, Dinh Phu Tran, An Vo, and Daeyoung Kim. 2024. *Reference-based post-ocr processing with llm for precise diacritic text in historical document recognition*. *Preprint*, arXiv:2410.13305.
- Soline Doat, Elsa Falcoz, Margaux Faure, Ana s Mazou , and Ariane Menu. 2022. *Projet notre-dame*.
- Hans J.G.A. Dolfing, Jerome Bellegarda, Jan Chorowski, Ricard Marxer, and Antoine Laurent. 2020. *The “ScribbleLens” Dutch historical handwriting corpus*. In *International Conference on Frontiers of Handwriting Recognition (ICFHR)*, page To Appear. <http://www.openslr.org/84/>.
- Senka Drobac, Pekka Kauppinen, and Krister Lind n. 2017. *OCR and post-correction of historical Finnish texts*. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 70–76, Gothenburg, Sweden. Association for Computational Linguistics.
- Alain Dubois, Thibault Cl rice, Delphine Mamie, Darius Schlaeppli, Cl mence Rudaz, and Marie-Caroline Schmied. 2024. *Tables du recensement du valais*.
- Malamatenia Vlachou Efstathiou, No  Leroy, and Marco Maulu. 2022. *git-project-boccace*.
- Nicole Eichenberger and Hedwig Suwelack. 2021. *Faithful transcriptions data set: Tei/xml-encoded transcriptions of medieval theological manuscripts*.
- David Fern ndez-Mota, Joan Almaz n, N ria Cirera, Alicia Forn s, and Josep Llad s. 2014. *Bh2m: The barcelona historical handwritten marriages database*. In *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR)*, pages 256–261.
- Andreas Fischer, Valentin Frinken, and Horst Bunke. 2009. *Parzival database*. <https://fki.tic.heia-fr.ch/databases/parzival-database>. Accessed: 2025-03-25.
- Andreas Fischer, Alex Keller, Valentin Frinken, and Horst Bunke. 2012. *Saint gall database*. <https://fki.tic.heia-fr.ch/databases/saint-gall-1-database>. Accessed: 2025-03-25.
- Marc Frincu, Simina Frincu, and Marius E. Penteliuc. 2023. *Challenges and solutions in transliterating 19th century Romanian texts from the transitional to the Latin script*. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 226–231, Vienna, Austria. NOVA CLUNL, Portugal.

- Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Mingxin Huang, and 1 others. 2025. [Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning](#). *ArXiv preprint*, abs/2501.00321.
- Simon Gabay. 2024. [Fondue-fr-print-16](#).
- Simon Gabay and Carmen Carrasco Luján. 2024. [Fondue-es-print-19](#).
- Simon Gabay and Sophie Dolto. 2024. [Fondue-fr-print-20](#).
- Simon Gabay, Peter Nahon, Marco Cicchini, Yvan Jau-reguy, and Loraine Chappuis. 2023a. [Fondue-fr-mss-18](#).
- Simon Gabay, Myriam Perregaux, and Jessica Da Silva Fernandes. 2023b. [Fondue-en-print-20](#).
- Simon Gabay, Myriam Perregaux, and Jessica Da Silva Fernandes. 2023c. [Fondue-en-print-20](#).
- Simon Gabay, Ariane Pinche, Eliott Fabert, and Kelly Christensen. 2023d. Données imprimés du 18e siècle. Dataset published by HTR United. Edited by Alix Chagué and Thibault Clérice. Available at <https://github.com/Gallicorpora/HTR-imprime-18e-siecle>.
- Simon Gabay, Ariane Pinche, Eliott Fabert, malamatenia Vlachou-Efstathiou, and Kelly Christensen. 2023e. Imprimés 17e siècle. HTR United (publisher). Edited by Alix Chagué and Thibault Clérice. Dataset. Available at <https://github.com/Gallicorpora/HTR-imprime-17e-siecle>.
- Simon Gabay, Ariane Pinche, Noé Leroy, and Kelly Christensen. 2023f. Données htr incunables du 15e siècle. <https://github.com/Gallicorpora/HTR-incunable-15e-siecle>. Edited by Alix Chagué and Thibault Clérice. Dataset. Publisher: HTR United.
- Simon Gabay, Ariane Pinche, Malamatenia Vlachou-Efstathiou, and Kelly Christensen. 2022. [Données htr imprimés du 16e siècle](#).
- Gemini Team Google. 2023. [Gemini: A family of highly capable multimodal models](#). *ArXiv preprint*, abs/2312.11805.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. [Gemma: Open models based on gemini research and technology](#). *ArXiv preprint*, abs/2403.08295.
- Kioumars Ghahreghlou, editor. 2018. *Persian Text*, volume 98, pages 1–372. American Oriental Society. Accessed 20 May 2025.
- Anthony Glaise, Thibault Clérice, Federico Boschetti, Franz Fischer, and Alix Chagué. 2023. [Htromance, medieval latin corpus of ground-truth for handwritten text recognition and layout segmentation](#).
- Tobias Grüning, Gundram Leifert, Johannes Michael, Tobias Strauß, Max Weidemann, and Roger Labahn. 2016. [read_dataset_german_konzilsprotokolle](#).
- Estelle Guéville and David Joseph Wrisley. 2022. [Transcribing medieval manuscripts for machine learning](#). *ArXiv preprint*, abs/2207.07726.
- Ingrid Guimarães, Perrine Maurel, Yagmur Ozturk, and Alix Chagué. 2022. [Memorials for jane lathrop stanford](#).
- Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, Jingji Chen, Jingjia Huang, Kang Lei, Liping Yuan, Lishu Luo, Pengfei Liu, Qinghao Ye, Rui Qian, Shen Yan, and 178 others. 2025. [Seed1.5-vl technical report](#).
- Estelle Guéville and David Joseph Wrisley. 2021. [Ground truth used in htr for the paris bible project](#). Released on 2021-10-10.
- Colette Halter-Pernet, Simon Teuscher, Tobias Hodel, Lukas Barwitzki, Salome Egloff, Fabian Henggeler, Michael Nadig, Anina Steinmann, Sabine Stettler, and Ismail Prada Ziegler. 2021. [Charters and records of königsfelden abbey and bailiwick \(1308-1662\)](#).
- Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. [Glottolog 5.1](#). <https://glottolog.org/>.
- Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 991–995. IEEE.
- Tobias Hodel, David Schoch, and Peter Dängeli. 2021. [Handwritten text recognition ground truth set: Stabs ratsbücher o10, urfedenbuch x](#).
- Laurie Hoeben. 2024. [Antoine verard extracts](#). Transcriptions of parts of Antoine Vérards editions principes of "Tristan", "Merlin" and "Gyron le Courtoys".
- Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. [Vision-r1: Incentivizing reasoning capability in multimodal large language models](#). *Preprint*, arXiv:2503.06749.
- Marco Humbel, Andreas Vlachidis, Julianne Nyhan, and The Trustees of the British Museum. 2024. Sloane lab htr model. Handwritten Text Recognition training data (layout segmentation and transcriptions) for the Sloane Lab HTR model.

- International Organization for Standardization. 2022. Information and documentation codes for the representation of names of scripts. <https://iso639-3.sil.org/>. Accessed: 2025-03-25.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. *Openai o1 system card*. *ArXiv preprint*, abs/2412.16720.
- Claire Jahan and Simon Gabay. 2021. *OCR17 +*.
- Béatrice Joyeux-Prunel, Simon Gabay, Martina Rizzello, Valéry Berlincourt, Elena Maria Rizzi, Stefania Tesser, Victoria Bukvic, Jaime Diaz, Guillaume Aebi, and Raoul Bickel. 2023. *Fondue-mlt-art*.
- Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. *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 04, pages 19–24.
- Jan Kamlah, Thomas Schmidt, Renat Shigapov, and Stefan Weil. 2024. *reichsanzeiger-gt*.
- Romain Karpinski, Devashish Lohani, and Abdel Belaid. 2018. *Metrics for complete evaluation of ocr performance*. In *Proceedings of the 22nd International Conference on Image Processing, Computer Vision, and Pattern Recognition (ICPV)*, Las Vegas, United States. HAL Id: hal-01981731.
- Majeed Kassis, Alaa Abdalhaleem, Ahmad Droby, Reem Alaasam, and Jihad El-Sana. 2017. *Vml-hd: The historical arabic documents dataset for recognition systems*. In *1st International Workshop on Arabic Script Analysis and Recognition*. IEEE.
- Duan Katuák, Lucia Niníková, Michaela Mikuková, Nikola Halfarová, Terezie Gajdoová, Lenka Málková, Nikol Taufrová, Imrich Nagy, Klára Kováková-Pohlová, Matej mida, and Nela Kociánová. 2024. *Slovenský supermodel p&t1 (sspt1) : Matej bel university skriptor project datasets*.
- Koray Kavukcuoglu. 2025. Gemini 2.5: Our most intelligent ai model. <https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-thinking>. Accessed: 2025-05-18.
- Liesbeth Keijser. 2024. *6000 ground truth of voc and notarial deeds 3.000.000 htr of voc, wic and notarial deeds*.
- Simon Carlos Kemper. 2021. *Gado2: multilingual newspapers from the netherlands indies*.
- Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. *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.
- Andreas Kölsch. 2025. Handwritten annotation detection dataset (annotationdb). https://tc11.cvc.uab.es/datasets/AnnotationDB_1. ID: AnnotationDB_1, Accessed: 2025-05-18.
- Badische Landesbibliothek, Katharina Ost, Annika Stello, and Gerrit Heim. 2024. *Training data in-cunabula reichenau*.
- David Lassner, Julius Coburger, Clemens Neudecker, and Anne Baillot. 2021. *Data set of the paper "publishing an ocr ground truth data set for reuse in an unclear copyright setting"*.
- LECTAUREP, Aurélia Rostaing, Marc Durand, and Alix Chagué. 2021. *Notaires de paris répertoires: Ground truth for various parisian registries of notary deeds (french 19th and 20th centuries)*. Released on 2021-12-07.
- Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvasi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. *Pix2struct: Screenshot parsing as pretraining for visual language understanding*. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 18893–18912. PMLR.
- Lith Lefranc. 2024. *Arletta: open-source handwritten text recognition models for historic dutch*.
- Vladimir I. Levenshtein. 1965. *Binary codes capable of correcting deletions, insertions, and reversals. Soviet physics. Doklady*, 10:707–710.
- Matthias Gille Levenson. 2023. *Towards a general open dataset and model for late medieval castilian text recognition (htr/ocr)*.
- David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. 2006. *Building a test collection for complex document information processing*. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 665–666.
- Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. 2020. *DocBank: A benchmark dataset for document layout analysis*. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 949–960, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Library of Congress. 2025. *ALTO: Analyzed Layout and Text Object*. <https://www.loc.gov/standards/alto/>. Accessed: 2025-3-22.
- Bernhard Liebl and Manuel Burghardt. 2020. *On the accuracy of crnns for line-based ocr: A multi-parameter evaluation*.

- Marie-Françoise Limon-Bonnet, Aurélie Rostaing, and Alix Chagué. 2024. [lectaurep-bronod, ground truth for maitre bronod's documents \(french xviiiith century\)](#).
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. [Llava-next: Improved reasoning, ocr, and world knowledge](#).
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. [Visual instruction tuning](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023b. [Mmbench: Is your multi-modal model an all-around player?](#) *Preprint*, arXiv:2307.06281.
- Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024b. [Ocr-bench: on the hidden mystery of ocr in large multi-modal models](#). *Science China Information Sciences*, 67(12):220102.
- Zoé LEveque, Kate Ekaterina, and Anahide Kasparian. 2022. [Projet kovalswky 1893](#).
- Ayush Maheshwari, Nikhil Singh, Amrith Krishna, and Ganesh Ramakrishnan. 2022a. [A benchmark and dataset for post-OCR text correction in Sanskrit](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6258–6265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Ayush Maheshwari, Nikhil Singh, Amrith Krishna, and Ganesh Ramakrishnan. 2022b. [A benchmark and dataset for post-OCR text correction in Sanskrit](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6258–6265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Souvik Mandal, Ashish Talewar, Paras Ahuja, and Prathamesh Juvatkar. 2025. [Nanonets-ocr-s: A model for transforming documents into structured markdown with intelligent content recognition and semantic tagging](#).
- Isabelle Marthot-Santaniello and Tobias Hodel. 2022. [Ground-truthed data set of zenon papyri for handwritten text recognition](#).
- Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. [Docvqa: A dataset for vqa on document images](#). In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.
- Guénette Maxime, Verstraete Mathilde, Chagué Alix, and Vitali-Rosati Marcello. 2024. [Ground truth for the palatine anthology](#).
- Anaïs Mazoue, Thibault Clérice, and Alix Chagué. 2024. [Cremma-mss-16](#).
- Nicole Merkel-Hilf. 2022. [Ground truth data for printed devanagari](#).
- Anna Michalcová, Kamil Bazelides, Jan Haji, Erika Pnkavová, Laura Maniaková, Hana Kreisingerová, Jitka Filipová, Chi hung Lu, and Martina Dvoáková. 2022. [Padeov-bible-handwriting-ground-truth: Initial release](#).
- Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. [Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras](#). *Preprint*, arXiv:2503.01743.
- Microsoft. 2024. [Azure AI Document Intelligence](#). <https://learn.microsoft.com/en-us/azure/ai-services/document-intelligence/>. Accessed: 2025-03-20.
- Mistral AI. 2025. [Mistral OCR](#). <https://mistral.ai/news/mistral-ocr>. Accessed: 2025-03-24.
- Guenter Muehlberger and Guenter Hackl. 2019. [News-eye / read ocr training dataset from austrian newspapers \(19th c.\)](#).
- Guenter Muehlberger and Guenter Hackl. 2020. [News-eye / read ocr training dataset from french newspapers \(18th, 19th, early 20th c.\)](#).
- Guenter Muehlberger and Guenter Hackl. 2021a. [Newseye / read ocr training dataset from finnish newspapers \(18th, 19th, early 20th c.\)](#).
- Guenter Muehlberger and Guenter Hackl. 2021b. [Newseye / read ocr training dataset from swedish newspapers \(18th, 19th, early 20th c.\)](#).
- Rayyan Najam and Safiullah Faizullah. 2024. [A scarce dataset for ancient arabic handwritten text recognition](#). Mendeley Data, V1. Accessed: 2025-03-25.
- Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. 2021. [A survey of ocr evaluation tools and metrics](#). In *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing, HIP '21*, page 1318, New York, NY, USA. Association for Computing Machinery.
- Lucas Noémie, Clément Salah, and Chahan Vidal-Gorène. 2022. [New Results for the Text Recognition of Arabic Maghribi Manuscripts - Managing an Under-resourced Script](#). Working paper or preprint.

- Jade Norindr, Anna Mikhalechuk, Thibault Clérice, and Alix Chagué. 2023. [Htromance, modern language corpus of ground-truth for handwritten text recognition and layout segmentation](#).
- Vít Novotný, Kristýna Seidlová, Tereza Vrabcová, and Ale Horák. 2021. [When tesseract brings friends: Layout analysis, language identification, and super-resolution in the optical character recognition of medieval texts](#). In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2021*, pages 91–100. Tribun EU.
- Numind. 2025. [NuMarkdown-8B-Thinking](#). Model page.
- NVIDIA. 2025. [Llama-3.1-Nemotron-Nano-VL-8B-V1](#). Model page.
- Sarah Oberbichler, Emanuela Boroş, Antoine Doucet, Jani Marjanen, Eva Pfanzelter, Juha Rautiainen, Hannu Toivonen, and Mikko Tolonen. 2022. Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians. *Journal of the Association for Information Science and Technology*, 73(2):225–239.
- OpenAI. 2025a. [Introducing gpt-5](#). Accessed: 2025-09-18.
- OpenAI. 2025b. [Introducing openai o3 and o4-mini](#). Accessed: 2025-09-18.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2023. [Gpt-4 technical report](#).
- OpenAI, Ananya Kumar, Jiahui Yu, John Hallman, Michelle Pokrass, Adam Goucher, Adi Ganesh, Bowen Cheng, Brandon McKinzie, Brian Zhang, Chris Koch, Colin Wei, David Medina, Edmund Wong, Erin Kavanaugh, Florent Bekerman, Haitang Hu, Hongyu Ren, Ishaan Singal, and 25 others. 2025. [Introducing gpt-4.1 in the api](#). Accessed: 2025-05-18.
- Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, and 1 others. 2024. [Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations](#). *ArXiv preprint*, abs/2412.07626.
- Alexander James O'Neill and Nathan Hill. 2022. [Text recognition for nepalese manuscripts in pracalit script](#). *Journal of Open Humanities Data*.
- Werner Pantke, Martin Dennhardt, Daniel Fecker, Volker Märgner, and Tim Fingscheidt. 2014. [An historical handwritten arabic dataset for segmentation-free word spotting - hadara80p](#). In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 15–20.
- Christos Papadopoulos, Stefan Pletschacher, Christian Clausner, and Apostolos Antonacopoulos. 2013. [The impact dataset of historical document images](#). In *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing, HIP '13*, page 123130, New York, NY, USA. Association for Computing Machinery.
- Aleksandros Papazoglou, Ioannis Pratikakis, Kleopatra Markou, and Lazaros Tsochatzidis. 2020. [Eparchos historical greek handwritten document dataset](#).
- Margot Pascual, Louis-Fiacre Franchet d'Espèrey, and Simon Gabay. 2022. [Château de chavigny](#).
- Marius Penteliuc. 2022. 19th-century romanian transitional script. <https://www.kaggle.com/datasets/mariuspenteliuc/rts-ocr>. Accessed: 2025-03-24.
- Kim Pham. 2020. [University of denver collections as data - htr train and validation set jcrs_2020_5_27](#).
- Ariane Pinche. 2022. [Cremma medieval](#).
- Ariane Pinche, Simon Gabay, Noé Leroy, and Kelly Christensen. 2022. [Données htr manuscrits du 15e siècle](#).
- Ariane Pinche and Corinne Pierreville. 2023. [Fabiliaux](#).
- Paraskevi Platanou, John Pavlopoulos, and Georgios Papaioannou. 2022. Handwritten paleographic greek text recognition dataset (century-based). This dataset is part of the work: Platanou, P., Pavlopoulos, J., and Papaioannou, G. (2022). Handwritten Paleographic Greek Text Recognition: A Century-Based Approach. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 65856589, Marseille, France: ELRA.
- Stefan Pletschacher and Apostolos Antonacopoulos. 2010. The PAGE (Page Analysis and Ground-Truth Elements) Format Framework. In *2010 20th International Conference on Pattern Recognition*, pages 257–260. IEEE.
- Jake Poznanski, Aman Rangapur, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Aman Rangapur, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. 2025. [olmocr: Unlocking trillions of tokens in pdfs with vision language models](#).
- Frederine Pradier, Simon Gabay, Juliette Janès, Esteban Sánchez Oeonomo, and Paul Kervégan. 2022. [Fondue datasets for historical catalogues](#).
- Ioannis Pratikakis, Aleksandros Papazoglou, Symeon Symeonidis, and Lazaros Tsochatzidis. 2021. [Stavronikita monastery greek handwritten document collection no.114](#).

- PRHLT Research Center. 2022. [Hisclima dataset](#).
- Lorenzo Quirós, Lluís Serrano, Vicente Bosch, Alejandro H. Toselli, Rosa Congost, Enric Saguer, and Enrique Vidal. 2018. [Oficio de hipotecas de girona. a dataset of spanish notarial deeds \(18th century\) for handwritten text recognition and layout analysis of historical documents](#).
- Irina Rabaev, Berat Kurar Bakarar, and Jihad El-Sana. 2019. [The pinkas dataset](#).
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Research Group Texts for the History of Spanish [GITHE]. 2022. CODEA+ 2022 (Corpus of Spanish documents prior to 1900). <https://doi.org/10.37536/CODEA.2015>. [online; accessed 2025-03-09].
- Christian Reul, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttnier, and Frank Puppe. 2019. Ocr4allan open-source tool providing a (semi-) automatic ocr workflow for historical printings. *Applied Sciences*, 9(22):4853.
- Justyna Rogos-Hebda. 2025. Making sense of abbreviation in late middle english literary manuscripts. In *Abbreviating Middle English: Scribal Practices, Visual Texts and Medieval Multimodalities*, pages 157–169. Springer.
- Matteo Romanello, Najem-Meyer Sven, and Bruce Robertson. 2021. [Optical Character Recognition of 19th Century Classical Commentaries: The Current State of Affairs](#). In *The 6th International Workshop on Historical Document Imaging and Processing (HIP '21)*, Lausanne. Association for Computing Machinery.
- C.A. Romein, Michel de Gruijter, and Sara Floor Veldhoen. 2020. The datafication of early modern ordinances. *DH Benelux Journal*, 2.
- Verónica Romero, Alejandro H. Toselli, Enrique Vidal, Joan Andreu Sánchez, Carlos Alonso, and Lourdes Marqués. 2019. [Modern vs diplomatic transcripts for historical handwritten text recognition](#). In *New Trends in Image Analysis and Processing ICIAP 2019: ICIAP International Workshops, BioFor, PatReCH, e-BADLE, DeepRetail, and Industrial Session, Trento, Italy, September 9-10, 2019, Revised Selected Papers*, page 103114, Berlin, Heidelberg. Springer-Verlag.
- Aurélia Rostaing, Nathalie Denis, and Alix Chagué. 2024. [Lectarep-mariages-et-divorces: Ground truth for the enregistrements des contrats de mariages et des séparations et divorces in paris \(french 19th century\)](#). Released on 2024-02-07.
- Sacro Convento di San Francesco in Assisi. 2024. [Assisi, fondo antico del sacro convento mediatheca franciscana](#). Internet Culturale. Central Institute for the Union Catalogue of Italian Libraries (ICCU), Ministry of Culture.
- Mehreen Saeed, Adrian Chan, Anupam Mijar, Joseph Moukarzel, Georges Habchi, Carlos Younes, Amin Elias, Chau-Wai Wong, and Akram Khater. 2024. [Muharaf-public](#).
- Rajkumar Saini, Derek Dobson, Jon Morrey, Marcus Liwicki, and Foteini Simistira Liwicki. 2019. [Icdar 2019 historical document reading challenge on large structured chinese family records](#). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1499–1504.
- Salamanca School Project. 2025. The school of salamanca. <https://www.salamanca.school/works.html>. Accessed: 2025-05-18.
- Vincent Sarbach-Pulicani, Théophile Miaille, Adrien Escoda, Violette Saïag, and Simon Gabay. 2022. [Ocr d’une poésie corse](#).
- John Schaefer and Alexis Litvine. 2023. [Joseph hooker htr model](#).
- Christian Schultze, Niklas Kerkfeld, Kara Kuebart, Princilia Weber, Moritz Wolter, and Felix Selgert. 2024. [Chronicling germany: An annotated historical newspaper dataset](#).
- Yongxin Shi, Chongyu Liu, Dezhi Peng, Cheng Jian, Jiarong Huang, and Lianwen Jin. 2023. [M5hisdoc: A large-scale multi-style chinese historical document analysis benchmark](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Sonia Solfrini, Simon Gabay, Geneviève Gross, Pierre-Olivier Beaulnes, Aurélia Marques Oliveira, Mylène Dejoux, and Daniela Solfaroli Camillocci. 2023. [Données ocr et segmentation des imprimés de jean michel \(projet setaf\)](#). OCR data for the SETAF project: 16th-century French prints in Gothic characters.
- Uwe Springmann and Fabian Fink. 2018. Ground truth (gt) data for fraktur/gothic prints. <https://github.com/ubtue/gt-fraktur>. Accessed: 2025-03-25.
- Peter Stotz and Phillip Ströbel. 2021. [bullinger-digital/gwalther-handwriting-ground-truth: Initial release](#).
- Phillip Ströbel and Simon Clematide. 2019. Improving ocr of black letter in historical newspapers: The unreasonable effectiveness of htr models on low-resolution images. In *Proceedings of the Digital Humanities 2019, (DH2019)*. Accepted.

- Daniel Stökl Ben Ezra, Bronson Brown-DeVost, Pawel Jablonski, Benjamin Kiessling, Elena Lolli, and Hayim Lapin. 2021. [Biblia - an open annotated dataset](#).
- Alaa Sulaiman, Khairuddin Omar, and Mohammad F. Nasrudin. 2019. [Degraded historical document binarization: A review on issues, challenges, techniques, and future directions](#). *Journal of Imaging*, 5(4).
- J.A. Sánchez, A.H. Toselli, V. Romero, and E. Vidal. 2017. [Icdar 2015 competition htrts: Handwritten text recognition on the transcriptorium dataset](#).
- Joan Andreu Sánchez, Verónica Romero, Alejandro H. Toselli, and Enrique Vidal. 2016. [Read dataset bozen](#).
- Granite Vision Team, Leonid Karlinsky, Assaf Arbelle, Abraham Daniels, Ahmed Nassar, Amit Alfassi, Bo Wu, Eli Schwartz, Dhiraj Joshi, Jovana Kondic, Nimrod Shabtay, Pengyuan Li, Roei Herzig, Shafiq Abedin, Shaked Perek, Sivan Harary, Udi Barzelay, Adi Raz Goldfarb, Aude Oliva, and 44 others. 2025. [Granite vision: a lightweight, open-source multimodal model for enterprise intelligence](#). *Preprint*, arXiv:2502.09927.
- Teklia. 2025a. [Finlam la liberté newspaper dataset](#). <https://huggingface.co/datasets/Teklia/Newspapers-finlam-La-Liberte>. Accessed: 2025-05-18.
- Teklia. 2025b. [Finlam newspaper segmentation dataset](#). <https://huggingface.co/datasets/Teklia/Newspapers-finlam>. Accessed: 2025-05-18.
- Tatiana Tommasi. 2024. [episearch-htr: Transcriptions, metadata, and htr models from giovanni antonio astoris correspondence and epigraphic manuscript](#).
- Alejandro H. Toselli and Enrique Vidal. 2021. [The finnish court records dataset](#).
- Michał Turski, Tomasz Stanisławek, Karol Kaczmarek, Paweł Dyda, and Filip Graliński. 2023. [Ccpdf: Building a high quality corpus for visually rich documents from web crawl data](#). In *International Conference on Document Analysis and Recognition*, pages 348–365. Springer.
- Tübingen University Library. 2023. [Ground truth data for printed malayalam](#).
- Dona Valy, Michel Verleysen, Sophea Chhun, and Jean-Christophe Burie. 2017. [A new khmer palm leaf manuscript dataset for document analysis and recognition: Sleukrith set](#). In *Proceedings of the 4th International Workshop on Historical Document Imaging and Processing*, HIP '17, page 16, New York, NY, USA. Association for Computing Machinery.
- Elsa Van Kote, Margaux Faure, Jade Norindr, Thibault Clérice, and Alix Chagué. 2024. [Cremma-mss-18](#).
- Caroline Vandyck, Jessie Dummer, Emmanuelle Kuhry, Zdzisław Koczarski, Sylvain Besson, Caroline Chevalier-Royet, and Matthias Gille Levenson. 2024. [Transcriboquest 2024 medieval literary](#).
- Chahan Vidal-Gorène, Noémie Lucas, Clément Salah, Aliénor Decours-Perez, and Boris Dupin. 2021. [Rasam – a dataset for the recognition and analysis of scripts in arabic maghrebi](#). In *Document Analysis and Recognition – ICDAR 2021 Workshops*, pages 265–281, Cham. Springer International Publishing.
- Malamatenia Vlachou-Efstathiou. 2024. [Eutyches "de uerbo" glossed](#). If you use this software, please cite it using the metadata from this file.
- De Digitale Bibliotheek voor de Nederlandse Letteren (DBNL). 2019. [Dbnl ocr data set](#).
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. [Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution](#). *ArXiv preprint*, abs/2409.12191.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025. [Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency](#).
- Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024b. [Charxiv: Charting gaps in realistic chart understanding in multimodal llms](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Maurice Weber, Carlo Siebenschuh, Rory Butler, Anton Alexandrov, Valdemar Thanner, Georgios Tso-lakis, Haris Jabbar, Ian T. Foster, Bo Li, Rick Stevens, and Ce Zhang. 2023. [Wordscape: a pipeline to extract multilingual, visually rich documents with layout annotations from web crawl data](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- K. S. Whetter. 2017. *The Manuscript and Meaning of Malory's Morte Darthur: Rubrication, Commemoration, Memorialization*. Arthurian Studies. Boydell and Brewer.
- Nick White, Thibault Clérice, and Antonia Karaisl. 2022. [Caroline minuscule by rescribe](#). Project website: <https://rescribe.xyz/>.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. [Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding](#).

M.D. Wyatt and W.R. Tymms. 1861. *The History, Theory, and Practice of Illuminating*. Day and son.

LLM-Core-Team Xiaomi. 2025. [Mimo-vl technical report](#).

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024a. [Qwen2.5 technical report](#). *ArXiv preprint*, abs/2412.15115.

Hailin Yang, Lianwen Jin, Weiguo Huang, Zhaoyang Yang, Songxuan Lai, and Jifeng Sun. 2018. [Dense and tight detection of chinese characters in historical documents: Datasets and a recognition guided detector](#). *IEEE Access*, 6:30174–30183.

Qi Yang, Bolin Ni, Shiming Xiang, Han Hu, Houwen Peng, and Jie Jiang. 2025. [R-4b: Incentivizing general-purpose auto-thinking capability in mllms via bi-mode annealing and reinforce learning](#).

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Yuliang Liu, and 1 others. 2024b. [Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy](#). *ArXiv preprint*, abs/2412.02210.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, and 3 others. 2024. [Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Taha Zerrouki. 2010. [pyarabic, an arabic language library for python](#).

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. [Sigmoid loss for language image pre-training](#). In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11941–11952. IEEE.

Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, and 1 others. 2023. [Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition](#). *ArXiv preprint*, abs/2309.15112.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 1015–1022. IEEE.

A CHURRO-DS Details

A.1 Dataset Statistics

Figures 3 and 4 show the distribution of the length of the gold text labels, and the dimension of images in CHURRO-DS respectively.

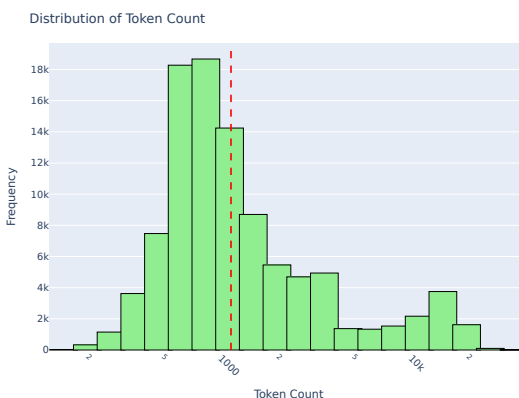


Figure 3: Distribution of Qwen 2.5 VL tokens in the gold text. Log-scale x-axis. Red dashed line shows the median.

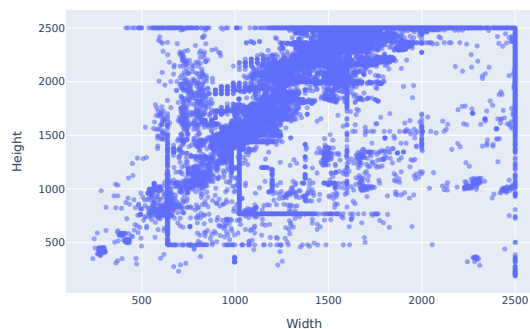


Figure 4: Plot of image dimensions in CHURRO-DS. Each point shows the width and height of a single image. Note that we resize large images to fit inside a 2500×2500 pixel box while maintaining their original aspect ratio.

A.2 Language and Script Statistics

Tables 3 and 4 show the number of examples per language and script, in CHURRO-DS.

B System Details

B.1 Model IDs

Table 5 lists the exact model IDs used in our experiments. For closed models, the IDs follow the providers’ APIs: Google Vertex AI and Azure OpenAI. For open-weight models, the IDs are taken from the HuggingFace hub. Throughout this paper, we refer to models by their total parameter counts to clarify the model sizes used. In cases where the official model ID differs (e.g., olmOCR), we use the total parameter count for consistency.

B.2 Hyperparameters

For all VLMs, we generate outputs with a temperature of 0 (i.e., greedy decoding). All results are reported from a single run.

We set the maximum number of generated tokens to 20,000 for non-reasoning models and 40,000 for reasoning models to accommodate the additional reasoning tokens they generate. The value 20,000 is chosen based on Figure 3 to allow for the generation of all gold outputs. Reasoning parameters for reasoning models are included in Table 5.

To create CHURRO, we fine-tuned Qwen 2.5 VL (3B) for 5 epochs on 32 NVIDIA H100 GPUs. Using gradient accumulation and a total effective batch size of 128, fine-tuning took approximately 25 hours. We used a learning rate of 5×10^{-5} with a cosine schedule. Overall, the computation for this paper, including fine-tuning and inference for all open-weight models, totaled about 6,000 GPU-hours on NVIDIA H100 GPUs.

The tokenizer of Qwen 2.5 VL outputs a variable number of tokens depending on the resolution of the input image. We resized all input images to a maximum of 5,120 image patches, each of size 28×28 pixels, for all fine-tuning experiments. This limits the total number of image and text tokens per example to under 25,000.

B.3 Prompt Used for Zero-Shot Evaluation of VLMs

The prompt used to evaluate zero-shot VLMs is shown in Table 6.

C Additional Experiments

Figure 5 plots normalized Levenshtein similarity against the number of tokens in the gold transcription for each page. We observe that, in general,

model accuracy decreases as the amount of text on the page increases.

D Example System Outputs

We present additional examples illustrating common error categories, with the correct gold output shown side by side with the outputs of the zero-shot Qwen 2.5 VL (3B) model and its fine-tuned counterpart, CHURRO.

D.1 Reading Order-Related Errors in Vertical Script Layouts

Figure 6 presents a Classical Chinese document that follows traditional layout conventions, where text is arranged vertically from top to bottom and columns are ordered from right to left.

The fine-tuned CHURRO model produces a transcription that closely matches the ground truth. Most discrepancies are minor, such as normalization of historical glyph variants or the substitution of rare characters not fully represented in modern Unicode. Importantly, the reading order is preserved, and the model correctly segments lines and parses content without inserting hallucinated tokens.

In contrast, the zero-shot VLM fails to generate a coherent transcription. Its output contains hallucinated markup symbols, fragmented character groupings, and widespread omissions or rearrangements. We attribute this to a fundamental misalignment in layout assumptions: the model appears to interpret the document as a modern layout with left-to-right, top-to-bottom reading order, rather than adhering to the traditional vertical, right-to-left format. This results in substantial degradation in both recognition accuracy and semantic structure.

³The English translation of the gold text is: “In the eighth year of the Guimao cycle, the 10th year of the Song dynasty Chunxi reign, in the first month of spring, Wu Lüxin was appointed as general to lead a campaign against Ailao. In the eleventh month, on the day of Renwu, a solar eclipse occurred. In the ninth year of Jiachen, the 11th year of Chunxi, in the third month of spring, Champa sent tribute, including three items from Siam. Merchants from Qi and neighboring countries came to Yuntun Garrison to offer goods and request permission to trade. In the tenth month, the Meng people and others rebelled and fought among themselves. In the 10th year of Yisi, the 12th year of Chunxi, in the first month of spring, civil service examinations were held for scholars aged fifteen and above. Those proficient in poetry and classical texts were admitted to study at the imperial court; thirty candidates, including Pei Guokai and Deng Yan, were selected. The rest remained in school. In the seventh month of autumn, the King of Jiankang, Long Yi, was ordered to lead an expedition against the southern tribes, which were subsequently

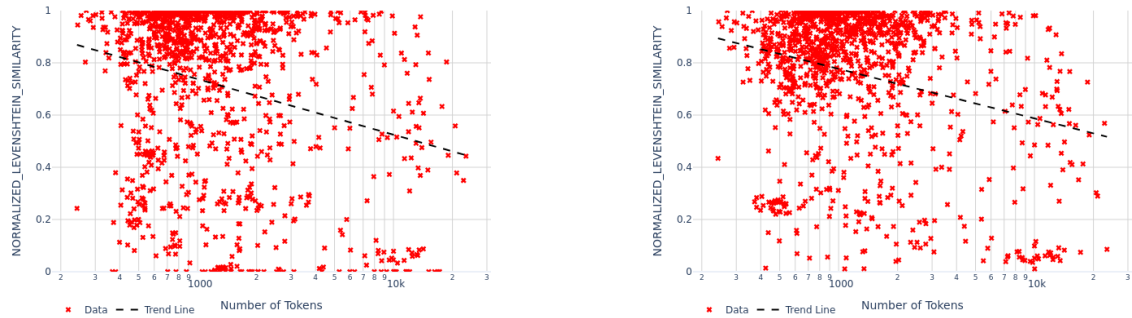


Figure 5: The effect of document length on normalized Levenshtein similarity. Here we show this effect for Gemini 2.5 Pro on the left and CHURRO on the right.

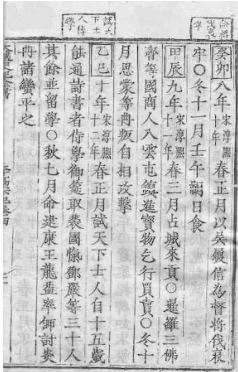
Document image	Gold	Finetuned	Zero-shot LMM
	癸卯八年 宋淳熙 十年 春正月以吳履信為督將伐袁 年冬十一月壬午朔日食 甲辰九年 宋淳熙 十一年 春三月占城來貢暹羅三佛 齊等國商人入雲屯鎮進寶物乞行買賣冬十 月思蒙等叛自相攻擊 乙巳十年 宋淳熙 十二年 春正月試天下士人自十五歲 能通詩書者侍學御筵取裴國慎鄧嚴等三十人 其餘並留學秋七月命建康王龍益率師討 諸蠻平之	癸卯八年 宋淳熙 十年 春正月以吳履信為督將伐袁 年冬十一月壬午朔日食 甲辰九年 宋淳熙 十一年 春三月占城來貢暹羅三佛 齊等國商人入雲屯鎮進寶物乞行買賣冬十 月思蒙等 叛自相攻擊 乙巳十年 宋淳熙 十二年 春正月試天下士人自十五歲 能通詩書者侍學御筵取裴國 鄧厥等三十人 其餘並留學秋七月命建康王龍益率師討炎 典諸蠻平之	{{合集}}(奏章){{卯}}(午) {{宋}}(淳熙){{十年}} {{冬}}(十一)月 {{壬午}}(正){{月}}以{{吳}}(履){{信}}為{{督}} 將{{代}}(蔡){{忠}}(臣){{等}}(國){{人}}入 {{雲}}屯{{進}}(寶){{物}}(乞)行 {{買}}(賣){{冬}}(十){{月}}(思){{蒙}}等 {{典}}(司)自{{相}}攻{{擊}} {{十}}(二){{乙}}(巳){{年}}(宋){{淳}} 熙){{春}}(正){{月}}試天下{{士}}人{{自}}十 五{{歲}} {{十}}(年) {{張}}通{{詩}}書{{者}}侍 {{學}}(御){{翼}}取{{裴}}國{{慎}}(鄧){{嚴}} 等{{三十}}人 {{其}}(俞)並{{留}}(學){{秋}}(七){{月}}命 {{建}}康{{王}}(龍){{益}}率{{師}}(討){{炎}} 等{{十}}人{{再}}(講){{平}}之

Figure 6: Document example where the unusual reading order causes the zero-shot VLM to make major mistakes.³

D.2 Recognition Challenges with Small Characters

Figure 7 illustrates another challenging Classical Chinese document, characterized by a mix of large title-like characters and densely packed columns of smaller body text. Like the previous example, this page follows a vertical writing style with right-to-left column ordering.

CHURRO transcribes the document with high accuracy. It preserves the reading order, successfully detects both large and small characters, and handles rare historical forms, with only minor mismatches due to character encoding limitations.

The zero-shot VLM, however, produces a significantly incomplete transcription. It primarily detects larger, visually prominent characters such as titles and headers, while failing to recognize most of the smaller text.

D.3 Reading Order in Western Languages

The postcard in Figure 8 begins on the right side of the page, as is customary with postcards, and continues on the top left. CHURRO recognizes the text mostly correctly, but it does not accurately identify the reading order. This leads to a prediction where the text order is reversed.

D.4 Major Hallucinations from Stereotypes

Some hallucinations due to misrecognition are based on stereotypes. In a mid-19th-century English newspaper (Figure 9), the first column begins with a report from India. The second column mentions the Queen traveling with the Princess

pacified.”

⁴The English translation of the gold text is: “To understand the harmony of Heaven and Earth is to discern signs of disorder. Each of the twelve earthly branches has its corresponding wind, and when the winds deviate from their natural rhythm, it is a sign that order is failing. In the eighteenth year of Duke Xiang, the state of Chu attacked Zheng. The musician Shi Kuang first sang of the north wind, then of the south wind. The south wind lacked strength and coherence, and Shi Kuang declared that Chu would not succeed. Indeed, many Chu soldiers suffered from cold, confirming the signs of disrupted harmony. To address such disruptions and govern effectively, five roles are established to offer counsel and maintain order. These roles help interpret omens and seasonal patterns, ensuring that state affairs proceed in accordance with cosmic and political balance. Among them are three minor emissaries and two senior diviners, appointed to advise the king. The Inner Historian manages eight categories of royal decrees: granting rank, bestowing emoluments, dismissal, appointment, execution, birth, reward, and confiscation. The Grand Minister also conveys royal edicts and serves as a secondary authority, helping oversee the laws and administrative affairs of the state. These officials work together to uphold governance, interpret laws, and plan state policy.”

Royal. The VLM is unable to decipher this text but captures enough words—particularly India and queen—that it generates a stereotypical passage about the Queen traveling to Calcutta with the Prince Royal. While it is statistically likely that those traveling to India from the UK in the 19th century went to Calcutta in large numbers and royal princes traveled more than royal princesses, text generation based on stereotypes results in factual mistakes.

D.5 Repetitive Degeneration

Figure 10 shows an example from a 19th-century German letter. The zero-shot VLM, upon gathering from the first line that this is a letter, repeatedly predicts the same sentence over a dozen times: Das Wetter ist hier sehr angenehm,” which translates to The weather here is really nice.” This is a very standard opening line for a letter.

In the case of Figure 11, a Latin manuscript, the zero-shot VLM predicts the word “arsagrium” and, without another word to follow, repeatedly generates the same word until it runs out of generation tokens. This manuscript remains difficult even for CHURRO. The fine-tuned VLM has several serious issues with recognizing the text and enters a degeneration loop with another phrase.

D.6 Errors Due to Historical Script Changes

Sometimes, in historical languages, it is not only the spelling that differs from contemporary usage but also the script itself. This is the case for many contemporary languages—for example, Azeri—which has historically been written in the Arabo-Persian script and continues to be written with that script in Iran. However, during the Soviet era in Azerbaijan, Azeri was written with the Cyrillic alphabet, and today it is written in the Latin alphabet. In our dataset, this diversity of scripts is showcased by transitional Romanian, a 19th-century alphabet that combined Cyrillic and Latin characters. Already a subject of computational linguistics research (Frincu et al., 2023), transitional Romanian poses challenges for VLMs. Figure 12 shows a 19th-century printed Romanian book. Without fine-tuning, the VLM does not recognize the Latin characters properly and defaults to Cyrillic spellings. These mistakes are drastically reduced in CHURRO.

Document image	Gold	Finetuned	Zero-shot LMM
	察天地之和命乖別之妖祥 十有二层皆有風吹其律以 知和不其遺亡矣春秋襄十 八年楚師伐鄭師曠曰吾觀北風又歌南風南風不 競多死聲楚無功是時楚師多謀其命乖別害之矣 凡此五物 者以詔教政訪序事 訪謀也見其象則當豫為之備以詔王教 其政且該歲天時占所宜序宣其謀 重言 凡此五物者三小人二 ○以詔教政二大卜一 內史掌王之八枋之灋以詔王治一曰辨二曰祿三 曰廢四曰置五曰殺六曰生七曰予八曰奪 大宰既以 詔王內史 又展布之○枋本又作 柄兵枋反○重吏反下同 重言 以詔王治二同上○一曰辨二曰 祿二見大掌○四曰置二大宰 執國…… 灋及國令之貳以攷政事以逆會計	察天地之和命乖別之妖祥 十有二层皆有風吹其律以 知和不其遺亡矣春秋襄十 八年楚師伐鄭師曠曰吾觀北風又歌南風南風不 競多死聲楚無功是時楚師多謀其命乖別害之矣 凡此五物 者以詔教政訪序事 訪謀也見其象則當豫為之備以詔王教 其政且該歲天時占所宜序宣其謀 重言 凡此五物者三小人二 ○以詔教政二大卜一 內史掌王之八枋之灋以詔王治一曰辨二曰祿三 曰廢四曰置五曰殺六曰生七曰予八曰奪 大宰既以 詔王內史 又展布之○枋本又作 柄兵枋反○重吏反下同 重言 以詔王治二同上○一曰辨二曰 祿二見大掌○四曰置二大宰 執國…… 灋及國令之貳以攷政事以逆會計	察天地之和命乖別之妖祥 凡此五物者以詔教政訪序事其政且該歲 歲天時占所宜序宣其謀 內史掌王之八枋之灋以詔王治一曰辨二曰 祿三 曰廢四曰置五曰殺六曰生七曰予八曰奪 凡史掌書內史讀之人事 方出之書為之為之辭也農去以方 出之以方成書而出之上 食五人處人在官者其祿是以是為諸從下 士視上實矣祿足以代 其耕也七倍下士十倍中士十倍上士 倍四夫十 如之內史掌書王命遂之 外史掌書外令王下書外掌四方之志 書名于四方 御史掌邦國都鄙及萬民之治令以贊家宰 治之令 吏及注及九治局凡治者受灋令焉為書寫 其治之法

Figure 7: A Classical Chinese document where the zero-shot VLM is not able to recognize the smaller text. ⁴

Document image	Gold	Finetuned	Zero-shot LMM
The document shows two pages of handwritten Danish text. The left page has dense cursive script covering most of its surface. The right page features several short entries separated by dates written at the top: "18-11-1907", "19-11-1907", and "20-11-1907". Each entry begins with "Kjære Jappel!". The handwriting is fluid and characteristic of early 20th-century personal correspondence.	Kjære Jappel! Jeg spekulerer næsten på om det skulde lønne sig at holde en hel Udstilling igjen til Høsten - i October - - Det var at tænke på - - Jeg har jo endnu mange Trapper der ikke har været i Ilden - Jeg antar en 30 Billeder der ikke er udstillede i Kristiania - og forresten er over 100 Billeder naturligvis skrige op - men jeg har i Berlin hvert År holdt Separat-udstillinger - og dertil store U Collectivudstillinger i Secession - Forrige Aar holdt jeg hos den fineste Kunsthandler Cassirer 2 og store Separatudstillinger, et År - og så vilde måtte ...	naturaligvis skryk op – men jeg har i Berlin hver år holdt separat- udstillinger – og disse store Collectivutstillinger i Secession - Forrige Aar holdt jeg hos den fineste Kunsthandler Cassier 2 store Separatudstillinger. et år – så vilde måske man vel hjemme finde sig hen – efter 5 Aars Ude- bleven – Min Billeder virker meget bedre her Kjære Jappel! Jeg spekulerer næsten på om det skulde lønne sig at holde en hel Udstilling igen til Høsten – i October – – Det var at tænke på – – Jeg har jo endnu mange Trapper der ikke har været i Ilden – Jeg antar en ...	[8.4.1909] Kjønn Dapp! Jeg spørger mesten po om det skulle lene seg at holde en rett tihelling oppe til Nosten - Oktober - Det var en endte po - Jeg har jo endto mange Troppen de seks man var Biden - Jeg anbær en 30 Billen de seks in nølling - Kustjama frede a

Figure 8: A 20th-century Danish postcard that begins on the right side of the page, which causes mistakes even in CHURRO predictions.

Document image	Gold	Finetuned	Zero-shot LMM
	The Birmingham Daily Post. VOL. I. No. 16. BIRMINGHAM, FRIDAY, DECEMBER 25, 1857. PRICE ONE PENNY. † A permanent issue of Exchequer notes, to the amount of fifteen or sixteen millions, might equally relieve the productive classes, but it would be less economical to the country. *An Exchequer bill (or a note, without interest,) "is to be current, and pass in any of the public revenues, aids, taxes, or supplies, or at the receipt of Exchequer in Westminster, after the 5th day of April next." Exchequer bills for £5. and £10. were issued in 1696, and for £10. continued to be issued till 1741, during which time bank notes were not issued for less than £20. Bank	The Birmingham Daily Post. VOL. I. No. 16. BIRMINGHAM, FRIDAY, DECEMBER 25, 1857. PRICE ONE PENNY. FOREIGN INQUIRIES. INDIA THE BELLE OF LUCKNOW. —We are glad that the Belle of Lucknow was sailing with the news of the fall of Lucknow, and that she was able to reach her destination in safety. The news of the fall of Lucknow was received with the most intense pleasure, and was greeted with the most enthusiastic cheering, as the Belle entered the harbour of the port of Calcutta.	The Birmingham Daily Post. Vol. I. No. 16. BIRMINGHAM, FRIDAY, DECEMBER 25, 1857. PRICE ONE PENNY. FOREIGN INTELLIGENT. INDIA. The Queen, with the Prince Royal, departs for Calcutta, on the 25th inst., to visit the Indian Army, and to confer with the Governor General on the progress of the campaign. The Prince of Wales, with the Princess Royal, departs for the East, on the 26th inst., to visit the Indian Army, and to confer with the Governor General on the progress of the campaign. The Prince of Wales, with the Princess Royal, departs for the East, on the 26th inst., to visit the Indian Army, and to confer with the Governor General on the progress of the campaign. The Prince of Wales, with the Princess Royal,

Figure 9: 19th-century English

D.7 Strict Evaluation

In Figure 10, words like beSSer” or ausgeschloSSen” (the character representing the German sharp S) are transcribed as besser” and ausgeschlossen” (with two modern s characters), following orthographic changes in current standard German spelling. Such normalizations, while diverging from the text on paper, are not the same

as actual recognition mistakes because they do not change the meaning of the text.

[illegible]

Figure 10: A 19th-century German letter where the zero-shot VLM exhibits repetitive degeneration.

This image shows folio 100v from the Lindisfarne Gospels. The page contains two columns of text in a dense, black, Carolingian minuscule script. The text is written on parchment that shows signs of age and wear. On the left margin, there are large, ornate initials in red ink, including a prominent 'B' at the top and a 'C' further down. The right margin contains smaller, less ornate initials and some red markings. The overall layout is typical of a medieval liturgical or biblical manuscript.

Figure 11: A Latin manuscript where both models exhibit repetitive degeneration.

[illegible]

Figure 12: A 19th-century transitional Romanian book where the zero-shot VLM is unable to recognize Latin characters.

Document Type	Cluster	Total	Train	Dev	Test
print	German	21,024	20,964	30	30
	French	14,648	14,588	30	30
	Spanish	8,426	8,366	30	30
	Latin	5,158	5,098	30	30
	Czech	3,145	3,085	30	30
	Dutch	3,016	2,956	30	30
	Polish	2,781	2,721	30	30
	Slovenian	2,723	2,663	30	30
	English	2,504	2,444	30	30
	Japanese	1,954	1,894	30	30
	Bulgarian	737	677	30	30
	Sanskrit	543	483	30	30
	Finnish	283	223	30	30
	Swedish	270	210	30	30
	Romanian	154	94	30	30
	Hindi	154	94	30	30
	Chinese	66	6	30	30
	Bangla	62	2	30	30
	Italian	50	50	0	0
	Serbian	48	48	0	0
	Malayalam	48	48	0	0
	Estonian	47	47	0	0
	Slovak	42	42	0	0
	Latvian	40	40	0	0
	Indonesian	38	38	0	0
	Corsican	35	35	0	0
	Portuguese	14	14	0	0
	Catalan	13	13	0	0
	Russian	9	9	0	0
	Greek	9	9	0	0
	Occitan	5	5	0	0
	Yiddish	3	3	0	0
	Hungarian	2	2	0	0
	Malay	1	1	0	0
	Javanese	1	1	0	0
	Croatian	1	1	0	0
handwriting	Chinese	5,113	5,053	30	30
	Spanish	4,512	4,452	30	30
	French	3,627	3,567	30	30
	English	3,439	3,379	30	30
	Dutch	3,262	3,202	30	30
	Arabic	2,367	2,307	30	30
	German	2,358	2,298	30	30
	Latin	1,596	1,536	30	30
	Japanese	892	832	30	30
	Khmer	622	562	30	30
	Swedish	552	492	30	30
	Norwegian	441	381	30	30
	Italian	430	370	30	30
	Greek	425	365	30	30
	Persian	417	357	30	30
	Vietnamese	416	356	30	30
	Turkish	237	177	30	30
	Sanskrit	182	122	30	30
	Hebrew	155	95	30	30
	Catalan	145	85	30	30
	Portuguese	72	12	30	30
	Czech	59	59	0	0
	Finnish	45	45	0	0
	Newari	19	19	0	0
	Russian	2	2	0	0
	Pali	2	2	0	0
	Danish	2	2	0	0
	Yiddish	1	1	0	0

Table 3: Language statistics in CHURRO-DS.

Script Name	Script Family	Script Type	Writing Direction	ISO 15924 Code	Count in CHURRO-DS
Latin	European	alphabet	LTR	Latn	71,749
Latin (Fraktur variant)	European	alphabet	LTR	Latf	12,987
Latin (Gaelic variant)	European	alphabet	LTR	Latg	117
Cyrillic	European	alphabet	LTR	Cyrl	952
Greek	European	alphabet	LTR	Grek	434
Hebrew	Middle Eastern	abjad	RTL (some bidirectional)	Hebr	159
Arabic	Middle Eastern	abjad	RTL (bidirectional)	Arab	3,021
Devanagari (Nagari)	Indic	abugida	LTR	Deva	698
Bengali (Bangla)	Indic	abugida	LTR	Beng	64
Malayalam	Indic	abugida	LTR	Mlym	50
Newa	Indic	abugida	LTR	Newa	195
Khmer	Southeast Asian	abugida	LTR	Khmr	624
Japanese (alias for Han + Hiragana + Katakana)	East Asian	logo-syllabary	vertical (RTL) and horizontal (LTR)	Jpan	2,779
Han (Hanzi, Kanji, Hanja)	East Asian	logo-syllabary	vertical (RTL) and horizontal (LTR)	Hani	5,662

Table 4: Script statistics in CHURRO-DS. Script family, type and writing direction are from Glottolog 5.1 (Hammarström et al., 2024).

Model Name	Model ID	Reasoning Budget	How Accessed
GPT-4o	gpt-4o-2024-11-20		
GPT-4o Mini	gpt-4o-mini-2024-07-18		
GPT-4.1	gpt-4.1-2025-04-14		
GPT-4.1 Mini	gpt-4.1-mini-2025-04-14		
GPT-4.1 Nano	gpt-4.1-nano-2025-04-14		
O1	o1-2024-12-17	medium	Azure OpenAI API
O3	o3-2025-04-16	medium	
O4 Mini	o4-mini-2025-04-16	medium	
GPT-5	gpt-5-2025-08-07	medium	
GPT-5 Mini	gpt-5-mini-2025-08-07	medium	
GPT-5 Nano	gpt-5-nano-2025-08-07	medium	
Claude Sonnet 3.7	claude-3-7-sonnet-20250219	2048 tokens	Google Vertex AI API
Claude Sonnet 4	claude-sonnet-4-20250514	2048 tokens	
Claude Opus 4.1	claude-opus-4-1-20250805	2048 tokens	
Gemini 2.5 Flash	gemini-2.5-flash	2048 tokens	
Gemini 2.5 Pro	gemini-2.5-pro	2048 tokens	
Qwen 2.5 VL 3B	Qwen/Qwen2.5-VL-3B-Instruct		Local GPU using vLLM v0.10.2
Qwen 2.5 VL 72B	Qwen/Qwen2.5-VL-72B-Instruct		
Phi 4 Multimodal	microsoft/Phi-4-multimodal-instruct		
InternVL 3.5 (30B)	OpenGVLab/InternVL3_5-30B-A3B		
MiMo VL (7B)	XiaomiMiMo/MiMo-VL-7B-RL-2508		
Nanonets OCR (4B)	nanonets/Nanonets-OCR-s		
Nemotron Nano VL (8B)	nvidia/Llama-3.1-Nemotron-Nano-VL-8B-V1		
NuMarkdown (8B)	numind/NuMarkdown-8B-Thinking		
Gemma 3 (27B)	google/gemma-3-27b-it		
R (4B)	YannQi/R-4B		
Skywork R1v3 (38B)	Skywork/Skywork-R1V3-38B		
RolmOCR (8B)	reducto/RolmOCR		
olmOCR (8B)	allenai/olmOCR-7B-0825		

Table 5: Model names, IDs, reasoning budget, and access methods.

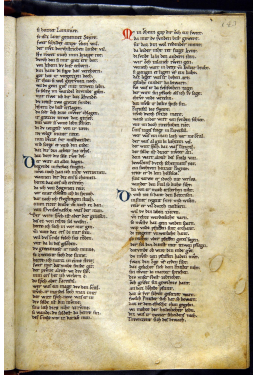
Document image	Gold	Finetuned	Zero-shot LMM
	er wirt an allez bagen. begynde in frvraz fragen. neve. noch han ich niht vernomen. wannen dir diz orf fi chomen. herre. daz ort ich erstreit. do ich von Sygvnen reit. vor einer chlosen ich di sprach. dar nach ich flvglingen stach. einen riter drabe in doch ez dan. von Mymselvasche was der man. Der wirt sich int aber der genesen. deß ez von rehte sol wesen. herre. ich sach in vor mir gen. in vant daz orf bi mir sten. wil dv Grals solch svs ræben. vnt da bi deß geleoben. dv gewinnest ir noch minne. so zweient sich die sinne. herre. ich namz in eime strit. swer mir dar vmbe vnde git. der proeve alrest wi dir ste. min orf het ich verwilvde. do sprach aber Parcfial.	si heizzet Lammire. so istz lant genennet Styre. swer schides ampt vben wil. der rfvz dvrchstrichen lande vil. Nv riwet mich min knappe rot. dvrch den si mir groz ere bot. von lthern dv bist erborn. din hant di lippe hat verchorn. got hat ir vergrezen doch. er chan si wol geprüven noch. w lv gein got mit triwen lebn. so solen im wandel drvmbe gebn. mit riwe ich dir daz khvnde. dv treit zwo groze fvnde: lthern dv hast erlagen. dv solt och dine mïver chlagen. ir groziiv triwe daz geries. din vart si vomë lebn schiet. di dv igveit von ir tate. nv volge miner rate. min fvze fvr müsswende. vnd sorge er vmb din ende: daz dir din arbet hie erhol.	st herzer Lammire. fo lthx lairt genemmer Styre. iwer schildet amptt vben wul. der mkrx dreyfstrichen lande vil. hy wweer mich min knappe rot. dych den m mir grox erx horx. von fehren dy brex erkorn. dun hatte typpx hat verehorn. gur hat ir vergessen doch. er chn si wol gerbrunx noch. hy w gern gor mit wriem lehm. fo feyex im wandel dremxe gebn. my rwey ich dir daz lkhnde. dy trestet two groxte frede. fehren dy halx elfaggen. ... einen ritter drade xoch rech dan. von Oynxelxelxelxelxelxelxelxelxelxel xelxelxelxelxelxelxelxelxelxelxel elxelxelxelxelxelxelxelxelxelxel lxelxelxelxelxelxelxelxelxelxelxel xelxelxelxelxelxelxelxelxelxelxel elxelxelxe ...

Figure 13: Middle High German

You are an expert in diplomatic transcription of historical documents from various languages. Your task is to extract the full text from a given page. Only output the transcribed text between <answer> and </answer> tags.

Follow these instructions:

1. You will be provided with a scanned document page.
2. Perform transcription on the entirety of the page, converting all visible text into the following format. Include handwritten and print text, if any. Include tables, captions, headers, main text and all other visible text.
3. If you encounter any non-text elements, simply skip them without attempting to describe them.
4. Do not modernize or standardize the text. For example, if the transcription is using “U+017F” instead of “s” or “U+0430” instead of “a”, keep it that way.
5. When you come across text in languages other than English, transcribe it as accurately as possible without translation.
6. Output the OCR result in the following format:

<answer>
extracted text here
</answer>

Remember, your goal is to accurately transcribe the text from the scanned page as much as possible. Process the entire page, even if it contains a large amount of text, and provide clear, well-formatted output. Pay attention to the appropriate reading order and layout of the text.

Table 6: The prompt used for all zero-shot experiments. “U+017F” is the Unicode for long S, used in 18th and early 19th century English. “U+0430” is the Unicode for cyrillic small letter “a”. We added these characters to the prompt after we noticed most VLMs try to normalize the text to modern English characters.

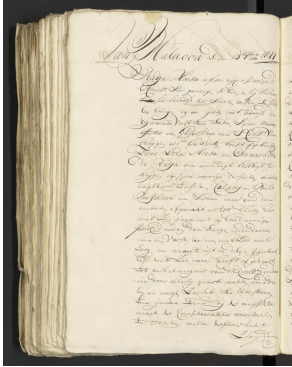
Document image	Gold	Finetuned	Zero-shot LMM
	Van Malacca Den 13 8-ber 1747 Radja Moeda af den rijx op volger E Comp,e Seer geneegen scheen en bij succes len brengen op een goeden voet, terwijl den regeerende vorst ten teeken, sijner waare affectie in Geselschap van 's Comp,s vaar thuijgen als Gezanten heeft over gezonden Ziere Lela Moeda, en Annachoda de Radja om eens regt beschijft te krijgen op sijne voorige versoecken, dewelke nogthans tusschen Calang en Poelo Passelaar in Swaar weer van den anderen afgeraakt, en tot heeden hier niet zijn g'arriveert op hoedaanige proceduuren dan teegen onderdaanen van een Vorst, die, men met zoo veel Zorg en moeite uijt een diepe vergeetent, heijd reets Zoo verre heeft over gehaalt, tot de belangens van d. E. Comp,s en daar in door allerlij praefiraable midde- len en weegen zoekt te koesteren ten jinde Eens bij de verijichte magt de Consideraable voordeelen te trekken, welke deselve met Zoo	Van Malacca Den 13 8ber 1747 Radja Moeda af den rijx op volgens d' E Comp: seer geneegen scheen en bij subven tie, sie beloofd de saaken werden te sullen ten brengen op een goeden voet, terwijl den regence vorst ten saeken, sijner Waare affectie in Geselschap van s Comp: Vaar, thuijgen als gezanten heeft overzonden Ziere Lela Moeda, en Annachoda de Radja om eens regt beschijft te krijgen op sijne voorige versoecken, dewelke nagthans Lisschen Calang en Poelo Parselaar in swaar weer van den anderen afgeraakt, en tot heeden hier met zijn garnisoen op hoed aanige proceduuren dan Leegen onderdaanen van een vorst, die, men met zoo veel zorg en moeite uijt en diverse vergeeten, heijd reets zoo verre heeft overgehaalt, tot de belangens van dE Comp,e, en daar in door allerlij practicaable middelen bij en weegen zaekt te koesteen ten ginder Eens bij de vereijichte ...	Van Malacca den 15.6.1787 Dit is een brief van de Heer genoemde, die is bij Subes Lee gebleven, de Saanke welke hij heeft gezocht, en is, geest voort terwijl de Regent heeft Fort te Sake, Syne Waar afwijst in Gezelschap van Heer D. van Kruyff als Ge Zantz heeft opgekomen. Hier is de Briefe, welke Annachtd De Raja om een Nyf betekyft te krijgen en sijne voorige Verfache, deel van Franse Laffele, Calang en Melo Gestelde in Siaan met een dew anderen afgesakt, en tot deze Eer met Zijt gaaen vest en Geen aanige Proed wier dan Leege onderdaan van er Vorst, die met Lee weel Zoog en mogtig wiert er dit e ...

Figure 14: 18th-century Dutch letter, bound

E List of 155 Sources Used to Create CHURRO-DS

DBNL OCR Data set (voor de Nederlandse Letteren , DBNL)

Description: “A set of 220 books digitised by the Dutch DBNL The texts range from 1776 to 1878 and are primarily in Dutch.”

Link: <https://zenodo.org/records/3239290#.XPfIqhYza70> https://www.dbnl.org/titels/titels_pd.php

License: Creative Commons Zero v1.0 Universal

Paris Bible Project (Guéville and Wrisley, 2021)

Description: “Latin Bibles throughout 13th and 14th century Europe. In the interest of diversifying the data we have opted for a wide variety of manuscript folios.”

Link: https://github.com/parisbible/group_truth/tree/main/PBP%201.0

License: “public domain”

POPP Datasets : Datasets for handwriting recognition from French population census (Constum et al., 2022)

Description:

Link: <https://zenodo.org/record/6581158>

License: Creative Commons Attribution 4.0 International

Deutschen Textarchiv (DTA) (Deutsches Textarchiv)

Description:

Link: <https://www.deutschestextarchiv.de/download>

License: CC BY-SA 4.0

Early Chinese Periodicals Online (ECPO) (Arnold, 2022)

Description:

Link: <https://github.com/hcts-hra/ecpo-data> <https://heidata.uni-heidelberg.de/dataset.xhtml?persistentId=doi:10.11588/DATA/Z3J0DV>

License: CC BY 4.0

Atlas lingüístico diacrónico e interactivo de la Comunidad de Madrid (ald, 2023)

Description:

Link: <https://aldicam.corpuscodea.es/consultas.php>

License: permission granted for research use

Corpus of Spanish Documents Prior to 1700 (CODEA) (Research Group Texts for the His-

Here we list the datasets unified in order to build CHURRO-DS, alongside a short description of each. The quotations in the “Description” field are direct excerpts taken from the corresponding paper or dataset description, as cited.

tory of Spanish [GITHÉ], 2022)

Description: “1,500 documents from various archives, covering all non-bilingual provinces of peninsular Spain and spanning from the 12th to the 17th centuries.”

Link: <https://corpuscodea.es/corpus/corpus2022/consultas.php>

License: CC BY-NC-ND 4.0

VML-HD: The Historical Arabic Documents Dataset for Recognition Systems (Kassis et al., 2017)

Description: “Based on five books written by different writers from the years 1088-1451.”

Link: <https://github.com/majeek/vml-hd?tab=readme-ov-file>

License: MIT License

A Dataset of French Trade Directories from the 19th Century (FTD) (Abadie et al., 2022)

Description: “Composed of pages and entries extracted from French directories published between 1798 and 1861.”

Link: <https://zenodo.org/records/6394464>

License: The images were extracted from the original source <https://gallica.bnf.fr>, owned by the Bibliothèque nationale de France (French national library). Original contents from the Bibliothèque nationale de France can be reused non-commercially, provided the mention “Source gallica.bnf.fr / Bibliothèque nationale de France” is kept. Researchers do not have to pay any fee for reusing the original contents in research publications or academic works. Original copyright mentions extracted from <https://gallica.bnf.fr/edit/und/conditions-dutilisation-des-contenus-de-gallica> on March 29, 2022.

The original contents were significantly transformed before being included in this dataset. All derived content is licensed under the permissive Creative Commons Attribution 4.0 International license.

Information Extraction in Historical Handwritten Records (IEHHR 2017) BH2M: the Barcelona Historical Handwritten Marriages database (Fernández-Mota et al., 2014)

Description:

Link: <https://rrc.cvc.uab.es/?ch=10&com=introduction> <http://dag.cvc.uab.es/the-historical-marriages-database/>

License: CC BY-NC-ND 4.0

Leopardi Dataset (Cascianelli et al., 2021)

Description: “Collection of early 19th Century letters written in Italian by Giacomo Leopardi”

Link: <https://github.com/aimagelab/HWD/releases/tag/leopardi>

License: CC BY 4.0

Modern Magazine OCR Training Dataset (Created by CODH and NINJAL) (Center for Open Data in the Humanities (CODH) and National Institute for Japanese Language and Linguistics (NINJAL), 2025)

Description: “Data from the modern magazines Kokumin no Tomo (Issues 1–27) and Meiroku Zasshi (Issues 1–43)”

Link: <https://codh.rois.ac.jp/modern-magazine/dataset/>

License: CC BY 4.0

Japanese Classical Books (Center for Open Data in the Humanities, 2019)

Description: “In addition to the field of Japanese literature, this collection includes classical books from various fields such as medicine, science, and industry held by the National Institute of Japanese Literature. It also includes cookbooks and other materials held by the Ajinomoto Foundation for Dietary Culture, and features classical books photographed by the National Institute of Japanese Literature.”

Link: <https://codh.rois.ac.jp/pmjt/>

License: CC BY-SA 4.0

HisClima Dataset (PRHLT Research Center, 2022)

Description: “208 descriptive text page images of the Jeannette logbook and with 72 table images coming from different log books of a ship called Albatross.”

Link: <https://zenodo.org/records/7442971>

License: CC BY 4.0

StABS Ratsbücher O10, Urfehdenbuch X (Hodel et al., 2021)

Description:

Link: <https://zenodo.org/records/5153263>

License: CC-BY-NC-SA

The git-project-Boccace by the Boccace team (Efsthathiou et al., 2022)

Description: “it is composed of two documents:

1. A Middle French incunabulum, namely BnF Rés. J-845 (département Réserve des livres rares, Rés. J-845) written between 1498-1499, scanned in high definition in black and white (notice <http://ark.bnf.fr/ark:/12148/cb30116914c>).

2. The editio princeps of the work, namely Bibliothèque Mazarine Inc. 59 (notice <https://www.sudoc.fr/135345618>) scanned in high definition in color.”

Link: <https://github.com/PSL-Chartes-HTR-Students/HN2021-Boccace>

License: MIT License

19th-Century Romanian Transitional Script (Penteliuc, 2022)

Description: “Romanian texts written in the Romanian Transitional Script (RTS). RTS is a mix of Latin and Cyrillic characters that were used in the 19th century in the Romanian provinces to facilitate the transition from the Romanian Cyrillic Script to the modern Latin Script. The images cover the period between 1833 and 1864. The selected texts cover a diverse range of literary genres, including poems, novels, dramas, stories, newspapers, and religious texts.”

Link: <https://www.kaggle.com/datasets/mariuspenteliuc/rts-ocr>

License: “You may use the dataset freely as long as you mention this page or the project below.”

NomNaOCR (Dang et al., 2022)

Description: st

Link: <https://www.kaggle.com/datasets/quandang/nomnaocr>

License: CC BY-NC-SA 4.0

University of Denver Jewish Consumptives Relief Society Medical Records (Pham, 2020)

Description: “All individuals for whom records are provided have been deceased for at least 70 years, but were they still living today, these records would be recognized as being protected health information under the US Health Insurance Portability and Accountability Act of 1996 (HIPAA).”

Link: <https://zenodo.org/records/4243023>

License: CC BY 4.0

Vienna, Österreichische Nationalbibliothek - HTR Winter School 2022 (Attwood et al., 2022)

Description: “This is ground truth for Vienna, Österreichische Nationalbibliothek, 2160, which is dated to the 3rd quarter of the 9th century (North

Italy), probably in the vicinity of Rome (Kaiser, Liebs).”

Link: <https://zenodo.org/records/7537204>

License: CC BY 4.0

6000 ground truth of VOC and notarial deeds 3.000.000 HTR of VOC, WIC and notarial deeds (Keijser, 2024)

Description: “17th and 18th century documents from the Dutch East-India Company (VOC). And 19th century notarial deeds from Noord-Hollands Archief and other archives in the provinces.”

Link: <https://zenodo.org/records/11209325>

License: CC BY 4.0

Incunabula Reichenau (Landesbibliothek et al., 2024)

Description: “Gothic and Antiqua typefaces found in Latin incunabula and early prints. based on a collection of incunabula and post-incunabula of the former Reichenau monastery, now held at the Badische Landesbibliothek in Karlsruhe.”

Link: <https://zenodo.org/records/11046062>

License: CC-BY-SA 4.0

Towards a general open dataset and model for late medieval Castilian text recognition (HTR/OCR) (Levenson, 2023)

Description: “All manuscripts are known to have been produced in the 15th century. The incunabula was printed in 149 The corpus is a good basis for the production of a general corpus for the late medieval period (14th-15th or even 13th-15th centuries), given its extension and the diversity of hands represented.”

Link: <https://zenodo.org/records/8340483>

License: CC BY-NC-SA 4.0

Ground truth of a sample Rudolph Gwalther’s handwriting from “Lateinische Gedichte” (Stotz and Ströbel, 2021)

Description: “Rudolph Gwalther’s (1519-1586) handwriting taken from his book Lateinische Gedichte, where he accumulated writings between 1540 and 1580.”

Link: <https://zenodo.org/records/4780947>

License: CC BY-NC 4.0

ARletta: open-source handwritten text recognition models for historic Dutch (Lefranc,

2024)

Description: “Manu: At the end of 2023, no models were publicly available in the kraken model repository (Kiessling, 2020) for the automated transcription of handwritten sources in historic Dutch, except the multilingual medieval CATMuS model that was partially trained on Middle Dutch (Clérice et al., 2023). Of all the models on HTR-united, Manu McFrench came closest to our Antwerp focus material because it was trained on French materials with a similar temporal scope (Chagué et al., 2023).

IJsberg: Two publicly available training sets created by the Dutch national archives seemed valuable to us because they share the same language, historical period, and administrative character (Keijser, 2020): VOC: 4,735 pages of the 17th and 18th century East-Asia Company archive

Notarial: 1,615 pages of 19th-century Dutch notarial deeds Antwerp incident books: We prepared a ground truth dataset from the Antwerp incident books. One expert (PhD student) annotated an initial subset of the data, consisting of 271 randomly selected pages.² This subset was later supplemented with 3444 pages annotated by students.³ The student-contributed materials are much larger in scope than the materials annotated by the expert, but the formers quality is unfortunately much less consistent.”

Link: <https://zenodo.org/records/1119145>
7

License: CC-BY-NC-SA

Diario del soldato Bruno Celestino (, Belgium)

Description: “diary of Italian soldier Bruno Celestino in World War I”

Link: <https://zenodo.org/records/1376058>
6

License: CC0 1.0

Dagboek Ernest Clarysse (, Belgium)

Description: “diary of Belgian citizen Ernest Clarysse in World War I”

Link: <https://zenodo.org/records/1376922>
2

License: CC0 1.0

Mémoire sur St Domingue par H ? M. Michel (, Belgium)

Description: “Transcription of monography Mémoire sur St Domingue par H ? M. Michel, dating

from 1797 and dealing on slavery in Haiti”

Link: <https://zenodo.org/records/1378441>
1

License: CC0 1.0

TKH and MTH Datasets MTHv2 (Yang et al., 2018)

Description: “The Tripitaka Koreana in Han (TKH) Dataset and the Multiple Tripitaka in Han (MTH) Dataset for the research of Chinese character detection and recognition in historical documents”

Link: https://github.com/HCIILAB/TKH_MTH_Datasets_Release https://github.com/HCIILAB/MTHv2_Datasets_Release

License: “The TKH dataset and MTH dataset can only be used for non-commercial research purpose.”

M5HisDoc (Shi et al., 2023)

Description: “M5HisDoc consists of 5,000 historical Chinese handwritten document images collected from ancient books and manuscripts. The documents are typically structured with form-like layouts”

Link: <https://github.com/HCIILAB/M5HisDoc>
c

License: CC BY-NC-ND 4.0

The Pinkas Dataset (Rabaeu et al., 2019)

Description: “Pinkas dataset is created from a historical Hebrew manuscript that contains records of Jewish communities in Europe in the early modern period (c. 1500-1800). The dataset consists of 30 pages digitized by full color digital images in JPG format with high resolution. The pages exhibit numerous degradations, complex layout and different handwritings. They are written in a mixture of Medieval Hebrew by different writers. The writers were not professional scribes. This adds additional challenge, since very often the same letters are written in different shapes.”

Link: <https://zenodo.org/records/3569694>

License: CC BY 4.0

Saint Gall database (Fischer et al., 2012)

Description: “The Saint Gall database presented in [1] contains a handwritten historical manuscript with following characteristics:

9th century, Latin language, single writer, Carolingian script, ink on parchment”

Link: <https://fki.tic.heia-fr.ch/databas>

[es/saint-gall-database](#)

License: “The Saint Gall database may be used for non-commercial research and teaching purposes only. If you are publishing scientific work based on the Saint Gall database, we request you to include a reference to our paper [1] A. Fischer, V. Frinken, A. Fornés, and H. Bunke: "Transcription Alignment of Latin Manuscripts using Hidden Markov Models," in Proc. 1st Int. Workshop on Historical Document Imaging and Processing, pages 29-36, 2011.

With kind permission of Prof. Ernst Tresp from the Abbey Library of Saint Gall, the original manuscript images [3] provided by e-codices can be used for non-commercial research and teaching purposes explicitly as follows:

Show and print sample manuscript images in scientific publications Show sample manuscript images during talks Show sample manuscript images online For any purposes other than non-commercial research and teaching, the Abbey Library of Saint Gall has to be contacted first.

With kind permission of Max Bänziger from the Monumenta project, the aligned text edition [4] is also included in the Saint Gall database.”

Parzival Database (Fischer et al., 2009)

Description: “The Parzival database described in [4] contains a handwritten historical manuscript with following characteristics:

13th century Medieval German language three writers Gothic script ink on parchment”

Link: <https://fki.tic.heia-fr.ch/databases/parzival-database>

License: “The Parzival database may be used for non-commercial research and teaching purposes only. If you are publishing scientific work based on the Parzival database, we request you to include a reference to [1] A. Fischer, A. Keller, V. Frinken, and H. Bunke: "Lexicon-Free Handwritten Word Spotting Using Character HMMs," in Pattern Recognition Letters, Volume 33(7), pages 934-942, 2012.

With kind permission of Prof. Ernst Tresp from the Abbey Library of Saint Gall, the original manuscript images can be used for non-commercial research and teaching purposes explicitly as follows:

Show and print sample manuscript images in scientific publications Show sample manuscript images during talks Show sample manuscript images online For any purposes other than non-

commercial research and teaching, the Abbey Library of Saint Gall has to be contacted first.”

The “ScribbleLens” Dutch Historical Handwriting Corpus (Dolfing et al., 2020)

Description: “one of the first manuscript recognition corpus for Early Modern Dutch, 400 years old material from the Dutch East India company (VOC) with ship journals and other written Material” “80 writers” “The manuscripts span over 150 years of significant journeys by captains and traders from the Vereenigde Oost-indische Company (VOC) such as Tasman, Brouwer and Van Neck”

Link: <https://openslr.org/84/>

License: CC-BY-NC-ND

SLOVAK Supermodel print & typewriter (SSPT1) (Katuák et al., 2024)

Description: “92 s. GT of historical printed book J.A. Comenius’ Orbis Pictus (1798 edition). the book has many illustrations, it is written in 4 languages (Latin, Hungarian, German, Czech), in addition in the form of tables and in antique and Swabian script

a whole series of historical newspapers, magazines and books from the 19th and early 20th centuries (Moravské noviny (1849), Program bulletins of the Slovak Philharmonic (1849-1970), Opavský Besedník (1863), Jitrenka (1840), I. Palugya: Kde jest pravda (1854), Lusatian Serbian magazine Luica (1909), labikár (1872), J.M. Hurban: Cirkev Ewanjelicko- Lutheránska (1861), J.N. Bobula: Jánošík (1862), D. Lichard: Obzor (1866) and others

typewritten documents, using samples of various fonts in Slovak, Czech, and German languages (ca. 150 pp.)

ca. 80 s. of GT transcriptions from various historical prints from the 18th and 19th centuries written in Czech (Svabian).”

Link: <https://zenodo.org/records/11218527>

License: CC BY 4.0

konzilsprotokolle (Grüning et al., 2016)

Description: “this dataset contains 8770 transcribed text lines of handwritten historical documents from the late 18th century.”

Link: <https://zenodo.org/records/215383>

License: CC BY 4.0

Charters and Records of Königsfelden Abbey and Bailiwick (Halter-Pernet et al., 2021)

Description: During its existence (1308-1528), the double monastery of Königsfelden was one of the central clerical institutions of Aargau and continued to exert a great attraction on the people of the region and beyond for centuries. The project makes the rich tradition from the Middle Ages and the early modern period accessible to researchers and interested laypeople in the form of a digital edition.

Link: <https://zenodo.org/records/5179361>

License: CC BY 4.0

NorHand v3 (Beyer and Solberg, 2023)

Description: “Norwegian letter and diary documents from 19th and early 20th century”

Link: <https://zenodo.org/records/10255840>

License: CC BY 4.0

The Finnish Court Records Dataset (Toselli and Vidal, 2021)

Description: 600 random sample of “The Finnish Court Records (FCR) collection encompasses 785 manuscripts from the “Renovated District Court Records” held by the National Archives of Finland. Many of these manuscripts were scanned into double-page images, amounting to 630,388 images of about one million pages in total. The manuscripts date from the 18th century and consist of records of deeds, mortgages, traditional life annuity, among others. They were written by many hands mostly in Swedish”

Link: <https://zenodo.org/records/4767732>

License: CC BY 4.0

ICDAR 2015 Competition HTRtS: Handwritten Text Recognition on the tranScriptorium Dataset (Sánchez et al., 2017)

Description: “drawn from the English “Bentham collection” dataset used in the TRANSCRIPTORIUM project. The selected data has been written by several hands and entails significant variabilities and difficulties regarding the quality of text images, writing styles and crossed-out text.”

Link: <https://zenodo.org/records/248733#.WH3zMczhBTY>

License: CC BY 4.0

The e-NDP project : collaborative digital edi-

tion of the Chapter registers of Notre-Dame of Paris (Claustre et al., 2023)

Description: According to our estimates no fewer than 18 main hands were involved in the writing of the registers during the medieval period. More than 98% of the content of the registers was written in Latin, the rest in French. The exact percentage is hard to estimate because the vernacular language is often used in formulae, notes and comments. It is rare to find entire pages or blocks written in French.

Script family : The registers were written using a Cursive script (ca. late XIIIe - XVIe).

Documental typology : The volumes containing the chapter conclusions were conceived to serve as memorial records, but above all as documents for regular use and consultation in the daily practice of administration and management. In diplomatics the notion of “documentary manuscripts” is used to describe this kind of sources also by opposition to books and literary or normative manuscripts.

Link: <https://zenodo.org/records/7575693>

License: CC BY 4.0

Faithful Transcriptions Data Set: TEI/XML-encoded Transcriptions of Medieval Theological Manuscripts (Eichenberger and Suwelack, 2021)

Description: “181 pages with 8,952 text lines from 12 manuscripts in German, Dutch, and Latin. The medieval scripts include Textura, Textualis, Gothic Cursiva, and Bastarda”

Link: <https://zenodo.org/records/5582483>

License: CC0 1.0

A Benchmark and Dataset for Post-OCR text correction in Sanskrit (Maheshwari et al., 2022b)

Description: “a multi-domain dataset from 30 different books and have 218,000 manually verified sentences in it” “We consider printed versions of these books, most of them reprinted in the first half of the twentieth century. While, these books are widely accessible to the public via libraries and academic institutions, we manually had to scan several of them as part of its digitisation process. These books vary widely in their vocabulary and stylistic usage owing to the differences in the domain and the original time period of publication, where the latter can be as old as the fifth century AD.”

Link: <https://github.com/ayushbits/pe-o>

cr-sanskrit

License: N/A

ICDAR2017 Competition on Handwritten Text Recognition on the READ Dataset (Maheshwari et al., 2022a)

Description: “Most of the dataset was taken from the Alfred Escher Letter Collection (AEC) 5 which is written in German but it also has pages in French and Italian. The selected dataset also included handwritten images drawn from other German collections, and thus characterized by being written by several hands.”

Link: <https://zenodo.org/records/835489>

License: CC BY 4.0

ICFHR2016 Competition on Handwritten Text Recognition on the READ Dataset (Sánchez et al., 2016)

Description: “a subset of documents from the Ratsprotokolle collection composed of minutes of the council meetings held from 1470 to 1805 (about 30.000 pages), which will be used in the READ project. This dataset is written in Early Modern German. The number of writers is unknown. Handwriting in this collection is complex enough to challenge the HTR software.”

Link: <https://zenodo.org/records/218236>

License: CC BY 4.0

RASAM 1 and 2 (Vidal-Gorène et al., 2021)

Description: “a diverse collection of Maghrebi Arabic manuscripts from the BULAC Library, featuring a wide variety of handwriting styles, layouts, states of preservation, and other characteristics representative of Arabic Maghrebi manuscript production.”

Link: <https://github.com/calfa-co/rasam-dataset>

License: Apache 2.0

Fraktur-GT (Springmann and Fink, 2018)

Description: “Ground Truth (GT) data for Fraktur/Gothic prints from the 19th Century, released by UB, Uni-Tübingen”

Link: <https://github.com/ubtue/gt-fraktur>

License: CC0 1.0

Ground Truth transcriptions for training OCR of historical Arabic handwritten texts - RASM2018 and RASM2019 (Clausner et al.,

2019b)

Description: “drawn from a selection of historical Arabic scientific manuscripts (10th-19th century) digitised through the British Library Qatar Foundation Partnership”

Link: <https://www.primaresearch.org/RASM2018/resources>

License: “All ground truth resources created for RASM competitions are freely available under an open license”

Prima Europeana (Clausner et al., 2015)

Description: Newspapers from 17th to 20th century

Link: <https://www.primaresearch.org/datasets>

License: “Resources are available exclusively for personal research, if intended to use any software or data in the scope of commercial activities, please contact enquiries@primaresearch.org to discuss individual licensing options.”

Prima Impact (Papadopoulos et al., 2013)

Description: “texts from as early as 1500, and containing material from newspapers, books, pamphlets and typewritten notes”

Link: <https://www.primaresearch.org/datasets>

License: “Resources are available exclusively for personal research, if intended to use any software or data in the scope of commercial activities, please contact enquiries@primaresearch.org to discuss individual licensing options.”

BLN600: A Parallel Corpus of Machine/Human Transcribed Nineteenth Century Newspaper Texts (Booth et al., 2024)

Description: “nineteenth-century newspaper text focused on crime in London, derived from the Gale British Library Newspapers corpus parts 1 and 2”

Link: https://orda.shef.ac.uk/articles/dataset/BLN600_A_Parallel_Corpus_of_Machine_Human_Transcribed_Nineteenth_Century_Newspaper_Texts/25439023

License: CC BY-NC-ND 4.0

REID2019 ICDAR Competition on Recognition of Early Indian printed Documents (Clausner et al., 2019a)

Description: “printed books written in Bengali (1713-1914)”

Link: <https://www.primaresearch.org/data-sets>

License: “Resources are available exclusively for personal research, if intended to use any software or data in the scope of commercial activities, please contact enquiries@primaresearch.org to discuss individual licensing options.”

ICDAR2017 Competition on Recognition of Early Indian Printed Documents , REID2017 (Clausner et al., 2017)

Description: “Books in Bengali language dating between 1785 and 1909. For the most part, the scanned images contain single column lines of text, with a small amount containing illustrations as well as text. Some pages contain marginal data such as numbers, handwritten notes, and decorative frames.”

Link: <https://www.primaresearch.org/data-sets>

License: “Resources are available exclusively for personal research, if intended to use any software or data in the scope of commercial activities, please contact enquiries@primaresearch.org to discuss individual licensing options.”

Gado2: multilingual newspapers from the Netherlands Indies (Kemper, 2021)

Description:

Link: <https://zenodo.org/records/4980170>

License: CC BY 4.0

Klosterneuburg, Stiftsbibl., Cod. 48 (Berger et al., 2022)

Description: “collection of sermons of Nikolaus von Dinkelsbühl (ca. 1360 to 17th March 1433), translated and reorganised by a German redactor, from the 15th century has never been edited until now. It consists of 361 folios of parchment and paper. The text speaks about various topics such as fasting and other religious practices”

Link: <https://zenodo.org/records/7466928>

License: CC BY 4.0

The Datafication of Early Modern Ordinances (Romein et al., 2020)

Description: “All books have been published between 1532 and 1789, within the seventeen federation-states of the Habsburg Netherlands and the Dutch Republic”

Link: <https://doi.org/10.5281/zenodo.3556795> <https://doi.org/10.5281/zenodo.3556788>

<https://doi.org/10.5281/zenodo.3552470> <https://doi.org/10.5281/zenodo.3612505> <https://doi.org/10.5281/zenodo.3612561> <https://doi.org/10.5281/zenodo.3612571> <https://doi.org/10.5281/zenodo.3612577> <https://doi.org/10.5281/zenodo.3612606> <https://doi.org/10.5281/zenodo.3612613> <https://doi.org/10.5281/zenodo.3613585> <https://doi.org/10.5281/zenodo.3613937> <https://doi.org/10.5281/zenodo.3624664> <https://doi.org/10.5281/zenodo.3614031> <https://doi.org/10.5281/zenodo.3625184> <https://doi.org/10.5281/zenodo.3625615> <https://doi.org/10.5281/zenodo.3625186> <https://doi.org/10.5281/zenodo.3625188> <https://doi.org/10.5281/zenodo.3625763> <https://doi.org/10.5281/zenodo.3614025> <https://doi.org/10.5281/zenodo.3613987> <https://doi.org/10.5281/zenodo.3625168> <https://doi.org/10.5281/zenodo.3613916> <https://doi.org/10.5281/zenodo.3613893> <https://doi.org/10.5281/zenodo.3625073> <https://doi.org/10.5281/zenodo.3625117> <https://doi.org/10.5281/zenodo.3625101> <https://doi.org/10.5281/zenodo.3625119> <https://doi.org/10.5281/zenodo.3624700> <https://doi.org/10.5281/zenodo.3625079> <https://doi.org/10.5281/zenodo.3625081> <https://doi.org/10.5281/zenodo.3625121> <https://doi.org/10.5281/zenodo.3625653> <https://doi.org/10.5281/zenodo.3625659> <https://doi.org/10.5281/zenodo.3625661> <https://doi.org/10.5281/zenodo.3613895> <https://doi.org/10.5281/zenodo.3625097> <https://doi.org/10.5281/zenodo.3625767> <https://doi.org/10.5281/zenodo.3625105> <https://doi.org/10.5281/zenodo.3625109> <https://doi.org/10.5281/zenodo.3625111> <https://doi.org/10.5281/zenodo.3625113> <https://doi.org/10.5281/zenodo.3625790> <https://doi.org/10.5281/zenodo.3625115> <https://doi.org/10.5281/zenodo.3624770> <https://doi.org/10.5281/zenodo.3624882> <https://doi.org/10.5281/zenodo.3625107> <https://doi.org/10.5281/zenodo.3625513> <https://doi.org/10.5281/zenodo.3625655> <https://doi.org/10.5281/zenodo.3625657> <https://doi.org/10.5281/zenodo.3625085> <https://doi.org/10.5281/zenodo.3624902> <https://doi.org/10.5281/zenodo.3625083> <https://doi.org/10.5281/zenodo.3625093>

<https://doi.org/10.5281/zenodo.3625497>
<https://doi.org/10.5281/zenodo.3613994>
<https://doi.org/10.5281/zenodo.3614003>
<https://doi.org/10.5281/zenodo.3624898>
<https://doi.org/10.5281/zenodo.3625091>
<https://doi.org/10.5281/zenodo.3625772>
<https://doi.org/10.5281/zenodo.3613925>
<https://doi.org/10.5281/zenodo.3613908>
<https://doi.org/10.5281/zenodo.3613914>
<https://doi.org/10.5281/zenodo.3614019>
<https://doi.org/10.5281/zenodo.3624563>
<https://doi.org/10.5281/zenodo.3614005>
<https://doi.org/10.5281/zenodo.3625172>
<https://doi.org/10.5281/zenodo.3625174>
<https://doi.org/10.5281/zenodo.3625170>
<https://doi.org/10.5281/zenodo.3625178>
<https://doi.org/10.5281/zenodo.3625509>
<https://doi.org/10.5281/zenodo.3625087>
<https://doi.org/10.5281/zenodo.3625176>
<https://doi.org/10.5281/zenodo.3625075>
<https://doi.org/10.5281/zenodo.3556788>
<https://doi.org/10.5281/zenodo.3625645>
<https://doi.org/10.5281/zenodo.3625649>
<https://doi.org/10.5281/zenodo.3625651>
<https://doi.org/10.5281/zenodo.3625507>
<https://doi.org/10.5281/zenodo.3625511>
<https://doi.org/10.5281/zenodo.3625505>
<https://doi.org/10.5281/zenodo.3625647>
<https://doi.org/10.5281/zenodo.3625099>
<https://doi.org/10.5281/zenodo.3625089>
<https://doi.org/10.5281/zenodo.3613523>
<https://doi.org/10.5281/zenodo.3613601>
<https://doi.org/10.5281/zenodo.3613617>
<https://doi.org/10.5281/zenodo.3613663>
<https://doi.org/10.5281/zenodo.3613927>
<https://doi.org/10.5281/zenodo.3613555>
<https://doi.org/10.5281/zenodo.3613518>
<https://doi.org/10.5281/zenodo.3624906>
<https://doi.org/10.5281/zenodo.3624908>
<https://doi.org/10.5281/zenodo.3624910>
<https://doi.org/10.5281/zenodo.3624912>
<https://doi.org/10.5281/zenodo.3624914>
<https://doi.org/10.5281/zenodo.3624916>
<https://doi.org/10.5281/zenodo.3624918>
<https://doi.org/10.5281/zenodo.3624920>
<https://doi.org/10.5281/zenodo.3624922>
<https://doi.org/10.5281/zenodo.3613553>
<https://doi.org/10.5281/zenodo.3625103>
<https://doi.org/10.5281/zenodo.3613974>
<https://doi.org/10.5281/zenodo.3614027>
<https://doi.org/10.5281/zenodo.3625182>

<https://doi.org/10.5281/zenodo.3625501>
<https://doi.org/10.5281/zenodo.3625503>
<https://doi.org/10.5281/zenodo.3625190>
<https://doi.org/10.5281/zenodo.3625643>

License: CC BY 4.0

NewsEye / READ OCR training dataset from Swedish Newspapers (18th, 19th, early 20th C.) (Muehlberger and Hackl, 2021b)

Description: “The dataset comprises swedish newspaper pages from late 18th till early 20th century with carefully corrected text. The page images were provided by the National Library Finland (NLF)”

Link: <https://zenodo.org/records/4599624>

License: CC BY 4.0

NewsEye / READ OCR training dataset from Austrian Newspapers (19th C.) (Muehlberger and Hackl, 2019)

Description: “The dataset comprises Austrian newspaper pages from 19th and early 20th century with carefully corrected text. The page images were provided by the Austrian National Library”

Link: <https://zenodo.org/records/3387369>

License: CC BY 4.0

NewsEye - READ OCR training dataset from French Newspapers (18th, 19th, early 20th C) NewsEye - READ AS training dataset from French Newspapers (19th, early 20th C) (Muehlberger and Hackl, 2020)

Description: “The dataset comprises French newspaper pages from 18th, 19th and early 20th century with carefully corrected text. The page images were provided by the French National Library”

Link: <https://zenodo.org/records/4293602>

<https://zenodo.org/records/5654841>

License: CC BY 4.0

NewsEye / READ OCR training dataset from Finnish Newspapers (18th, 19th, early 20th C.) NewsEye / READ AS training dataset from Finnish Newspapers (19th C.) (Muehlberger and Hackl, 2021a)

Description: “The dataset comprises finnish newspaper pages from late 18th till early 20th century with carefully corrected text. The page images were provided by the National Library Finland (NLF)”

Link: <https://zenodo.org/records/4599472>

<https://zenodo.org/records/5654858>

License: CC BY 4.0

Ground truth for Neue Zürcher Zeitung black letter period (Ströbel and Clematide, 2019)

Description: “The Neue Zürcher Zeitung (NZZ) has been publishing in black letter from its very first issue in 1780 until 1947. From this time period, we randomly sampled one frontpage per year, resulting in a total of 167 pages. We chose frontpages because they typically contain highly relevant material and because we want to make sure not to sample pages containing exclusively advertisements or stock information. During certain periods, the NZZ was published several times a day, and there were supplements, too. Due to incomplete metadata, the sampling included frontpages from supplements.”

Link: <https://github.com/impresso/NZZ-black-letter-ground-truth>

License: CC BY-NC 4.0

OCR17+ - Layout analysis and text recognition for 17th c. French prints (Jahan and Gabay, 2021)

Description: “17th c. French print”

Link: <https://github.com/Heresta/OCR17plus>

License: Data is CC-BY, except images which come from Gallica (cf. conditions d’utilisation).

The OCR-D project (Boenig, 2024)

Description: “Union Catalogue of Books of the 16th–18th century (VD 16, VD 17, VD 18) published in the German-speaking countries”

Link: https://github.com/OCR-D/gt_structure_text

License: CC-BY-SA-4.0

A dataset of Spanish notarial deeds (18th Century) for Handwritten Text Recognition and Layout Analysis of historical documents (Quirós et al., 2018)

Description: “a subset of 596 documents from the Registre d’Hypoteques de Girona of 1769 collection, guarded by the Arxiu Històric de Girona. This collection, is composed by hundreds of thousands of notarial deeds from the XVIII-XIX century (1768-1862). Sales, redemption of censuses, inheritance and matrimonial chapters are among the most common documentary typologies in the

collection.”

Link: <https://zenodo.org/records/1322666#.Ypi6Ty8RoUE>

License: CC BY-NC 4.0

Chronicling Germany (Schultze et al., 2024)

Description: “693 annotated historical newspaper pages from the time period between 1852 and 1924”

Link: <https://github.com/Digital-History-Bonn/Chronicling-Germany-Code/blob/master/script/download.py>

License: EUROPEAN UNION PUBLIC LICENCE v. 1.2

Manuscripts of Handwritten Arabic dataset (Muharaf) for cursive text recognition (Saeed et al., 2024)

Description: “The Muharaf dataset consists of a diverse set of images, ranging from individual personal letters, poems, and dialogues to legal consensus records, correspondences, and church records. The manuscripts date from the early 19th century to the early 21st century. The quality of page images varies, from writing on a clean white background to illegible sentences on creased pages with ink bleeds”

Link: <https://zenodo.org/records/11492215>

License: CC BY-NC-SA 2.0

A scarce dataset for ancient Arabic handwritten text recognition (Najam and Faizullah, 2024)

Description: “The books and their authors, orderly, as appearing in the dataset, are the following: Rafaa Al Niqab An Kitab Al Shahab, Al Husain Al Shushawi; Nuzul Al Saereen Ela Allah Rabb Al Aalamin Fee Ahadith Said Al Mursalin, Mahmoud Al Darkazini; Kitab Al Azamah, Ibn Hayyan Al Asbahani; Al Rawd Al Nadheer, Muhammad Mutawalli; Sirr Al Fosoos, Muhammad Abdulbaqi; Al Juzz Alkhamis Min Jamii Al Masanid Wa Al Sunan Al Hadi Li Aqwa Sunan, Ismail Ibn Kathir; Kitab Tareekh Madinat Dimashq Wa Mn Banaha Min Al Mutaqaddimin, Ali Al Rabi; Dirham Al Suarrah Fee Wad Al Yadain Taht Al Surrah, Muhammad Hashim Al Sindi.”

Link: <https://data.mendeley.com/datasets/xz6f8bw3w8/1>

License: CC BY 4.0

Caroline Minuscul (White et al., 2022)

Description: Manuscripts from 800 to 1100

Link: <https://github.com/rescribe/carolineminuscul-groundtruth>

License: “The ground truth contained in this repository should be considered Public Domain, or licensed under Apache License 2.0, whichever suits your needs better.”

The metadata.txt files in each manuscript directory describes its provenance and the licensing of the images. We have only used manuscript images which are freely redistributable and reusable.

(Ghereghlou, 2018)

Description: Transcription is adopted from the book "Qsim Beg ayt Tabrz: A Chronicle of the Early Safavids and the Reign of Shah Isml (907–930/1501–1524): Persian Edition and Introduction", then aligned at the page-level with the manuscript images by the authors of this paper.

Link:

License: permission granted for research use

Ground Truth data for printed Devanagari (Merkel-Hilf, 2022)

Description:

Link: <https://heidata.uni-heidelberg.de/dataset.xhtml?persistentId=doi:10.11588/data/EG0KEI>

License: CC BY 4.0

Les Papiers Barye (Claass et al., 2021)

Description: a collection of correspondence and documents relating to the activity of the sculptor and painter Antoine-Louis Barye (1795-1875). These documents are kept in the Antoine-Louis Barye archive of the Library of the Institut National d'Histoire de l'Art, Jacques Doucet collections. The collection is composed of 350 documents (918 pages) and organized into 6 collections:

Correspondence during the lifetime of Antoine-Louis Barye
Anonymous documentation on Antoine-Louis Barye
Documents relating to Barye's activity during his lifetime
Biographical documentation on Antoine-Louis Barye
Documents relating to the trade and exhibition of Barye's works after his death
Iconographic documentation

Link: <https://gitlab.inha.fr/snr/LesPapiersBarye>

License: CC-BY 4.0

Padeov-Bible Handwriting Ground Truth (Michalcová et al., 2022)

Description: the Padeov Bible (Vienna, Austrian National Library, shelfmark Cod. 1175, 1432–1435), the bible of the third redaction of the Old Czech Bible translation

Link: <https://zenodo.org/records/7467034#.Y6LQZBWZM2w>

License: CC BY 4.0

TRANSCRIPTIONS OF THE INTERNATIONAL CONGRESS OF ETHNOGRAPHIC SCIENCES (PARIS, 1878) (Christensen et al., 2022)

Description: The International Congress of Ethnographic Sciences of 1878 took place on the occasion of the 1878 Universal Exposition in Paris. Edited in 1881 by the National Printing Office, the proceedings of this congress have been made available by the Digital Conservatory of Arts and Crafts. These proceedings allow us to revisit the often problematic beginnings of several scientific disciplines: ethnography, anthropology, and prehistory. They also bear witness to the development of archaeology, folklore studies, and French popular culture.

Link: https://github.com/PSL-Chartes-HTR-Students/TNAH-2021-Expositions_Universelles

License: CC-BY 4.0

Project for Transcribing the Active Correspondence of Hector Berlioz to His Sister Nanci Berlioz (Ceard et al., 2022)

Description: “the active correspondence of Hector Berlioz addressed to his sister Anne-Marguerite "Nanci" Berlioz”.

Link: <https://github.com/PSL-Chartes-HTR-Students/TNAH-2021-Projet-Correspondance-Berlioz>

License: CC-BY 4.0

TAPUSCORPUS (Chagué, 2021)

Description: Ground Truth for French 20th century typewritten documents collected on Gallica and Europeana

Link: <https://github.com/HTR-United/tapuscopus>

License: CC BY 4.0

TIMEUS CORPUS (Chagué et al., 2023a)

Description: Registers from the Prud'hommes

Court for the Textile Industry in Paris, january to june 1858 Registers from the Prud'hommes Court for the Textile Industry in Paris, january 1878

Link: <https://github.com/HTR-United/timeuscorpus>

License: CC BY 4.0

Notaires de Paris - Bronod (Limon-Bonnet et al., 2024)

Description: The lectaurep-bronod corpus consists of 100 pages from the register of Maître Louis Bronod (1719–1765), a notary in Paris from December 13, 1719, to July 23, 1765. The selected pages were written between the years 1742 and 1745.

Link: <https://github.com/HTR-United/lectaurep-bronod>

License: CC BY 4.0

Lectaurep-Mariages-et-Divorces, ground truth for the Registres des Contrats de Mariages et des Séparations et Divorces (French 19th century) (Rostaing et al., 2024)

Description:

Link: <https://github.com/HTR-United/lectaurep-mariages-et-divorces>

License: CC BY 4.0

Ground truth for various Parisian notary's repertoires (French 19th and 20th century) (LECTAUREP et al., 2021)

Description:

Link: <https://github.com/HTR-United/lectaurep-repertoires>

License: CC BY 4.0

(Alba et al., 2023)

Description:

Link: <https://github.com/HTRomance-Project/medieval-italian>

License: CC-BY-4.0

(Glaise et al., 2023)

Description:

Link: <https://github.com/HTRomance-Project/medieval-latin>

License: CC-BY 4.0

(Bordier et al., 2023)

Description:

Link: <https://github.com/HTRomance-Project/middle-ages-in-spain>

License: CC-BY 4.0

(Norindr et al., 2023)

Description:

Link: <https://github.com/HTRomance-Project/modern-roman-languages>

License: CC-BY 4.0

Digital Peraire (Chagué and Pérez, 2023)

Description: The documents are handwritten, dating from the second half of the 20th century, written in French with a blue ink pen or, more frequently, with a blue pencil. Occasional marginal notes appear in red.

Link: <https://github.com/alix-tz/peraire-ground-truth>

License: CC-BY 4.0

Ground truth for German newspaper Deutscher Reichsanzeiger und PreuSSischer Staatsanzeiger (1819–1945) (Kamlah et al., 2024)

Description: “Ground truth for German newspaper "Deutscher Reichsanzeiger und PreuSSischer Staatsanzeiger" (German Imperial Gazette and Prussian Official Gazette), which was published under changing names from 1819 to 1945”

Link: <https://github.com/UB-Mannheim/reichsanzeiger-gt>

License: CC0-1.0 license

Données du recensement du Valais (Dubois et al., 2024)

Description:

Link: <https://github.com/PonteIneptique/valais-recensement>

License: CC BY-NC 4.0.

Ground Truth for ÖNB, Cod. 3891 (Ainonen et al., 2022)

Description:

Link: <https://zenodo.org/record/7467249>

License: CC-BY 4.0

Stavronikita Monastery Greek handwritten document Collection (Pratikakis et al., 2021)

Description: It comprises manuscripts made of paper, written in the 16th century and its dimensions are 220X165 mm. The manuscript is embellished with epititles and red initials. Tachygraphical symbols and abbreviations are encountered in the manuscript as well. The collection is one of the oldest Stavronikita Monastery on Mount Athos. It is a parchment, four-gospel manuscript which has

been written between 1301 and 1350. It comprises 54 pages with dimensions that are approximately 250x185 mm. The script is elegant minuscule and the use of majuscule letters is rare. Tachygraphical symbols and abbreviations are encountered in the manuscript as well. Furthermore, the manuscript is enriched with chrysography, elegant epititles and initials. It comprises manuscripts made of paper, written at the end of the 15th century and its dimensions are 218X150 mm. In various pages, we find red initials and epititles which enrich the manuscripts decoration.

Link: <https://zenodo.org/records/5578251>
<https://zenodo.org/records/5595669> <https://zenodo.org/records/5578136>

License: CC-BY 4.0

(Vlachou-Efstathiou, 2024)

Description: Eutychès was a Latin Grammarian, active in the mid-6th century AD Constantinople, a disciple of Priscianus Caesariensis. His contribution to the field of grammar consists mainly of a treatise called *De uerbo*, addressed to his diligentissimum discipulorum Craterus. The *ars* elaborates on the criteria for the classification of the conjugation of verbs, in short, if the genitive of dico is dicas or dicis. Itself a rather arid development of principles that do not escape the general rule, nevertheless thanks to *De uerbo* 141 citations of classical authors have been passed down to us -some passages being otherwise lost-, mostly early Augustan poets. Adding to this, exhaustive lists of examples of verbs and derived nouns make an integral part of the treatise, that have been recycled from posterior *Grammatici*.

De uerbo was last edited at the end of the 19th century by Heinrich Keil in his fifth (out of seven) tome of the *Grammatici Latini*, alongside an enormous corpus of *artes* dating from the 4th to the 7th century. This monumental edition lacks methodological rigor and exhaustive paleographical research, is, for the most part, outdated, leaving a substantial amount of grammatical works in need of critical editions, including Eutychès' own contribution.

The manuscript tradition of the work, as reported by the excellent work of Colette Jeudy, is rather modest, as it is comprised of 32 manuscripts transmitting partially or entirely the text, and span from the end of the 8th to the 11th century AD. *De uerbo* enjoyed a fulgurous posterity, that inspired two medieval scholars, namely Sedulius Scottus

and Remigius of Auxerre to make a commentary rendition out of it, attesting to its importance in the educational milieu.

The typology of the manuscripts comes with two interesting features that pose at the same time the biggest difficulty for perspective editors (and, as we suppose, for Keil). Firstly, as many other works on grammar, actively used in the "medieval classroom" for the teaching of Latin, lists of examples are incorporated in the narrative between bits of theory, forming columns or/and tabular boards, a *mise-en-page* apt for the memorization of these exceptions. At first, this feature may not seem as big of a deal, but a thorough examination of these tables and of the posterior works that use them indicate that it yields important information for the understanding of the manuscript tradition (see especially Conduché 2019 "La mise en page d'Eutychès"). Secondly, several witnesses are richly annotated with multiple layers of interlinear and marginal notes, the progressive accumulation of which finds its climax in Remigius of Auxerre's commentary. Glosses in grammatical manuscripts have only recently (better late than ever!) started to interest editors (see Monella/Rossellini for Priscian and Evina Steinová for Isidore of Seville's *Etymologies*) most of them making use of digital tools. And that is because a typology as complex as that of grammatical glossed manuscripts cannot be easily handled and, better, be utilized for research purposes without the flexibility, absence of spatial constraints and ability of to handle of big and multilayered data that digital tools offer.

Link: <https://github.com/malamatenia/Eutyches>

License: Apache-2.0 license

(Gabay et al., 2023c)

Description:

Link: <https://github.com/FoNDUE-HTR/FONDUE-EN-PRINT-20>

License: Annotation is CC-BY. Images belong to the digital libraries.

(Gabay and Carrasco Luján, 2024)

Description:

Link: <https://github.com/FoNDUE-HTR/FONDUE-ES-PRINT-19>

License: CC BY 4.0

(Gabay et al., 2023a)

Description:

Link: <https://github.com/FoNDUE-HTR/FONDUE-FR-MSS-18>

License: CC BY 4.0

(Gabay, 2024)

Description:

Link: <https://github.com/FoNDUE-HTR/FONDUE-FR-PRINT-16>

License: CC BY 4.0

(Gabay and Dolto, 2024)

Description:

Link: <https://github.com/FoNDUE-HTR/FONDUE-FR-PRINT-20>

License: CC BY 4.0

(Gabay et al., 2023b)

Description:

Link: <https://github.com/FoNDUE-HTR/FONDUE-IT-PRINT-20>

License: CC BY 4.0

Swiss (German and French) art catalogues of the 19th c. (Joyeux-Prunel et al., 2023)

Description: Catalogues from Swiss art exhibitions from the Turnus period (1842-1961) published by the Société suisse des beaux-arts (SSBA). More information on the website of the SIK-ISEA.

Link: <https://github.com/FoNDUE-HTR/FONDUE-MLT-ART>

License: The catalogues are in the public domain, images are made available by SIK-ISEA and transcriptions are CC-BY.

Datasets for catalogs OCR and segmentation (Pradier et al., 2022)

Description: transcription of catalogues, printed mainly in the 19th c. but not only

Link: <https://github.com/FoNDUE-HTR/FONDUE-MLT-CAT>

License: CC BY 4.0

HTR data sets from medieval manuscripts (13th-14th c.) collecting "fabliaux" (Pinche and Pierreville, 2023)

Description: medieval manuscripts (13th-14th c.) collecting "fabliaux" in Old French

Link: <https://github.com/CIHAM-HTR/Fabliaux>

License: CC BY 4.0

(Carta et al., 2022)

Description: 374 Spanish chapbooks published by the same printer (José María Moreno) during the 19th century. 5 pliegos from the Varios corpus (32 pages)

Link: <https://github.com/DesenrollandoElCordel/FoNDUE-Spanish-chapbooks-Dataset>
<https://github.com/DesenrollandoElCordel/Moreno-OCR-files>

License: CC0-1.0 license

Ground Truth data for printed Malayalam (Tübingen University Library, 2023)

Description:

Link: <https://heidata.uni-heidelberg.de/dataset.xhtml?persistentId=doi:10.11588/data/L2KRZO>

License: CC BY 4.0

Ground-Truthed Data Set of Zenon Papyri for Handwritten Text Recognition (Marthot-Santaniello and Hodel, 2022)

Description:

Link: <https://zenodo.org/records/6565706>

License: CC BY 4.0

HN2021-ChateauChavigny (Pascual et al., 2022)

Description: Testimony of the looting of the Château de Chavigny in Lerné in Touraine in 1468 by the Protestants

Link: <https://github.com/PSL-Chartes-HTR-Students/HN2021-ChateauChavigny>

License: CC BY 4.0

HN2021-Kovalewsky-1893 (LEveque et al., 2022)

Description: Transcription of the chapter "On the regime of property among the Ossetians", in: Kovalewsky M. Contemporary custom and ancient law: Ossetian customary law, enlightened by comparative history. Paris, 1893

Link: <https://github.com/PSL-Chartes-HTR-Students/HN2021-Kovalewsky-1893>

License: CC BY 4.0

Memorials for Jane Lathrop Stanford (Guimarães et al., 2022)

Description:

Link: https://github.com/PSL-Chartes-HTR-Students/HN2021-Memorials_Jane_Lathrop_Stanford

License: CC BY 4.0

HN2021-OCR-Poesie-Corse (Sarbach-Pulicani et al., 2022)

Description: the collection of poems "Pon-tênôvu", written by Petru Rocca and published by the "Stamparia di a Muvra" in 1927 in the name of the Corsican Action Party.

Link: <https://github.com/PSL-Chartes-HTR-Students/HN2021-OCR-Poesie-Corse>

License: CC BY 4.0

Handwritten Paleographic Greek Text Recognition: A Century-Based Approach (Platanou et al., 2022)

Description: The Barocci collection consists of 244 volumes and it is the largest acquisition of the Bodleian collection. The dates of these manuscripts range from the 8th century AD to the 17th century AD. folios serve as characteristic examples of the writing style they represent. They include both Greek minuscule script and the cursive style of the minuscule script. In this way, our work involves examination of different styles and is not limited to one style of script.

Link: <https://github.com/vivianpl/hpgtr>

License: CC BY-NC-SA 3.0

15th century manuscript HTR data (Pinche et al., 2022)

Description: French manuscripts from the 15th century.

Link: <https://github.com/Gallicorpora/HTR-MSS-15e-Siecle>

License: CC0-1.0 license

(Humbel et al., 2024)

Description: handwriting of Hans Sloane (1660-1753) Sloanes Catalogue of Miscellanies (folio 3-152, recto and verso). The Catalogue of Miscellanies was chosen for the training data creation because it is known to be predominantly written by Sloane.

Link: <https://github.com/sloanelab-org/HTR-Model>

License: CC BY-NC-SA 4.0

The SETAF project (Solfrini et al., 2023)

Description: works published by Jean Michel, master printer in Geneva from 1538 to 1544, who bought Pierre de Vingle's typographical equipment. The texts are sixteenth-century French prints in Gothic characters and the list of texts with more details can be found in the CSV table of the

repository.

This repository contains OCR data from works published by Pierre de Vingle, master printer in Lyon from 1525 to 1532, in Geneva in 1532-1533 and in Neuchâtel in 1533-1535. The texts are sixteenth-century French prints in Gothic characters and the list of texts with more details can be found in the CSV table of the repository.

This repository contains OCR data from various editions of the Facts of Jesus Christ and the Pope. It is a singular work of religious polemic, the only illustrated book of the French-speaking Reformation preserved for the first half of the sixteenth century. It was published anonymously, in three successive editions, in [Neuchâtel], [Geneva] and [Lyon], between the years 1530 and 1560.

Link: <https://github.com/SETAFDH/HTR-SETAF-Jean-Michel> <https://github.com/SETAFDH/HTR-SETAF-Pierre-de-Vingle> <https://github.com/SETAFDH/HTR-SETAF-LesFaictzJCH>

License: CC BY 4.0.

(Gabay et al., 2022)

Description:

Link: <https://github.com/Gallicorpora/HTR-imprime-16e-siecle>

License: CC0-1.0 license

(Gabay et al., 2023e)

Description:

Link: <https://github.com/Gallicorpora/HTR-imprime-17e-siecle>

License: CC-BY 4.

(Gabay et al., 2023d)

Description:

Link: <https://github.com/Gallicorpora/HTR-imprime-18e-siecle>

License: CC-BY 4.

(Gabay et al., 2023f)

Description:

Link: <https://github.com/Gallicorpora/HTR-incunable-15e-siecle>

License: CC-BY 4.

JosephHookerHTR (Schaefer and Litvine, 2023)

Description: the correspondence of Joseph Dalton Hooker (1817-1911), primarily letters to William Turner Thiselton-Dyer (1843-1928) during the late-19th/early-20th century. Many tran-

scriptions in this dataset were generated by a small team of anonymous volunteers as part of the Joseph Hooker Correspondence Project based at Kew Gardens.

Link: <https://github.com/jschaefer738b/JosephHookerHTR.git>

License: CC BY 4.0

LiDi 1.0 Project (Agostini, 2024)

Description: 16th antiquarian Pirro Ligorio time: notBefore: '1568' notAfter: '1580'

hands: count: '1' precision: estimated

Link: <https://github.com/Giorgiaagostini/LiDi1.0-project>

License: CC-BY-SA 4.0

Liber (Aruta et al., 2023)

Description:

Link: <https://github.com/CIHAM-HTR/Liber>

License: CC-BY 4.0

Dresdner Hofdiarium 1665 (Mscr.Dresd.K.80) - 17th century Kurrent manuscript (Beckert, 2024)

Description:

Link: <https://zenodo.org/records/14356190>

License: CC BY-NC-SA 4.0

NuBIS OCR (Bibliothèque Interuniversitaire de la Sorbonne, 2024)

Description: a sample of 3 pages taken from 19 printed books, making a total of 57 pages. These books, in French and Latin, cover a period from 1602 to 1989, with roughly one book every 20 years. They are thus representative of the printed documents available in the digital library.

Link: <https://github.com/ksefil/NuBIS-OCR>

License: CC-BY 4.0

TNAH-2021-ArgusDesBrevets (Craene et al., 2022)

Description: The 1910 patent argus is presented in the form of a contemporary print, organized into sections grouping chronologically and then thematically the patents filed in France. This list and brief presentation of the patents is divided into two columns and presents standardized abbreviations.

Link: <https://github.com/PSL-Chartes-HTR-Students/TNAH-2021-ArgusDesBrevets>

License: CC-BY 4.0

TNAH-2021-DecameronFR (Biay et al., 2022)

Description: The project aims to create ground truth data for training HTR (Handwritten Text Recognition) models based on a French manuscript from the years 1430–1455: manuscript 5070 from the Bibliothèque de l'Arsenal (available on Gallica). This manuscript contains the French translation of Boccaccio's Decameron by Laurent de Premierfait. Our ground truth data covers the description of the plague in Florence, located in the prologue of the work.

Link: <https://github.com/PSL-Chartes-HTR-Students/TNAH-2021-DecameronFR>

License: CC-BY 4.0

Notre-Dame Project (Doat et al., 2022)

Description: the transcription of the daily journals from the year 1860, documenting the restoration work carried out between 1844 and 1865 at the Notre-Dame Cathedral in Paris under the direction of Eugène Viollet-le-Duc and Jean-Baptiste Lassus.

Link: <https://github.com/PSL-Chartes-HTR-Students/TNAH-2021-Projet-Notre-Dame>

License: CC-BY 4.0

TranscriboQuest 2024 Medieval Literary (Vandyck et al., 2024)

Description: The aim of this dataset was to contribute to underrepresented aspects of the manuscripts used in the CATMuS project. We opted to focus on medieval scientific documents that are damaged, in several different languages.

Link: <https://zenodo.org/records/13757440>

License: CC BY 4.0

Verard Corpus (Hoeben, 2024)

Description: Parts (10 pages) of Vérard's Editions princeps of the following texts:

Premier Volume de Tristan (1489) shelfmark: IFN-8600174, folios 12r-16v Second Volume de Tristan (1489) shelfmark: IFN-8600175, folios 12r-16v Premier Volume de Merlin [1498/1503] shelfmark: Res Y2 26, folios 12v-17r Second Volume de Merlin [1498/1503] shelfmark: Res Y2 26, folios 12v-17r Prophecies de Merlin [1498/1503] shelfmark: Res y2 27, folios 12r-16v Gyron le Gourtoys [1501/1503] shelfmark: RES-Fol-BI-

922, folios 12r-16v

Link: <https://github.com/LaurieHoeben/Verald-corpus>

License: Etalab Open License 2.0

EPARCHOS - Historical Greek handwritten document dataset (Papazoglou et al., 2020)

Description: The dataset originates from a Greek handwritten codex that dates from around 1500-1530. This is the subset of the codex British Museum Addit. 6791, written by two hands, one by Antonius Eparchos and the other by Camillos Zanettus (ff. 104r-174v) and delivers texts by Hierocles (In Aureum carmen), Matthaeus Blastares (Collectio alphabetica) and, notably, texts by Michael Psellos (De omnifaria doctrina). The writing delivers the most important abbreviations, logograms and conjunctions, which are cited in virtually every Greek minuscule handwritten codex from the years of the manuscript transliteration and the prevalence of the minuscule script (9th century) to the post-Byzantine years.

Link: <https://zenodo.org/records/4095301>

License: CC BY 4.0

Episearch HTR (Tommasi, 2024)

Description: “Diplomatic trascription of Giovanni Antonio Astoris letters to Ludovico Antonio Muratori by Tatiana Tommasi; images of Astoris letters to Muratori (from Internet Culturale)”

Link: <https://github.com/vedph/episearch-htr>

License: CC BY-SA 4.0

Ground Truth Dataset Medieval Greek Manuscripts – Codex heidelbergensis palatinus graecus 23 (Maxime et al., 2024)

Description:

Link: https://gitlab.huma-num.fr/ecrinum/anthologia/htr_cpgr23

License: CC-BY 4.0

The project iForal: Portuguese municipal charters in the Middle Ages: an historical and linguistic approach in the digital era (Alexandre et al., 2025)

Description:

Link: <https://github.com/Arch-W/iForal-Dataset>

License: CC-BY 4.0

Biblia (Stökl Ben Ezra et al., 2021)

Description: “Medieval Hebrew manuscripts from the Bibliothèque nationale de France (BnF, National Library of France) and the Biblioteca Apostolica Vaticana (BAV, Vatican Library)”

Link: <https://zenodo.org/records/5167263>

License: CC-BY-NC-SA 4.0

Transcription corpora for training HTR models for medieval manuscripts from the 12th to the 15th century. (Pinche, 2022)

Description:

Link: <https://github.com/HTR-United/cremma-medieval>

License: CC-BY 4.0

CREMMA HTR GT for medieval latin manuscripts (Clérice et al., 2022b)

Description:

Link: <https://github.com/HTR-United/CREMMA-Medieval-LAT>

License: CC-BY 4.0

CREMMA - A repository of 20th century manuscripts (Clérice and Chagué, 2023)

Description:

Link: <https://github.com/HTR-United/CREMMA-MSS-20>

License: CC-BY 4.0

CREMMA - A repository of 18th century manuscripts (Van Kote et al., 2024)

Description:

Link: <https://github.com/HTR-United/CREMMA-MSS-18>

License: CC-BY 4.0

CREMMA - A repository of 19th century manuscripts (Clérice et al., 2023)

Description:

Link: <https://github.com/HTR-United/CREMMA-MSS-19>

License: CC-BY 4.0

CREMMA - A repository of 16th century manuscripts (Mazoue et al., 2024)

Description:

Link: <https://github.com/HTR-United/CREMMA-MSS-16>

License: CC-BY 4.0

CREMMA - A repository of 17th century manuscripts (Clérice et al., 2022a)

Description:

Link: <https://github.com/HTR-United/CREMMA-MSS-17>

License: CC-BY 4.0

CREMMA Testament De Poilus (Chagué et al., 2023b)

Description: Each document (1 or more images sharing on the same side) corresponds to a hand. As these are the wills of soldiers who died for France during the First World War, all the documents were written between 1898 and 1918.

Link: <https://github.com/HTR-United/CREMMA-AN-TestamentsDePoilus>

License: CC-BY 4.0

CREMMA Early Modern Books (Clérice, 2021)

Description:

Link: <https://github.com/HTR-United/cremma-16-17-print>

License: CC-BY 4.0

DAHN Corpus (Chiffolleau, 2021)

Description: Ground Truth dataset for French 20th typewritten OCR

Link: https://github.com/HTR-United/dahn_corpus

License: CC-BY 4.0

Celestine Doniau-Danest (Chagué et al.)

Description: The dataset contains pages randomly selected from the digitization of the "Journal de Célestine Doniau-Danest sur les débuts de la Guerre 1914-1918" (Diary of Célestine Doniau-Danest on the beginning of the 1914–1918 war) The dataset features only one handwriting style, with little variation, for a text written between 1914 and 1915. The digitizations are double-page spreads.

Link: <https://github.com/alix-tz/dataset-celestine-doniau-danest>

License: CC-BY 4.0

Assisi, Fondo Antico presso la Biblioteca del Sacro Convento, 408 (Sacro Convento di San Francesco in Assisi, 2024)

Description:

Link: <https://www.internetculturale.it/it/1175/assisi-fondo-antico-del-sacro-convento-mediatheca-franciscana>

License: CC BY-NC-SA 4.0

La Correspondance Jacques Doucet - René

Jean (Cugy et al., 2022)

Description: This collection, donated by René-Jean in 1946, is one of the main sources on the relationship between Doucet and the man he hired as librarian on 2 June 1908. The letters and documents that make up the book evoke various episodes relating to the formation and development of the Art and Archaeology Library; they also provide more general information on the life, functioning and ideas of Jacques Doucet.

Link: <https://gitlab.inha.fr/snr/LaCorrespondanceDoucetReneJean>

License: Etalab Open License 2.0

The school of Salamanca (Salamanca School Project, 2025)

Description: “The digital collection of sources of the project "The School of Salamanca" comprises 107 works in total,”

Link: <https://www.salamanca.school/works.html>

License: CC BY 4.0

A Human-Annotated Dataset of Scanned Images and OCR Texts from Medieval Documents (Novotný et al., 2021)

Description: This is an open dataset of scanned images and OCR texts from 19th and 20th century letterpress reprints of documents from the Hussite era.

Link: <https://nlp.fi.muni.cz/trac/ahisto/wiki/OcrDataset>

License: CC0 1.0

Old Books Dataset (Barcha, 2025)

Description: “built with Project Gutenberg ebooks. They were selected among the following books:

-Betrayed Armenia, de Diana Agabeg Aparcar - The Boy Apprenticed to an Enchanter, de Padraic Colum -The Child of the Moat, de Stoughton Holborn -The Corset and the Crinoline, de W.B.L - Engraving of Lions, Tigers, Panthers, Leopards, Dogs, &C., de Thomas Landseer -Half-Hours with Highwaymen, de Charles G. Harper -Historical Sketches of Colonial Florida, de Richard L. Campbell -Horton Genealogy, de Geo. F. Horton -The Lusitania’s Last Voyage, de Charles E. Lauriat - Seat Weaving, de L. Day Perry”

Link: <https://github.com/PedroBarcha/old-books-dataset>

License: Project Gutenberg License for annota-

tions

An Historical Handwritten Arabic Dataset for Segmentation-Free Word Spotting - HADARA80P (Pantke et al., 2014)

Description:

Link: <https://cloud.tu-braunschweig.de/s/anM8gnrLC66aTAf>

License: permission granted for research use

SleukRith Set (Valy et al., 2017)

Description: the first dataset specifically created for Khmer palm leaf manuscripts Unlike the writing of most Latin languages where characters are sequenced from left to right, in Khmer writing, vowels are positioned either on the left, on the right, below, or above the consonant they are spelled with. Two or more consonants can also be merged together by transforming into different shapes called low-consonant or subscript form which are placed below the main consonant. These variations of how Khmer letters are formed produce multiple levels (sometimes more than three) of characters. Moreover, some writers tend to exaggerate their writing by elongating the upper or lower part of a character which makes it go far out of its main line, touch, or overlap with other characters from adjacent lines. In addition, the ambiguity of certain characters in Khmer alphabet is a big challenge as well for character recognition problem. Some groups of characters can only be distinguishable by a mere difference of a small hole or a short stroke. Some types of symbols contain multiple parts whose shapes are identical or very similar to other characters.

Link: <https://github.com/donavalys/SleukRith-Set>

License: permission granted for research use

annotationdb (Kölsch, 2025)

Description:

Link: https://tc11.cvc.uab.es/datasets/AnnotationDB_1

License: permission granted for research use

Ground Truth Model for Pracalit for Sanskrit and Newar MSS 16th to 19th C. (O'Neill and Hill, 2022)

Description:

Link: <https://heidata.uni-heidelberg.de/dataset.xhtml?persistentId=doi:10.11588/data/WI9184>

License: CC BY 4.0

ICDAR 2019 Historical Document Reading Challenge on Large Structured Chinese Family Records (ICDAR2019HDRC) (Saini et al., 2019)

Description: mainly written in Chinese traditional Han script. The document images have been taken from different books.

Link: https://tc11.cvc.uab.es/datasets/ICDAR2019HDRC_1 https://tc11.cvc.uab.es/datasets/ICDAR2019HDRC_1/gt_1_1

License: CC BY-NC-SA 3.0

(Lassner et al., 2021)

Description: historical prints from around 1830

Link: <https://github.com/millawell/ocr-d-ata>

License: CC BY 4.0

TariMa dataset (Noémie et al., 2022)

Description: This new dataset is focused on specific contents or vocabulary lacking in previous RASAM models.

Link: <https://github.com/calfa-co/tarima>

License: Apache-2.0 license

(Teklia, 2025b)

Description: The Finlam dataset includes 149 French newspapers from the 19th to 20th centuries.

Link: <https://huggingface.co/datasets/Teklia/Newspapers-finlam>

License: MIT License

(Teklia, 2025a)

Description: The Finlam La Liberté dataset includes 1500 issues from La Liberté, a French newspaper, from 1925 to 1928.

Link: <https://huggingface.co/datasets/Teklia/Newspapers-finlam-La-Liberte>

License: MIT License

The Osman Agha dataset (osm, 2025)

Description: The Osman Agha dataset consists of 243 images of an 18th-century Ottoman manuscript and its Arabic-script transcription, aligned at the page level by the authors of this paper. The manuscript contains the memoirs of Osman Agha, who was a prisoner of war in Austria between 1688 and 1699. The memoir was completed on May 18, 1724. The manuscript survives as a single authorial copy, currently held at the

British Library (MS. Or. 3213). The manuscript is commonly known in English as Prisoner of the Infidels: The Memoirs of Osman Agha of Timioara, the title of its 2021 English translation (Casale 2021). A transcription of the manuscript was published in 1980 (Kreutel, 1980).

Link:

License: permission granted for research use

American Stories (Dell et al., 2023)

Description: “newspaper scans from Library of Congress’s Chronicling America collection”

Link: https://huggingface.co/datasets/dell-research-harvard/AmericanStoriesTraining/blob/main/gold_data/gold_hand_transcription.json

License: Apache license 2.0