# Data Report — OCR Arabic Manuscript


Prepared by

Hussam Aldossary

Supervised by

Prof. Samer Samarah

# Table of Contents


**1. Introduction** ..... **4**

1.1 Purpose ..... 4

1.2 Scope ..... 4

**2. Dataset Analysis and Preparation** ..... **6**

2.1 AHDQ ..... 6

2.1.1 Description ..... 6

2.1.2 Features ..... 7

2.1.3 Analysis ..... 8

2.1.4 Results ..... 12

2.2 Omar-al-Saleh ..... 14

2.2.1 Description ..... 14

2.2.2 Features ..... 15

2.2.3 Analysis ..... 17

2.2.4 Results ..... 19

2.3 AraMs-28k-HTR ..... 22

2.3.1 Description ..... 22

2.3.2 Features ..... 24

2.3.3 Analysis ..... 27

2.3.4 Results ..... 30

2.4 OpenITI-Makhzan ..... 33

2.4.1 Description ..... 33

2.4.2 Features ..... 35

2.4.3 Analysis ..... 38

2.4.4 Results ..... 41

2.5 AMIDDA ..... 44

2.5.1 Description ..... 44

2.5.2 Features ..... 52

2.5.3 Analysis ..... 55

2.5.4 Results ..... 61

2.6 KHATT ..... 66

2.6.1 Description ..... 66

2.6.2 Features ..... 68

2.6.3 Analysis ..... 69

2.6.4 Results ..... 75

2.7 Muharaf ..... 78
2.7.1 Description ..... 78
2.7.2 Features ..... 80
2.7.3 Analysis ..... 81
2.7.4 Results ..... 84
**3. Summary Across the Subsets ..... 87**
3.1 The seven subsets side by side ..... 87
3.2 Main findings ..... 88
3.3 Licenses and redistribution ..... 89
3.4 Splits and how to reproduce them ..... 91
3.5 Limits of this report ..... 92
**4. References ..... 93**

# 1. Introduction

## 1.1 Purpose

This report documents how the data for our Arabic manuscript project was gathered, analysed and prepared. For each dataset it records what the source contains, which attributes it has, what the exploratory analysis found, and what the dataset looks like after preparation.

The report has three readers. The supervisor and examiners can check that every preparation step has a reason. The team members who train and evaluate the model can see exactly what each published subset contains. Anyone who reuses the published collection can repeat or question our choices.

It is the data report listed among the data module's deliverables in the project plan, next to the unified dataset, the build scripts and the split manifests.

## 1.2 Scope

The report covers the seven subsets of the Arabic Manuscript Collection that we published on Hugging Face. All figures refer to the version of 12 September 2026 (revision `809d854`).

| Subset | Unit | Size | How it is published |
|---|---|---|---|
| AHDQ | Page | 642 pages | Converted; our own split (90 / 5 / 5) |
| AMIDDA | Line | 66,779 lines | Unchanged upstream Parquet files |
| AraMs-28k-HTR | Line | 27,969 lines | Converted; official split by book |
| KHATT | Line | 11,398 lines | Converted; test set made by us |
| Muharaf | Line | 24,495 lines | Converted; official split |
| Omar-al-Saleh | Line | 18,057 lines | Converted; validation set made by us |
| OpenITI-Makhzan | Line on a page image | 23,079 lines on 1,497 pages | Whole pages with line coordinates |

Each subset has its own section with the same four parts:

1. **Description:** the source, its content and license, its original state, and its place in the collection.
2. **Features:** the files, the attributes, and any null or invalid values.
3. **Analysis:** what the exploratory analysis of the labels and the images shows.
4. **Results:** the subset after preparation, the decisions we took, and the limitations that remain.

Model training and evaluation are outside the scope. The report covers the data only: where it comes from, how it was prepared and split, and what the analysis found. Section 3 compares the seven subsets side by side, and section 4 lists the references.

In this report, "our card" and "the collection card" both mean the dataset card of the Arabic Manuscript Collection.

# 2. Dataset Analysis and Preparation

## 2.1 AHDQ

### 2.1.1 Description (original state, place in the collection)

AHDQ (Arabic Handwriting Dataset from Quran) contains 642 binarised page images of handwritten Quran text, each paired with its full transcription. The pages come from the Wolof community in Senegal, where writing the Quran by hand from memory is a long tradition. They cover six handwriting styles.

| Property | Value |
|---|---|
| Source | Kaggle, published by XIDMA_AI, last updated 27 September 2024 |
| Content | Handwritten Quranic verses, several lines per image |
| Unit | Full page, not a line crop |
| Size | 642 images, about 112 MB |
| Label file | One CSV with two columns, `file_name` and `text` |
| Text | Unicode Arabic, fully diacritised |
| Image format | Binarised PNG |
| License | MIT |
| Upstream split | None |

The prefix of each file name gives the handwriting style of the page:

| Prefix | Style |
|---|---|
| `dmn_` | Darou Minan |
| `sal_` | Serigne Amidoun Lo |
| `sas_` | Serigne Abdoulahat Souhaibou |

| Prefix | Style |
|---|---|
| sbi_ | Serigne Bousso Imaam |
| sdc_ | Serigne Djeumb Cisse |
| sfm_ | Serigne Fallou Mbacke |

**In the collection.** AHDQ is published as 642 whole-page images with their text, in a split of our own: 579 pages for training, 31 for validation and 32 for test, stratified by style.

### 2.1.2 Features (files, attributes, null and invalid values)

The published subset has two attributes per row, and neither has a missing value. One row describes one page.

**Files**

| File | Content |
|---|---|
| images.zip | The page images, in images/train/, images/val/ and images/test/ |
| train.csv | 579 rows |
| val.csv | 31 rows |
| test.csv | 32 rows |
| labels.csv | The original label file for all 642 pages, with the columns file_name and text |

**Attributes**

| Attribute | Type | Meaning | Example |
|---|---|---|---|
| image_path | String | File name of the page image inside images/<split>/ | dmn_132.png |
| text | String | Transcription of the whole page, in Unicode Arabic with diacritics | سُورَةُ النِّسَاءِ يَوْمَئِذٍ يَوَدُّ الَّذِينَ كَفَرُوا … |

- **image_path** is always 11 characters long: a three-letter style code, an underscore, a three-digit page number and .png. It holds no folder name, because the name of the CSV file already gives the split.
- **text** is one string for the whole page. It records everything written on the page, not only the verses. The dmn labels we sampled all start with the surah name from the top of the page, and surah headings such as the verse count appear inside the text. The Quranic pause marks are kept.
- **Handwriting style** is not a column. It is read from the first three letters of image_path.

**Null and invalid values**

| Check | Train | Validation | Test |
|---|---|---|---|
| Rows | 579 | 31 | 32 |
| Null image_path | 0 | 0 | 0 |
| Null text | 0 | 0 | 0 |
| image_path not 11 characters long | 0 | 0 | 0 |
| Shortest text (characters) | 313 | 771 | 573 |

No row is empty, and no file name appears in both the validation and the test split. Two labels have a damaged first letter. In dmn_003 the first letter of the basmala is missing. In dmn_130 the first letter of the page stands before the surah name instead of after it. Both were found in the first 100 training labels.

One pair of pages carries the same label. dmn_006 (validation) and dmn_007 (training) start and end with the same words. Either the page was scanned twice or one of the two labels is wrong.

### 2.1.3 Analysis (split sizes, handwriting styles, text content per style, text length, diacritics and special characters, images)

AHDQ is small and uneven across styles, and its pages are long and fully diacritised. The three splits have the same mix of styles and almost the same text length.

**Split sizes**

| Split | Pages | Share of pages | Characters of text |
|---|---|---|---|
| Train | 579 | 90.2% | 653,009 |
| Validation | 31 | 4.8% | 35,178 |
| Test | 32 | 5.0% | 35,977 |
| Total | 642 | 100% | 724,164 |

**Handwriting styles**

| Style | Train | Validation | Test | Total | Share of pages |
|---|---|---|---|---|---|
| dmn | 136 | 7 | 8 | 151 | 23.5% |
| sal | 96 | 5 | 5 | 106 | 16.5% |
| sas | 105 | 6 | 6 | 117 | 18.2% |
| sbi | 74 | 4 | 4 | 82 | 12.8% |
| sdc | 133 | 7 | 7 | 147 | 22.9% |
| sfm | 35 | 2 | 2 | 39 | 6.1% |
| Total | 579 | 31 | 32 | 642 | 100% |

Every style appears in every split, and each style keeps between 89.7% and 90.6% of its pages in training. The styles are not balanced: dmn has 3.9 times as many pages as sfm, and dmn and sdc together hold 46% of all pages. Validation and test have only 2 to 8 pages per style, so any figure per style is only indicative.

**Text content per style**

In the sampled pages, each style covers its own part of the Quran, in order. The same passage is therefore not written in two styles.

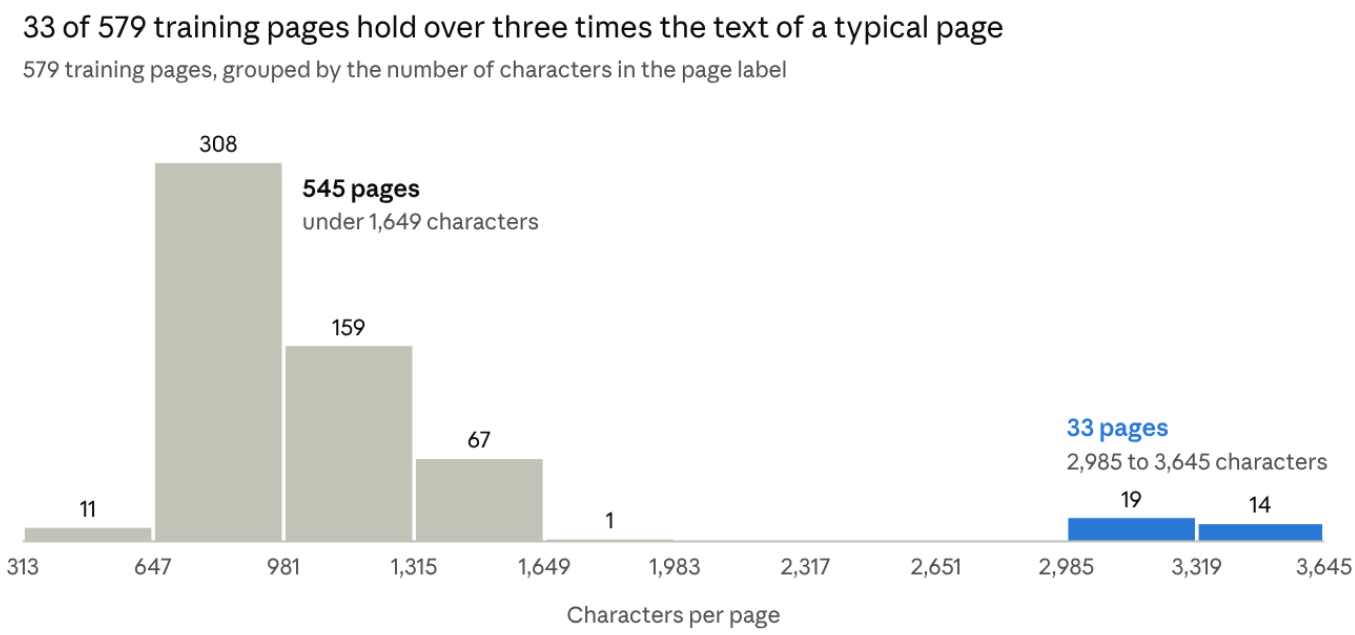
| Style | Pages sampled | Surahs seen |
|---|---|---|
| dmn | 115 | 1 to 4 (Al-Fatiha to An-Nisa) |
| sas | 12 | 5 to 9 (Al-Ma'idah to At-Tawbah) |
| sal | 10 | 10 to 18 (Yunus to Al-Kahf) |
| sbi | 8 | 20 to 27 (Ta-Ha to An-Naml) |
| sfm | 4 | 29 to 35 (Al-Ankabut to Fatir) |
| sdc | 14 | 46 to 98 (Al-Ahqaf to Al-Bayyinah) |

The sample is the 63 validation and test pages and the first 100 training pages. Style and content are tied together: a difference between two styles can come from the text as well as from the handwriting.

**Text length**

| Split | Shortest page | Median | Mean | Longest page |
|---|---|---|---|---|
| Train | 313 | 934 | 1,128 | 3,645 |
| Validation | 771 | 951 | 1,135 | 3,107 |
| Test | 573 | 944.5 | 1,124 | 3,323 |

Lengths are in characters per page, with diacritics and spaces counted. A typical page holds about 940 characters. That is 13 to 20 times the median line of the line-level subsets (47 to 70 characters).




Hugging Face dataset statistics, AHDQ train split, revision 809d854 · 579 pages in 10 bins

Validation and test show the same two groups, with 2 long pages each. The long pages point to the `sfm` style. It has 35, 2 and 2 pages in the three splits, close to the 33, 2 and 2 long pages, and it is written much more densely.

For example, `sfm_141` starts in verse 46 of Al-Ankabut and the next page starts in verse 5 of Ar-Rum, 28 verses later.

### Diacritics and special characters

All labels are diacritised. They also contain Quranic pause marks (ۖ ۗ ۚ ۘ ۙ ۛ ۜ) and the superscript alef (ٰ).

### Images

The pages are binarised PNG files, about 112 MB in all. Their pixel size was not measured (section 3.5). The published subset keeps the images as the source gives them, and it holds whole pages only.

### 2.1.4 Results (before and after, decisions, limitations that remain)

After preparation, AHDQ has the same 642 pages and the same text as the source. It now follows the common layout of the collection and has a split of its own.

**Before and after**

| Aspect | Source (Kaggle) | Prepared subset |
|---|---|---|
| Label files | One file, labels.csv | train.csv, val.csv and test.csv; the original labels.csv is kept |
| Label columns | file_name, text | image_path, text |
| Split | None | 579 / 31 / 32 pages (90 / 5 / 5%), stratified by style |
| Images | 642 PNG pages | 642 PNG pages in one images.zip, with one folder per split |
| Text | Arabic with diacritics | Unchanged |
| Rows | 642 | 642 |

**Decisions**

| Decision | Reason | Alternative |
|---|---|---|
| Rename file_name to image_path and write one CSV per split | One label format for all converted subsets, so they can be used together | Keep the source file and write a loader only for AHDQ |
| Create our own split, 90 / 5 / 5 | The source has no split | Use all pages for one purpose only |
| Stratify the split by style | All six styles appear in every split | A random split, which could leave sfm (39 pages) out of validation or test |

| Decision | Reason | Alternative |
|---|---|---|
| Keep whole pages instead of line crops | We cut the pages into lines for the evaluation and it worked, but the line labels are missing, and making them may take a lot of time | Cut every page into lines before publishing |
| Keep the text exactly as in the source | Project rule: keep the original text and normalise only when scoring | Strip the diacritics to match the other subsets |
| Ship the images as one zip file | To unify the data: one layout for all converted subsets | Keep loose image files in the repository |

**Limitations that remain**

- **Page-level only.** A line recogniser needs segmentation first.
- **Fully diacritised text.** Among the other converted subsets, KHATT and AraMs-28k-HTR have no diacritics, about 3% of the Omar-al-Saleh lines have some, and about a third of the Muharaf lines we read have some, so AHDQ follows a different labelling convention.
- **Small validation and test sets.** With 31 and 32 pages, and 2 to 8 pages per style, any figure per style is only indicative.
- **Image problems.** `sdc` includes scans tilted by 3 to 4 degrees, and `sas` has decorated pages whose medallions produce false line crops.
- **Label defects.** Two labels have a damaged first letter, and one pair of pages has the same label.
- **One text domain.** All pages are Quran text, and each style covers a different part of it.

## 2.2 Omar-al-Saleh

### 2.2.1 Description (original state, place in the collection)

Omar-al-Saleh is a set of line images cut from the handwritten memoirs of Omar Al-Saleh, each with an expert-checked transcription. It was released for the NAKBA NLP 2026 shared task.

**Original state**

Before any preprocessing, Omar-al-Saleh is a single gated Hugging Face dataset of three Parquet files, one per split. The table below describes it as the shared-task organisers published it.

| Property | Value |
|---|---|
| Source | U4RASD/omar-al-saleh-manuscripts-segments on Hugging Face (Doha Institute), published 8 April 2026 |
| Content | Arabic memoirs written by one person, 1951 to 1965 |
| Full collection | 16 documents, about 6,395 pages and 1.6 million words |
| Unit | Line crop |
| Size at source | 20,735 lines in three splits |
| Format at source | Parquet with the image inside each row, about 208 MB |
| Text | Unicode Arabic, about 3% of lines with diacritics |
| License | CC BY 4.0, with access after accepting the terms |
| Shared task | NAKBA NLP 2026, Arabic handwritten manuscript understanding |

The three files hold the lines of the three splits:

| File | Split | Lines | Size |
|---|---|---|---|
| data/train-00000-of-00001.parquet | train | 15,969 | 125.5 MB |
| data/test-00000-of-00001.parquet | test | 2,095 | 16.6 MB |
| data/blind_test-00000-of-00001.parquet | blind_test | 2,671 | 65.9 MB |

Every row holds the line image, its transcription and the original file name, and every line in all three splits is transcribed. The lines were cut from the manuscript pages and the transcriptions were checked by experts for the shared task. A blind-test row takes about 25 KB against about 8 KB in the other two splits, which points to a different image format or size.

Three things that our project needs are missing from the source:

- **No validation split.** Only train, test and blind test exist.
- **No document or page column.** Apart from the image and the text, a row has only `filename`.
- **No open download.** Access requires logging in and agreeing to share contact information.

The source asks to be cited through the NAKBA NLP 2026 shared-task paper (Hamoud et al., 2026).

**In the collection.** Omar-al-Saleh is published as 18,057 line images: 15,164 for training, 798 for validation and 2,095 for test. The training and test splits are the official ones. The validation lines are 5% of the official training split, drawn at random by us. The blind test of the source is not included.

### 2.2.2 Features (files, attributes, null and invalid values)

The published subset keeps two attributes per row, and neither has a missing value. One row describes one line of handwriting.

**Files**

| File | Content |
|---|---|
| images.zip | The line images, in images/train/, images/val/ and images/test/ |
| train.csv | 15,164 rows |
| val.csv | 798 rows |
| test.csv | 2,095 rows |

**Attributes**

The source has three attributes: image (the cropped line image, JPG or PNG), text (the Arabic transcription) and filename (the original file name). Our copy has two:

| Attribute | Type | Meaning | Example |
|---|---|---|---|
| image_path | String | File name of the line image inside images/<split>/ | 013105.png |
| text | String | Transcription of one line, in Unicode Arabic | قطعت واتلفت الثلوج اسلاك الكهرباء وأصلحوها سريعاً. |

- **image_path** is always 10 characters long: a six-digit number and .png. The first 25 test rows are numbered 000000.png to 000024.png in row order. Training and validation use the same kind of numbers in shuffled order, and validation includes 000078.png. Test numbering therefore very likely repeats names used in the other two folders. Until this is checked on the files, a file should be identified by its split and its name together.
- **text** holds one line as the source gives it. It is not trimmed or normalised.
- **filename** from the source is not in our copy, and no other column identifies the document or the page of a line.

**Null and invalid values**

| Check | Train | Validation | Test |
|---|---|---|---|
| Rows | 15,164 | 798 | 2,095 |
| Null image_path | 0 | 0 | 0 |
| Null text | 0 | 0 | 0 |
| image_path not 10 characters long | 0 | 0 | 0 |
| Shortest text (characters) | 1 | 1 | 1 |

No value is missing. Three points about the values still need a full count on the files:

- **Spaces at the edges.** Some labels start or end with a space. In a sample of 75 labels, 4 did.
- **Line breaks.** A handful of labels contain a line break inside the text, so the CSV files need a real CSV parser.
- **Seven lines fewer than the source.** Train and validation together have 15,962 lines, against 15,969 in the source train split. The dataset card says the gap comes from the source. The seven lines are not identified.

### 2.2.3 Analysis (split sizes, line length, what the short lines are, digits, script and diacritics, images)

Omar-al-Saleh has 18,057 lines and about one million characters of text. The three splits have the same line-length profile, and about a fifth of the lines in each split are very short.

**Split sizes**

| Split | Lines | Share of lines | Characters of text |
|---|---|---|---|
| Train | 15,164 | 84.0% | 860,871 |
| Validation | 798 | 4.4% | 45,573 |

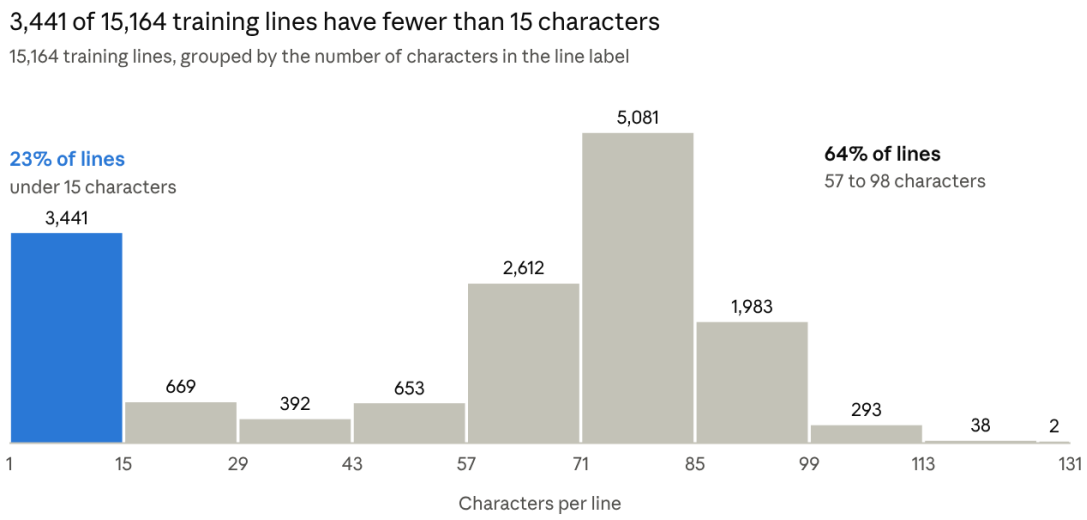
| Split | Lines | Share of lines | Characters of text |
|---|---|---|---|
| Test | 2,095 | 11.6% | 120,311 |
| Total | 18,057 | 100% | 1,026,755 |

The validation set is 5.0% of the 15,962 lines that came from the source train split.

**Line length**

| Split | Shortest line | Median | Mean | Longest line |
|---|---|---|---|---|
| Train | 1 | 70 | 56.8 | 131 |
| Validation | 1 | 70 | 57.1 | 128 |
| Test | 1 | 70 | 57.4 | 125 |

Lengths are in characters per line, with spaces counted. The median is 70 in all three splits, and the means differ by less than one character.




Hugging Face dataset statistics, Omar-al-Saleh train split, revision 809d854 · 15,164 lines in 10 bins

Validation and test have the same shape: 20.3% and 20.2% of their lines are shorter than 14 characters.

**What the short lines are**

The short lines are mostly calendar items such as day numbers, dates, and month or weekday names. They are not running text. In a sample of 75 labels, 21 were shorter than 15 characters:

| Kind | Examples | Lines in the sample |
|---|---|---|
| Numbers only | ٢٣, 17, ١٩-٣٤٦ | 5 |
| Dates, months and weekdays in Arabic | شباط (فبراير), ١٠ رمضان, الاربعاء | 6 |
| Calendar words in Latin script | JANUARY, FRIDAY, 4Ragab | 4 |
| Short notes and headings | برد شديد جدا, سد الفرات, العنوان | 6 |

The sample is the first 25 rows of each split. Two things follow from it. On a two-character line, one wrong character is a 50% error, so these lines weigh much more in mean line CER than in corpus CER. Words such as JANUARY and FRIDAY also suggest printed diary pages, and whether these lines are printed or handwritten has to be checked on the images.

**Digits, script and diacritics**

The labels mix Eastern Arabic digits (٢٣) with Western digits (17) and contain some Latin-script words.

About 3% of the lines carry diacritics.

**Images**

The source stores each line image as a JPG or PNG file inside its Parquet row. A train or test row takes about 8 KB and a blind-test row about 25 KB, which points to a different image format or size in the blind test. Our copy writes every line as a PNG file named with a six-digit number (section 2.2.2). The pixel size of the lines was not measured (section 3.5).

### 2.2.4 Results (before and after, decisions, limitations that remain)

After preparation, Omar-al-Saleh keeps the official test set of the source and its text unchanged. It gains a validation set of its own and leaves out the blind test.

**Before and after**

| Aspect | Source (Hugging Face) | Prepared subset |
|---|---|---|
| Storage | Three Parquet files with the images inside | PNG files in one images.zip, plus three CSV files |
| Columns | image, text, filename | image_path, text |
| Train | 15,969 lines | 15,164 lines |
| Validation | None | 798 lines, taken from the source train split |
| Test | 2,095 lines | 2,095 lines |
| Blind test | 2,671 lines | Not included |
| Total | 20,735 lines | 18,057 lines |
| Text | Arabic, as transcribed | Unchanged |

**Decisions**

| Decision | Reason | Alternative |
|---|---|---|
| Keep the official train and test splits | Project rule: use official splits where they exist | Build a new split by document |
| Take 5% of the source train split as validation (798 lines, drawn at random with seed 0) | The source has no validation split | Hold out whole documents, which this copy cannot do because it has no document ID |
| Leave out the blind test (2,671 lines) | So it cannot end up in training by accident | Keep it as a second test set |
| Store the images as files and the labels as CSV | One layout and label format for all converted subsets | Keep the source Parquet files |
| Do not keep the source column filename | To unify the data format: image_path and text only | Keep it as a link back to the original file |

| Decision | Reason | Alternative |
|---|---|---|
| Keep the text exactly as in the source | Project rule: keep the original text and normalise only when scoring | Trim the spaces and remove the line breaks in the labels |
| Ship the images as one zip file | To unify the data: one layout for all converted subsets | Keep loose image files in the repository |

**Limitations that remain**

- **One writer.** Every line comes from the same hand. A good score on this subset therefore says little about other hands.
- **Validation is not independent of training.** The validation lines were drawn at random from the training lines, so both sets hold lines from the same documents and very likely the same pages. Validation shows how well a model fits this collection, not how well it generalises.
- **No document or page identifier.** A split by document and results per document are not possible with this copy.
- **Many very short lines.** 22.7% of the training lines have fewer than 15 characters, mostly calendar items.
- **Public test labels.** The test split is published with its transcriptions, so a system that was trained on it would score unfairly well there.
- **Unresolved details.** Seven lines are missing against the source, some labels have spaces at the edges, and a handful contain a line break.

## 2.3 AraMs-28k-HTR

### 2.3.1 Description (original state, place in the collection)

AraMs-28k-HTR is the line-recognition form of AraMS-28k, a dataset of 14 historical Arabic books. It holds 27,969 main-text lines in three scripts, and its official split is by book.

**Original state**

Before any preprocessing, AraMS-28k is one zip file on Zenodo that holds full page images and line annotations, not line images. The table below describes it as its authors published it.

| Property | Value |
|---|---|
| Source | Zenodo record 22095333, version 1.0, published 25 August 2026 |
| Authors | Guechaoui, Zellagui, Chaib and Dhelim (ESI Sidi Bel Abbes and Dublin City University) |
| File | AraMS-28k.zip, 2.0 GB |
| Content | 14 books: 13 hand-copied manuscripts in Naskh, Ruq'ah and Maghrebi script, and 1 lithographed edition |
| Pages | 3,043 page images |
| Lines | 28,600 annotated lines: 27,971 main-text and 629 margin |
| Unit | Page image with line annotations |
| Official split | By book: 9 train, 2 validation and 3 test |
| License | CC BY-NC-SA 4.0 |
| Paper | arXiv 2608.26921 |
| Code | ArchaText/AraMS-28k-Dataset on GitHub |

The zip holds four folders:

| Folder | Content |
|---|---|
| images/ | The original page images, one folder per book |
| annotations/ | The line records, one file per book |
| schema/ | The JSON schema of a line record |
| splits/ | The book-level train, validation and test lists |

A line record carries much more than a transcription:

| Field | Meaning |
|---|---|
| line_uid, book_id, page_id, line_idx | Identity of the line and its position on the page |
| page_image | Path of the full page image |
| line_type | main or margin |
| split | train, val or test |
| text.gt_raw | Reference transcription from a scholarly edition, fully vocalised |
| text.gt_normalized | The same text without diacritics, which the authors recommend as training target |
| text.gemini_raw | Unchecked OCR output of a multimodal model, kept for reference only |
| text.confidence | Alignment score from 0 to 100 between the OCR output and the reference |
| geometry | Baseline, boundary polygon and bounding box of the line on the page |

| Field | Meaning |
|---|---|
| margin_anchor | For a margin line, the main-text line it inserts after. It is set for 189 of the 629 margin lines |
| review | Flags for edited, validated, deleted and page reviewed |

The authors built the dataset with their RefLAM pipeline. A Kraken model segmented the lines, a multimodal model produced an OCR draft of each page, the draft was aligned to a clean reference text of the book, and people reviewed the result. The books were checked at two levels:

- **Page-validated books (7 books, 548 pages).** Two reviewers inspected every line on every page.
- **Line-validated books (7 books, 2,495 pages).** Only lines whose OCR draft matched the reference perfectly were kept. The other lines were dropped.

Three things that our project needs are missing from the source:

- **No line images.** The lines have to be cropped from the pages with the geometry in the records.
- **No single label.** Each line has three texts, and one has to be chosen as the training target.
- **No commercial use.** The license allows non-commercial use only.

**In the collection.** AraMs-28k-HTR is published as 27,969 line images from 14 books: 19,739 lines from 9 books for training, 1,486 from 2 books for validation and 6,744 from 3 books for test. This is the official split by book. The 629 margin lines are not included.

### 2.3.2 Features (files, attributes, null and invalid values)

The published subset has two attributes per row and no missing value. A metadata file keeps the book, page and line of every row. One row describes one main-text line.

**Files**

| File | Content |
|---|---|
| images.zip | The line images, in images/train/, images/val/ and images/test/ |
| train.csv | 19,739 rows |
| val.csv | 1,486 rows |
| test.csv | 6,744 rows |
| metadata.csv | One row per line, with eight columns of provenance |
| split_used.json | The books of each split |
| dataset_stats.json | Lines, pages and books per split and per book |

**Attributes**

| Attribute | Type | Meaning | Example |
|---|---|---|---|
| image_path | String | File name of the line image inside images/<split>/ | book_03__book_03_page_001_line000.png |
| text | String | Transcription of one line, in Arabic letters and spaces | بسم الله الرحمن الرحيم وصلى الله على سيدنا محمد وآله وسلم |

- **image_path** is always 37 characters long. It names the book, the page and the line number, so the document of every line can be read from its file name.
- **text** holds only Arabic letters and spaces in the labels we sampled. This matches the cleaning rule of the authors' build script, which removes diacritics, tatweel, digits and punctuation and keeps the different forms of alef and hamza.
- **Row order** follows the books: consecutive rows are consecutive lines of a page. Line numbers on a page can have gaps, where the build skipped a margin, deleted or empty line.

metadata.csv has eight columns:

| Column | Meaning | Example |
|---|---|---|
| book_id | Book | book_03 |
| page_id | Page | book_03_page_001 |
| line_idx | Position of the line on the page, counted from 0 | 0 |
| line_type | main or margin | main |
| split | Split of the book | test |
| crop_src | How the line image was cut from the page: bounding box or polygon | bbox |
| gt_text | The same text as in the split CSV | قال الشيخ الفقيه الأجل الحسيب الأصيل |
| img_path | Path of the image when the file was built, before our split folders | ./AraMs-28k-HTR/images/book_03__book_03_page_001_line000.png |

**Null and invalid values**

| Check | Train | Validation | Test |
|---|---|---|---|
| Rows | 19,739 | 1,486 | 6,744 |
| Null image_path | 0 | 0 | 0 |
| Null text | 0 | 0 | 0 |
| image_path not 37 characters long | 0 | 0 | 0 |
| Shortest text (characters) | 2 | 1 | 2 |

No value is missing. One kind of label noise shows in the first 25 test labels: in 3 of them two words are joined without a space, for example بكرعبد. A full count needs the files.

### 2.3.3 Analysis (split sizes, books, script mix, validation level, line length, images)

AraMs-28k-HTR has 27,969 lines from 14 books and 3,043 pages. Its three splits hold different books, and they differ in script and in how the lines were checked.

**Split sizes**

| Split | Books | Pages | Lines | Share of lines | Characters of text |
|---|---|---|---|---|---|
| Train | 9 | 2,659 | 19,739 | 70.6% | 1,187,789 |
| Validation | 2 | 71 | 1,486 | 5.3% | 89,425 |
| Test | 3 | 313 | 6,744 | 24.1% | 412,307 |
| Total | 14 | 3,043 | 27,969 | 100% | 1,689,521 |

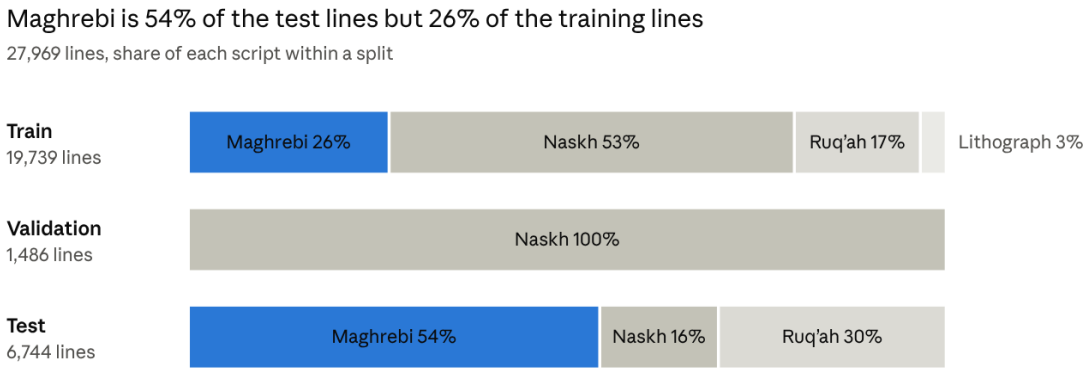
**Books**

| Book | Split | Script | Checked by | Pages | Lines | Lines per page |
|---|---|---|---|---|---|---|
| book_03 | Test | Maghrebi | Page | 172 | 3,661 | 21.3 |
| book_05 | Test | Ruq'ah | Page | 95 | 2,027 | 21.3 |
| book_09 | Test | Naskh | Page | 46 | 1,056 | 23.0 |
| book_06 | Validation | Naskh | Page | 41 | 874 | 21.3 |
| book_11 | Validation | Naskh | Page | 30 | 612 | 20.4 |
| book_10 | Train | Lithograph | Page | 13 | 674 | 51.8 |
| book_27 | Train | Naskh | Page | 151 | 2,532 | 16.8 |
| book_12 | Train | Naskh | Line | 144 | 2,015 | 14.0 |
| book_16 | Train | Naskh | Line | 392 | 2,276 | 5.8 |
| book_17 | Train | Naskh | Line | 584 | 3,731 | 6.4 |

| Book | Split | Script | Checked by | Pages | Lines | Lines per page |
|---|---|---|---|---|---|---|
| book_19 | Train | Maghrebi | Line | 122 | 666 | 5.5 |
| book_20 | Train | Maghrebi | Line | 312 | 2,163 | 6.9 |
| book_21 | Train | Ruq'ah | Line | 496 | 3,282 | 6.6 |
| book_24 | Train | Maghrebi | Line | 445 | 2,400 | 5.4 |

"Checked by" is the validation level of the source. Every line of a page-validated book was reviewed, while a line-validated book keeps only the lines that matched the reference perfectly. Script and validation level come from the authors' paper; pages and lines come from dataset_stats.json.

### Script mix




Lines per book from dataset_stats.json, script per book from the authors' paper · 27,969 lines in 14 books

Validation holds Naskh only, so it cannot show how a model does on Maghrebi or Ruq'ah. The test set has one book per script, and book_03 alone is 54% of it.

### Validation level

The seven line-validated books give 16,533 lines, which is 84% of the training lines. They keep about 6.6 lines per page, against about 21 in the page-validated books, because only perfectly matched lines survive. Validation and test are fully page-validated.

Training lines are therefore likely to be cleaner and more regular than test lines. The authors point out this bias themselves.

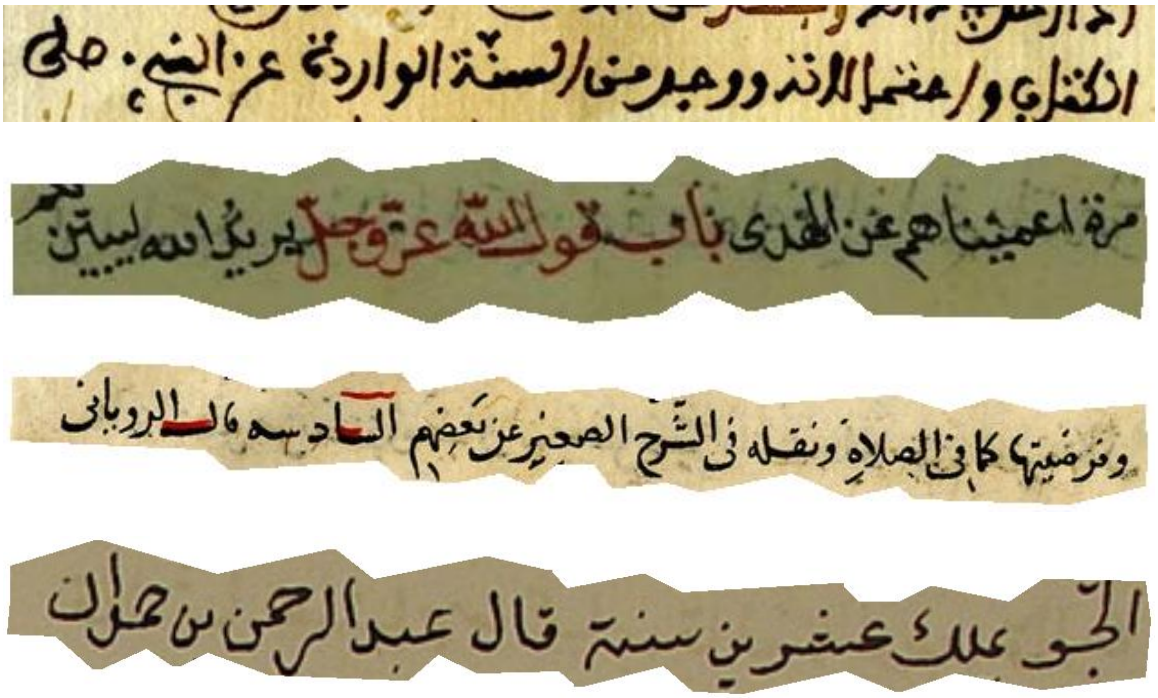
**Line length**

| Split | Shortest line | Median | Mean | Longest line |
|---|---|---|---|---|
| Train | 2 | 61 | 60.2 | 121 |
| Validation | 1 | 62 | 60.2 | 81 |
| Test | 2 | 58 | 61.1 | 105 |

Lengths are in characters per line, with spaces counted. The lines are regular: 73% of the training lines have 50 to 73 characters, and only 1.2% have fewer than 14. Omar-al-Saleh is very different here, with a fifth of its lines that short.

**Images**

Four sample lines published by the authors show what the line images look like:



| Sample | Book | Script | Split | Crop | Size in pixels |
|---|---|---|---|---|---|
| 1 | book_03 | Maghrebi | Test | Rectangle | 627 × 63 |
| 2 | book_17 | Naskh | Train | Polygon | 579 × 93 |

| Sample | Book | Script | Split | Crop | Size in pixels |
|---|---|---|---|---|---|
| 3 | book_05 | Ruq'ah | Test | Polygon | 636 × 86 |
| 4 | book_21 | Ruq'ah | Train | Polygon | 1,126 × 158 |

The line images are in colour and keep the tone of the paper and the red ink. A rectangle crop also shows parts of the lines above and below. A polygon crop paints everything outside the line white, so the line has an irregular white border. Size differs between books: sample 4 is about twice as wide as the others.

The samples come from the authors' repository and are under CC BY-NC-SA 4.0.

### 2.3.4 Results (before and after, decisions, limitations that remain)

After preparation, AraMs-28k-HTR is a set of 27,969 line images with one label each, in the common layout of the collection. The book-level split of the source is unchanged. The margin lines, the geometry and two of the three texts are no longer part of it.

**Before and after**

| Aspect | Source (Zenodo, AraMS-28k) | Prepared subset |
|---|---|---|
| Unit | Page image with line annotations | Line image |
| Files | One zip with page images, line records, schema and split lists | images.zip, three CSV files and three provenance files |
| Lines | 28,600: 27,971 main-text and 629 margin | 27,969 main-text lines |
| Label | Three texts per line | One text per line, letters and spaces only |
| Geometry | Baseline, polygon and bounding box | Not kept; only the crop remains |

| Aspect | Source (Zenodo, AraMS-28k) | Prepared subset |
|---|---|---|
| Margin anchors | 189 anchored margin lines | Not kept |
| Split | By book: 9 / 2 / 3 books | Same books: 19,739 / 1,486 / 6,744 lines |
| Line identity | line_uid, book_id, page_id, line_idx | In the file name and in metadata.csv |

**Decisions**

| Decision | Reason | Alternative |
|---|---|---|
| Use line crops, not pages with geometry | The project works at line level first | Keep the pages and the geometry for segmentation work |
| Take the line images as the authors' build script cuts them: a bounding-box crop, or a polygon crop with the outside painted white | To unify the data: the line images have the same form as in the other converted subsets | Cut the lines in another way, for example with more padding |
| Leave out the 629 margin lines | Margin notes act as layout and orientation noise | Add the margin lines as extra training lines |
| Keep the official book-level split | Project rule: use official splits where they exist. No book is in two splits | Re-split by page or by line |
| Move the images into one folder per split and write one CSV per split | One layout and label format for all converted subsets | Keep the builder's layout of text files and manifests |

| Decision | Reason | Alternative |
|---|---|---|
| Keep metadata.csv, split_used.json and dataset_stats.json | For later reference: they record the book, page and line of every row and the books of each split | Publish only the images and the three CSV files |
| Ship the images as one zip file | To unify the data: one layout for all converted subsets | Keep loose image files in the repository |

**Limitations that remain**

- **Script mix differs between splits.** Maghrebi is 54% of the test lines and 26% of the training lines, and validation is Naskh only.
- **Training lines are filtered.** 84% of them come from line-validated books, where only perfectly matched lines were kept. Test lines are not filtered in this way.
- **Labels differ from the other subsets.** AraMs labels have no diacritics, digits or punctuation, while Omar-al-Saleh and Muharaf keep them. Labels from different subsets therefore follow different conventions.
- **One book per script in the test set.** A score per script may reflect the book as much as the script. The authors note that book_09 has many faded lines.
- **Two lines fewer than the source.** The splits cover 27,969 of the 27,971 main-text lines. The authors held the other two out during quality control.
- **Label noise.** Some labels have two words joined without a space.
- **Non-commercial license.** CC BY-NC-SA 4.0 makes any use of the combined collection non-commercial.

## 2.4 OpenITI-Makhzan

### 2.4.1 Description (original state, place in the collection)

OpenITI-Makhzan is the OpenITI MAKHZAN release as a table with one row per transcribed line: 23,079 lines on 1,497 page images from 208 documents, in seven languages. It is the only subset that keeps whole pages together with the position of every line.

**Original state**

Before any preprocessing, OpenITI MAKHZAN is a Zenodo record with three files: a zip of page images with their transcription files, a metadata table and release notes. The tables below describe it as its authors published it.

| Property | Value |
|---|---|
| Source | Zenodo record 19861912, version 2026.1.2, published 28 April 2026 |
| Authors | The OpenITI team, at the University of Maryland, the Aga Khan University, UC San Diego, Northeastern University and Forman Christian College |
| Content | 208 documents: 200 manuscripts and 8 printed Urdu publications |
| Pages | 1,497 page images from 30 repositories |
| Period | 10th to 20th century |
| Languages | Arabic, Persian, Ottoman Turkish, Urdu, Malay, Javanese and Baleybelen |
| Scripts | About 20, led by naskh and nastaliq |
| Unit | Page image with line-level segmentation and transcription |
| Formats | Page images, ALTO XML for the lines and their text, and a metadata table |

| Property | Value |
|---|---|
| License | CC BY-NC-SA 4.0 |
| Paper | Journal of Open Humanities Data, 2026 |

The record holds three files:

| File | Size | Content |
|---|---|---|
| OpenITI-Makhzan_Data_2026-1-2.zip | 6.1 GB | The page images and their transcription files |
| OpenITI-Makhzan_Metadata_2026-1-2.tsv | 568.5 kB | The metadata table |
| OpenITI-Makhzan_ReleaseNotes_2026-1-2.pdf | 86.9 kB | The release notes |

The team produced the data between 2019 and 2025 in eScriptorium. Each page image was segmented into lines by hand, transcribed by hand, and reviewed by a second person before export. The pages come from three groups:

| Group | Pages | Documents | Design |
|---|---|---|---|
| ACDC training data | 400 | 70 manuscripts | Many copies of five well-known works, so the hand varies while the text stays the same |
| ACDC evaluation data | 123 | 80 manuscripts | A few pages each from 75 different works, for breadth |
| Research transcriptions | 974 | 50 manuscripts and 8 printed books | From single pages up to 315 pages of one document, mostly checked once |

The authors' transcription rules matter when this data is mixed with other sets. Diacritics are generally not transcribed, alif maqsura is written as ya, catchwords are left out, and nastaliq shorthands are written out in full. The authors add that not every transcriber applied the rules in the same way.

Three things that our project needs are missing from the source:

- **No line images.** A line exists only as coordinates on a page image.
- **No single table of lines.** The text and the coordinates sit in XML files, and the catalogue data in a separate table.
- **No commercial use.** The license allows non-commercial use only.

**In the collection.** OpenITI-Makhzan is published as whole page images, with a table of 23,079 lines (one row per line, with its coordinates on the page) in three splits by document: 98 documents for training, 54 for validation and 56 for test.

### 2.4.2 Features (files, attributes, null and invalid values)

The published subset has 22 attributes per row, against two in the other subsets. One row describes one line, and its image is the whole page that holds the line.

**Files**

| File | Content |
|---|---|
| images.zip | The page images, 6.1 GB |
| train.csv | 18,447 rows |
| val.csv | 2,343 rows |
| test.csv | 2,289 rows |

**Attributes**

| Group | Attribute | Type | Meaning | Values |
|---|---|---|---|---|
| Text | text | String | Transcription of the line | 1 to 199 characters |
| Image | image_file | String | File name of the page image | 2334_357981.tif |
| Image | page_id | Integer | Number of the page | 2,974 to 1,818,845 |
| Image | page_number | Integer | Page number inside the document | 0 to 589 |
| Geometry | bbox_x, bbox_y | Integer | Left and top edge of the line box, in pixels | 0 to 9,986 and 0 to 12,691 |
| Geometry | bbox_width, bbox_height | Integer | Width and height of the line box, in pixels | 11 to 6,088 and 9 to 2,452 |
| Geometry | polygon | String | Outline of the line as a list of points | 38 to 2,142 characters |

| Group | Attribute | Type | Meaning | Values |
|---|---|---|---|---|
| Provenance | document_id | Integer | Number of the document | 37 to 8,211 |
| Provenance | shelfmark | String | Library shelfmark of the manuscript | Wetzstein II 1763 |
| Provenance | line_id | String | Identifier of the line | Always 17 characters |
| Provenance | line_number | Integer | Number of the line on the page | 1 to 190 |
| Provenance | source_pdf | String | PDF file the page was taken from | leiden_or_1278.pdf |
| Catalogue | lang | String | Language code, or several codes | 12 values, such as ara |
| Catalogue | script | String | Script | 20 values, such as naskh |
| Catalogue | genre | String | Genre of the work | More than 50 values, such as Devotional works |
| Catalogue | form | String | Prose or poetry | 7 spellings, such as prose |
| Catalogue | document_type | String | Manuscript or printed publication | manuscript or publication (kitabat) |

| Group | Attribute | Type | Meaning | Values |
|---|---|---|---|---|
| Quality | transcriber | String | Who transcribed the line | Names of people |
| Quality | vetter | String | Who checked the transcription | Names of people |
| Quality | vetting_status | String | How often the line was checked | clean; once vetted or clean; twice vetted |

- **image_file** names a whole page, so many rows share one image. The name is built from document_id, page_id and the file type.
- **The geometry columns** are filled in every row. They are what a user needs to cut a line out of its page.
- **The page images** keep three file types: PNG, TIF and JPG.

**Null and invalid values**

Four attributes have missing values. The other 18 have none.

| Attribute | Train | Validation | Test | All | Share of rows |
|---|---|---|---|---|---|
| source_pdf | 14,115 | 1,760 | 1,952 | 17,827 | 77.2% |
| vetter | 4,379 | 1,603 | 1,196 | 7,178 | 31.1% |
| shelfmark | 822 | 140 | 93 | 1,055 | 4.6% |
| transcriber | 0 | 32 | 0 | 32 | 0.1% |

The missing shelfmarks are not an error. Their number equals the number of lines from printed Urdu publications in every split, and a printed book has no library shelfmark. The missing vetters are a real gap: every row has a vetting_status, but for 31% of the rows nobody is named.

Several catalogue columns hold the same value in more than one spelling:

- **form** has seven spellings for two values, among them prose, Prose and ProseProse.
- **genre** repeats genres in different case, such as Sufism and sufism, or Commentary and commentary.
- **lang** lists mixed languages in different orders, such as ota,ara,fas and ara,ota,fas.
- **transcriber** names one person in two ways.

The geometry has extreme values. A line box is 143 pixels high at the median in training, but the tallest is 2,452 pixels and the narrowest is 11 pixels wide. Such boxes are probably not ordinary text lines and should be checked on the images.

### 2.4.3 Analysis (split sizes, languages, scripts, documents, document type, form and checking, line length, images)

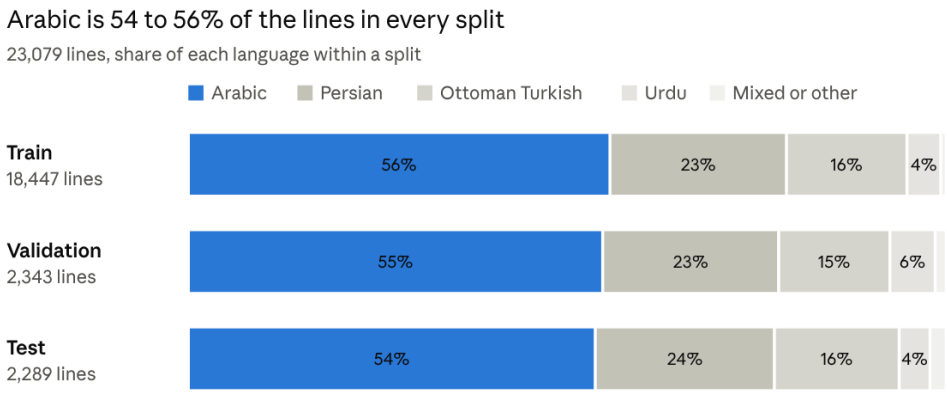
OpenITI-Makhzan has 23,079 lines on 1,497 pages from 208 documents. Arabic is just over half of it, six manuscripts hold half of the training lines, and the three splits have the same language mix.

**Split sizes**

| Split | Documents | Pages | Lines | Share of lines | Lines per document | Characters of text |
|---|---|---|---|---|---|---|
| Train | 98 | 1,185 | 18,447 | 79.9% | 188 | 932,430 |
| Validation | 54 | 165 | 2,343 | 10.2% | 43 | 125,276 |
| Test | 56 | 147 | 2,289 | 9.9% | 41 | 122,564 |
| Total | 208 | 1,497 | 23,079 | 100% | 111 | 1,180,270 |

Validation and test share no document and no page. A page holds 14 lines at the median in both, with at most 34 lines in validation and 64 in test.

## Languages




Dataset statistics for the lang column, revision 809d854 · 23,079 lines

Over all splits, Arabic has 12,773 lines, Persian 5,381, Ottoman Turkish 3,654 and Urdu 1,055. Another 216 lines carry two or three language codes, or Malay, Javanese or Baleybelen. For an Arabic recogniser, 45% of this subset is out of scope.

## Scripts

| Script | Train | Validation | Test | All | Share of lines |
|---|---|---|---|---|---|
| naskh | 12,668 | 1,420 | 1,244 | 15,332 | 66.4% |
| nastaliq | 3,283 | 521 | 550 | 4,354 | 18.9% |
| eurabic | 1,399 | 0 | 0 | 1,399 | 6.1% |
| maghribi | 288 | 183 | 124 | 595 | 2.6% |
| taliq | 300 | 61 | 205 | 566 | 2.5% |
| shikaste | 245 | 60 | 0 | 305 | 1.3% |
| 14 other labels | 264 | 98 | 166 | 528 | 2.3% |
| Total | 18,447 | 2,343 | 2,289 | 23,079 | 100% |

The script column has 20 labels, and 14 of them have fewer than 100 lines each. Seven labels never occur in training: kufic in validation, and bihari, riqa, jawi and

three mixed labels in test. They cover 21 validation lines and 132 test lines, which is 5.8% of the test set.

**Documents**

Training is concentrated in a few long documents. Six shelfmarks hold 9,376 lines, which is 50.8% of the training set, and the largest alone holds 2,386. Validation and test are made of many short documents: 13 to 142 lines in validation and 10 to 141 in test, with a median of about 35.

One transcriber produced 73% of all lines, so one person's habits shape most of the labels.

**Document type, form and checking**

| Property | Value | Lines | Share of lines |
|---|---|---|---|
| Document type | Manuscript | 22,024 | 95.4% |
| Document type | Printed Urdu publication | 1,055 | 4.6% |
| Form | Prose | 18,520 | 80.2% |
| Form | Poetry | 4,509 | 19.5% |
| Form | Mixed or other | 50 | 0.2% |
| Checking | Checked twice | 10,983 | 47.6% |
| Checking | Checked once | 12,096 | 52.4% |

The form counts merge the different spellings. The share of lines checked twice is 45.5% in training, 65.3% in validation and 46.1% in test.

**Line length**

| Split | Shortest line | Median | Mean | Longest line |
|---|---|---|---|---|
| Train | 1 | 48 | 50.5 | 199 |
| Validation | 1 | 52 | 53.5 | 161 |
| Test | 1 | 53 | 53.5 | 137 |

Lengths are in characters per line, with spaces counted. 87% of the training lines have 21 to 80 characters.

**Images**

The zip that holds the page images and their transcription files is 6.1 GB, about 4 MB per page. Of the 312 validation and test pages, 191 are PNG files, 68 are TIF and 53 are JPG; the file types of the training pages were not counted. The line boxes start as far as 9,986 pixels from the left and 12,691 from the top, so some pages are very large scans. At the median a line box is 1,018 pixels wide in training, 1,104 in validation and 1,084 in test, and 143 to 152 pixels high. The extremes are 11 to 6,088 pixels in width and 9 to 2,452 in height.

### 2.4.4 Results (before and after, decisions, limitations that remain)

After preparation, OpenITI-Makhzan is three CSV files with one row per line and 22 columns, next to the page images. No line was cropped and no catalogue value was normalised.

**Before and after**

| Aspect | Source (Zenodo) | Prepared subset |
|---|---|---|
| Lines and text | In ALTO XML files inside the data zip | One CSV row per line |
| Catalogue data | In a separate metadata table | Repeated on every line row |
| Images | Page images in a 6.1 GB zip | Page images in images.zip, 6.1 GB |
| Unit | Page | Line on a page |
| Size | 208 documents, 1,497 pages | 208 documents, 1,497 pages, 23,079 lines |
| Split | Not described in the data paper | 18,447 / 2,343 / 2,289 lines in 98 / 54 / 56 documents |

**Decisions**

| Decision | Reason | Alternative |
|---|---|---|
| Keep whole pages with line coordinates instead of line crops | It keeps the subset usable for line segmentation and layout work | Crop every line and use the two-column layout of the other subsets |
| Use the three splits as shipped | Project rule: use official splits where they exist. No document is in two splits | Re-split, for example by work |
| Leave the catalogue values as they are | Project rule: keep the source values and document their problems | Normalise the spelling of `form`, `genre` and `lang` |
| Keep all seven languages | The language of a line does not matter to the model, because it is not a classification model | Publish only the Arabic lines |

**Limitations that remain**

- **Only 55% Arabic.** The other lines are Persian, Ottoman Turkish, Urdu or mixed. An Arabic recogniser has to filter on `lang`.
- **Lines are not cropped.** Before a line recogniser can train on this subset, each line has to be cut from its page with the box or the polygon.
- **Splits are separate by document, not by work.** The source holds many copies of five well-known works, and the genres of those works appear in all three splits. A test line may repeat text that training shows in another hand.
- **Training is concentrated.** Six shelfmarks hold half of the training lines, and one transcriber produced 73% of all lines.
- **Rare scripts.** 14 of the 20 script labels have fewer than 100 lines, and 5.8% of the test lines are in a script that training never shows.
- **Uneven checking.** 52% of the lines were checked once only. The authors call their once-checked pages relatively less reliable.
- **Different transcription rules.** The source generally leaves out diacritics and writes alif maqsura as ya, which differs from the other subsets.

- **License note on the dataset card is out of date.** The card says the license is not verified. The Zenodo record states CC BY-NC-SA 4.0.

## 2.5 AMIDDA

### 2.5.1 Description (original state; the eight datasets inside AMIDDA: RASAM 1, RASAM 2, TARIMA, BAYBARS, ISKANDAR, RASM, KHATT, Muharaf; place in the collection)

AMIDDA is a ready-made collection of 66,779 line images from eight Arabic datasets, almost all of them handwriting. It is the only subset whose data we published without changing it. The name stands for Arabic Manuscript Images for Digital Deciphering and Analysis. The Calfa team put it on Hugging Face in May 2026 as a dataset for one general Arabic text-recognition model.

The eight datasets fall into three groups. Five datasets built by Calfa and its research partners hold 34,124 lines (51.1%). The British Library's RASM dataset holds 1,488 lines (2.2%). KHATT and Muharaf hold 31,167 lines (46.7%); the source card calls them modern cursive handwriting, and our collection already has both as separate subsets.

**Original state**

Before any step on our side, AMIDDA was one public Hugging Face repository of Parquet files, with the line images stored inside the rows.

| Property | Value at the source |
|---|---|
| Source | calfa-ai/AMIDDA on Hugging Face, open to everyone without a request |
| Publisher | Calfa (account calfa-ai) |
| Dates | Created 20 May 2026, last changed 11 September 2026 |
| Paper | Vidal-Gorène, Lucas, Salah and Decours-Perez, "Is a Generic Dataset and Foundation VLM for Arabic HTR Worth It? Lessons from AMIDDA", preprint of July 2026, presented at the HIP 2026 workshop of ICDAR 2026 (HAL hal-05693582) |

| Property | Value at the source |
|---|---|
| Unit | One cropped text-line image with its transcription |
| Size | 66,779 lines, 9.45 GB |
| Files | 20 Parquet files (16 train, 2 validation, 2 test) and one README.md |
| Columns | 9 |
| Splits | Train 53,420, validation 6,675, test 6,684 |
| License | "other": mixed, each dataset inside keeps its own license |
| Languages on the card | Arabic, Urdu, Persian |

The source card also reports error rates for six ways of reading the lines, from commercial vision-language models to small recognisers trained per dataset.

Four things are not documented at the source:

- **How the split was made.** The card gives only the three sizes. Section 2.5.3 shows that every dataset is cut 80/10/10 by line.
- **How the line images were cut.** Nothing is said about outlines, padding or resizing.
- **The license of each dataset.** The card sends the reader to the source repositories and gives no link for RASM. It marks RASM as "not redistributed", although the 1,488 RASM lines are in the files.
- **Where the Urdu and Persian are.** No column records the language, and none of the source descriptions we read mentions Urdu or Persian material.

**The eight datasets inside AMIDDA**

The table gives the identity of each dataset. A short part on each one follows, with its original state and how it appears in AMIDDA.

| Dataset | Built by | Material | Documents | Pages | License at its source |
|---|---|---|---|---|---|
| RASAM 1 | GIS MOMM, BULAC, Calfa | Maghrebi manuscripts: history and inheritance law | 3 | 299 | Apache 2.0 |
| RASAM 2 | GIS MOMM, DISTAM, BULAC, Calfa | Maghrebi manuscripts of many genres | 15 | 162 | Apache 2.0 |
| TARIMA | TariMa project: BULAC, MMSH, IREMAM, Calfa | Maghrebi history books: manuscripts and lithographs (books printed from handwriting on stone) | 26 | 103 | Apache 2.0 |
| BAYBARS | LiPoL project with the BnF, Ifpo, Université Saint-Joseph and Calfa | Notebooks of the Baybars story, in Middle Arabic | 3 | 546 | Etalab Open Licence 2.0 |

| Dataset | Built by | Material | Documents | Pages | License at its source |
|---|---|---|---|---|---|
| ISKANDAR | DISTAM, LiPoL project, Calfa | Manuscripts of the Romance of Alexander, in Middle Arabic | 13 volumes of 5 manuscripts | 251 | Etalab Open Licence 2.0 |
| RASM | British Library, PRImA lab | Scientific manuscripts | 4 | Not stated | "Open license" on the competition page |
| KHATT | KFUPM; the copy used here was uploaded by a Hugging Face user | Handwritten paragraphs by 1,000 writers | No document names | Not recorded | Research use only at KFUPM; the copy is tagged MIT |
| Muharaf | Saeed and others, NeurIPS 2024 | Letters, diaries, poems, church and legal records | No document names | 1,216 at the source | CC BY-NC-SA 4.0 |

Documents and pages of the five Calfa datasets are the figures on their own Hugging Face cards. BULAC and the BnF are libraries in Paris, and MMSH is a research centre in Aix-en-Provence. GIS MOMM and DISTAM are French research consortia.

**RASAM 1**

RASAM 1 is the oldest of the five Calfa datasets, and the four others follow its transcription rules. It was made in a transcription workshop (a "hackathon") at the BULAC library in Paris from January to April 2021, and presented at ICDAR 2021. It covers three Maghrebi manuscripts: two history books (MS.ARA.1977 and MS.ARA.417) and one book on inheritance law (MS.ARA.609).

- **Original state:** the GitHub repository holds 300 page files in PAGE XML, with text regions, line outlines, baselines and text. We counted 7,540 line annotations, of which 7,343 carry text (492,602 characters). The page images are not in the repository; they are fetched from the library's image server.
- **In AMIDDA:** 7,331 lines from 299 pages, the same as Calfa's line-level release on Hugging Face. That is 12 lines and one page fewer than the XML files, all from MS.ARA.417 (1,166 lines instead of 1,178).
- **To know:** MS.ARA.1977 alone gives 3,394 lines. Lines are long: the median is 70 characters.

**RASAM 2**

RASAM 2 widens RASAM 1 with many different hands. A second workshop at BULAC, from December 2021 to March 2022, transcribed 15 manuscripts chosen to be unlike each other: chronicles, law, literature, poetry, religious devotion and grammar. Most were copied in the 17th to 19th centuries, and some hold Berber written in Arabic script. It was presented at CHR 2024.

- **Original state:** the same GitHub repository holds 247 page files. We counted 6,550 line annotations, of which 3,750 carry text: exactly 250 lines per manuscript. The other lines are located but not transcribed.
- **In AMIDDA:** 3,720 lines from 162 pages, again the same as the line-level release. Thirty lines and two pages are missing against the XML files: MS.ARA.1946 has 230 lines and MS.ARA.1926 has 240.
- **To know:** the manuscripts have almost the same weight (230 to 250 lines each), so no single hand dominates. In the measured lines it is the only dataset with lines from tables (107 lines).

**TARIMA**

TARIMA comes from TariMa (Tārīkh al-Maghrib, "writing history in the Maghreb"), a project funded by CollEx-Persée and led by Antoine Perrier (CNRS)

with BULAC, MMSH, IREMAM and Calfa. It adds words and document types that RASAM covers little: chronicles, travel accounts and biographical dictionaries. It also adds lithographed books next to manuscripts.

- **Original state:** the GitHub repository holds 120 page files from 29 documents: 85 manuscript pages and 35 lithograph pages. We counted 2,673 line annotations, of which 2,365 carry text (the authors state 2,364). Three lithographs have layout only.
- **In AMIDDA:** 2,285 lines from 103 pages of 26 documents, the same as the line-level release: 15 manuscripts (5 from BULAC, 5 from the BnF, 5 from MMSH) and 11 lithographs. Eighty lines and three pages are missing against the XML files, all from the manuscript MMSH MS.11 (376 lines instead of 456).
- **To know:** the 597 lithograph lines are the only printed material we found in AMIDDA (0.9% of all lines). The authors call TariMa a working dataset to be used together with RASAM, and only main text is transcribed.

**BAYBARS**

BAYBARS is the largest of the Calfa datasets: 15,567 lines, 23.3% of AMIDDA. The text is the Sīrat Baybars, a long popular story about Sultan Baybars, in its Damascus version. It is written in Levantine Middle Arabic, a written form that mixes classical and spoken Arabic. It was made in the LiPoL project ("Popular Literatures of the Levant", funded by the French research agency ANR) with the BnF, Ifpo, Université Saint-Joseph in Beirut and Calfa. Sixteen participants transcribed it between September 2023 and July 2024.

- **Original state:** a separate Hugging Face dataset with 15,567 lines from 546 pages of three notebooks (502,534 characters). The card points to the project's GitLab server for the page-level files; we could not open that server from our workspace.
- **In AMIDDA:** the same number of lines, but another split. The separate release has 12,483 / 1,556 / 1,528 lines and other counts per notebook; AMIDDA has 12,453 / 1,556 / 1,558. So one split cannot be used in place of the other.
- **To know:** notebook 3 alone gives 7,996 lines, 12.0% of AMIDDA. Lines are short: the median is 32 characters. In the separate release, about 8% of the training lines are page numbers, catchwords (the first word of the next page, written at the foot of a page) and margin notes.

### ISKANDAR

ISKANDAR holds the Sīrat al-Iskandar, the Arabic Romance of Alexander, also in Middle Arabic. It covers five manuscripts in 13 volumes, from the 18th and 19th centuries. Three are at the BnF in Paris, one in nine volumes is at the Gotha Research Library, and one is at the University of Leeds. Twelve students from French universities transcribed it between November 2022 and May 2023.

- **Original state:** a separate Hugging Face dataset with 5,221 lines from 251 pages (230,450 characters). Its card points to the same GitLab server.
- **In AMIDDA:** 5,221 lines with the same split sizes (4,176 / 522 / 523) and, for the 12 volumes we could measure, the same training counts.
- **To know:** the BnF manuscripts give 2,523 lines, Gotha 1,923 and Leeds 775. One Gotha volume has only 29 lines and none in the test split. The line images are the widest of the Calfa datasets (median 1,558 px).

### RASM

Apart from KHATT and Muharaf, RASM is the only dataset that Calfa did not build. It comes from the British Library's competitions on the Recognition of Historical Arabic Scientific Manuscripts. The 2019 edition was run with the PRImA lab at ICDAR 2019. It released 120 page images, 20 as examples and 100 for evaluation, from the British Library and the Qatar Digital Library. Regions, lines and text are given in PAGE XML.

- **Original state:** page images in TIFF, each with one PAGE XML file. The AMIDDA card names the source only as "British Library"; Calfa's website points to the 2019 competition data.
- **In AMIDDA:** 1,488 line images from four manuscripts: Add MS 7474, Add MS 23494, Or 3366 and Delhi Arabic 1901. It has its own guideline label, `rasm`. Who cut the lines, and from which pages, is not documented.
- **To know:** it is the smallest dataset (2.2%). We read ten RASM labels, all from Add MS 23494: nine carry the dagger alif (a small alif written above the letter) and one has all its vowel marks. The AMIDDA card marks RASM "not redistributed" and gives no link, so its license inside AMIDDA needs checking.

### KHATT

KHATT (KFUPM Handwritten Arabic TexT) was collected at King Fahd University of Petroleum and Minerals from 1,000 writers. It has 2,000 paragraph

images with the same text for every writer, and 2,000 with unique text. AMIDDA did not take it from KFUPM but from a copy on Hugging Face.

- **Original state:** the Hugging Face copy has 6,673 line images with two columns, `image` and `text`, split 4,672 / 963 / 1,038 (70/15/15 according to its card).
- **In AMIDDA:** all 6,673 lines, split again 80/10/10 (5,338 / 667 / 668). The file name keeps the old split, for example `khatt_v1_test_000009`.
- **To know:** all 48 KHATT labels we read are pieces of one and the same paragraph, the fixed text that every writer copied. So this copy seems to hold the same few sentences in many hands, not many different texts. KFUPM's terms allow research use only and no redistribution, while the copy is tagged MIT. Our own KHATT subset (11,398 lines, 838 writers) was built from another copy of KHATT and has its own split.

**Muharaf**

Muharaf is the largest dataset inside AMIDDA: 24,494 lines, 36.7% of all lines. It was published at NeurIPS 2024 and holds letters, diaries, notes, poems, church records and legal papers from the early 19th to the early 21st century. So it is everyday handwriting of the last two centuries rather than books. The pages come from the Phoenix Center for Lebanese Studies (USEK) and the Khayrallah Center (NC State University). The full dataset has 1,644 page images and 36,311 lines; the public part has 1,216 pages.

- **Original state:** the public part on Hugging Face has 24,495 line images with two columns, `image` and `text`, in the official split 22,092 / 1,069 / 1,334. One training line has an empty label.
- **In AMIDDA:** 24,494 lines, split again 80/10/10 (19,595 / 2,449 / 2,450). One official training line is missing, very likely the one with the empty label. The file name keeps the official split, for example `train_5` or `test_1320`.
- **To know:** it is the same data as our Muharaf subset, in a different split. All 41 images we sampled are 60 pixels high. It has no script, support, region or manuscript label.

**In the collection.** AMIDDA was added to the collection on 12 September 2026 as a ready-made dataset. It is published as the 20 upstream Parquet files, without any change.

The useful part for the project is the 35,612 lines that do not come from Muharaf or KHATT. They are Maghrebi manuscripts, lithographed books, two long popular stories in Middle Arabic, and scientific manuscripts.

The AMIDDA authors fine-tuned Qwen3.5-4B with LoRA on the training split. They report 23.1% CER on the test split, from 9.8% on KHATT to 37.4% on RASM. Their small recognisers, trained one per dataset, reach 6.4% overall, with no result given for ISKANDAR.

### 2.5.2 Features (files, attributes, null and invalid values)

The published copy has nine attributes per line. Five are filled for every line we could measure; the other four are empty for 37% to 47% of the lines.

**Files**

| Path | Files | Size | Content |
|---|---|---|---|
| AMIDDA/data/train-*.parquet | 16 | 7.55 GB | 53,420 lines; single files from 173 MB to 1.9 GB |
| AMIDDA/data/validation-*.parquet | 2 | 0.96 GB | 6,675 lines |
| AMIDDA/data/test-*.parquet | 2 | 0.94 GB | 6,684 lines |
| AMIDDA/README.md | 1 | 7 KB | A copy of the source card (see 2.5.4) |

This layout differs from our five converted subsets, which ship one images.zip and three CSV files. Here the image is stored inside each Parquet row, at about 140 KB per line on average.

The rows are grouped by dataset, not shuffled: the five Calfa datasets come first, then RASM, KHATT and Muharaf. In the training split, Muharaf fills the last 19,595 rows, and the first 26,712 rows hold only Calfa datasets. A loader that stops early sees no KHATT or Muharaf line, so the rows must be shuffled before training.

**Attributes**

| Attribute | Type | Meaning | Example | Filled for |
|---|---|---|---|---|
| image | Image stored in the row | The cropped text line | 652 × 78 px (first training line) | All lines |
| transcription | Text | What the line says, as written | (a رحمهما الله Muharaf line) | All lines |
| dataset | Category, 8 values | The source dataset | rasam1, baybars, muharaf | All lines |
| transcription_guidelines | Category, 4 values | The rule set the transcribers followed | rasam, rasm, khatt, muharaf | All lines |
| script | Category, 3 values | Script family | maghrebi, oriental, contemporary | All but Muharaf |
| support_type | Category, 2 values | Handwritten or lithographed | manuscript, lithography | All but Muharaf |
| region_type | Category, at least 25 spellings | Part of the page the line belongs to | MainText, Marginalia_Catchword | All but KHATT and Muharaf |
| manuscript_name | Category, 64 values | The document the line comes from | BULAC_MS_ARA_1977, baybars-nt-ms3 | All but KHATT and Muharaf |
| image_name | Text | Page name, or file name in the source | 0012, c041-434, train_5 | All lines |

- **transcription is the label.** Our converted subsets call this column text and give the image as a file path. Code written for them does not read AMIDDA without a small adapter.
- **image_name means two different things.** For the Calfa datasets and RASM it names the page, so all lines of a page share it. RASAM 1 has 249 page names for 299 pages, so manuscripts also reuse names. For KHATT and Muharaf it is the file name of the line in the source, with the source split in it.
- **Most lines have no id.** For the Calfa datasets and RASM there is no line id and no position of the line on its page, so a line can only be pointed at by its row number. No dataset has a writer or a language attribute.

**Null and invalid values**

| Attribute | Train | Validation | Test | All lines | Empty for |
|---|---|---|---|---|---|
| manuscript_name, region_type | 24,933 | 3,116 | 3,118 | 31,167 (46.7%) | KHATT and Muharaf |
| script, support_type | 19,595 | 2,449 | 2,450 | 24,494 (36.7%) | Muharaf |
| The five other attributes | 0 in the first 26,712 lines | 0 | 0 | 0 found | No line |

Validation and test are measured on all their lines. For training, the dataset viewer measures only the first 26,712 lines, so the two training counts are worked out from the sizes of KHATT and Muharaf. Rows sampled at the end of the training split show the same pattern.

The empty values are not errors. They show that KHATT and Muharaf came in without the page-level labels of the other datasets. Five other points need care:

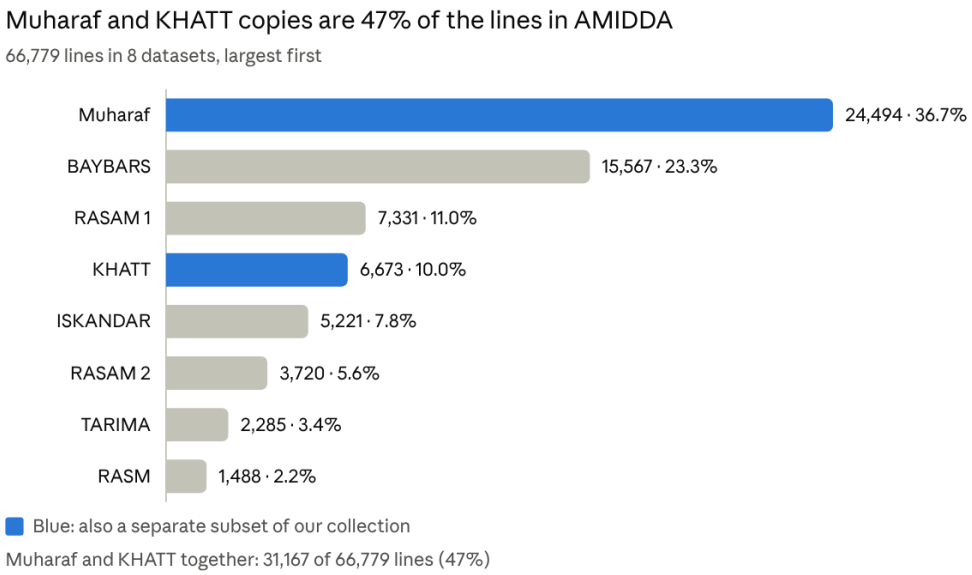
- **region_type has at least 25 spellings for about six ideas.** Each source kept its own words. Main text is MainText in four datasets and MainZone in two. A catchword is catchword, Marginalia_Catchword or MarginTextZone:catchword. A page number is page_number, Marginalia_PageNumber or Numerotation.

- **script and support_type are loose for KHATT and Muharaf.** KHATT is labelled contemporary and manuscript. Muharaf has no label at all.
- **One manuscript name does not follow the pattern.** It is written BULAC_ARA_MS_1982, while the 22 other BULAC manuscripts are written BULAC_MS_ARA_…. In the page files on GitHub the name follows the pattern, so the odd spelling came in with the line-level release.
- **Some labels contain #.** The RASAM 1 card says it marks a character that cannot be read. We also saw it in 5 of the 48 KHATT labels we read.
- **Some KHATT crops are much taller than a line.** Five of the 63 sampled KHATT images are 665 to 1,180 pixels high, while the others are 85 to 232. These crops probably cover more than one written line; this needs a look at the images.

### 2.5.3 Analysis (lines per dataset, split sizes, the split is by line and not by document, documents, scripts and material, overlap with our Muharaf subset, line length and image width by dataset, four labelling conventions, images)

AMIDDA is large but uneven. Copies of Muharaf and KHATT make up 47% of its lines, and the split cuts every dataset by line, so validation and test are not independent of training.

## Lines per dataset




Source card totals and dataset statistics of calfa-ai/AMIDDA · 66,779 lines in 8 datasets

The five Calfa datasets hold 34,124 lines (51.1%) and RASM 1,488 (2.2%). Without Muharaf and KHATT, AMIDDA has 35,612 lines, of which 28,487 are in training.

## Split sizes

| Dataset | Train | Validation | Test |
|---|---|---|---|
| RASAM 1 | 5,864 | 733 | 734 |
| RASAM 2 | 2,976 | 372 | 372 |
| TARIMA | 1,828 | 228 | 229 |
| BAYBARS | 12,453 | 1,556 | 1,558 |
| ISKANDAR | 4,176 | 522 | 523 |
| RASM | 1,190 | 148 | 150 |
| KHATT | 5,338 | 667 | 668 |

| Dataset | Train | Validation | Test |
|---|---|---|---|
| Muharaf | 19,595 | 2,449 | 2,450 |
| All | 53,420 | 6,675 | 6,684 |

Every dataset is cut the same way: training gets 80% of its lines rounded down, validation 10% rounded down, and test the rest. Validation and test counts are measured on all lines. The training counts of the first four datasets are measured too. Those of ISKANDAR, RASM, KHATT and Muharaf were checked against their row positions in the split.

**The split is by line, not by document**

All 64 named documents have lines in the validation split, and 63 of them in the test split. Pages are cut too. For example, page 0012 of MS.ARA.1977 has 32 lines: 24 are in training, 2 in validation and 6 in test.

So a model trained on this split is tested on hands and pages that it has also seen in training. Scores on AMIDDA's own test split are likely to look better than scores on unseen books, such as the three held-out books of AraMs-28k-HTR. Our project rule groups lines by document whenever we make a split ourselves, exactly to avoid this. Here the split is the source's own, so we kept it, and this weakness came with it.

**Documents**

| Dataset | Documents | Largest document | Lines in it |
|---|---|---|---|
| RASAM 1 | 3 | MS.ARA.1977 | 3,394 |
| RASAM 2 | 15 | 13 manuscripts with 250 lines each; the two others have 230 and 240 | 250 |
| TARIMA | 26 | MMSH MS.11 | 376 |
| BAYBARS | 3 | Notebook 3 | 7,996 |

| Dataset | Documents | Largest document | Lines in it |
|---|---|---|---|
| ISKANDAR | 13 | BnF volume btv1b10031486f | 1,378 |
| RASM | 4 | Not counted per manuscript | |

The 64 documents hold 35,612 lines, and the weight is very unequal. The three BAYBARS notebooks hold 15,567 lines (43.7%), and five documents hold 61.0%. At the other end, 22 of the 60 Calfa documents have fewer than 100 lines each: 19 in TARIMA and 3 in ISKANDAR.

**Scripts and material**

By dataset size, the script attribute gives 22,276 oriental lines (33.4%), 13,336 Maghrebi lines (20.0%) and 6,673 contemporary lines (10.0%). The 24,494 Muharaf lines (36.7%) have no script label. Lithographed lines are 597 (0.9%), all in TARIMA.

About 94% of the measured lines with a region label are main text. The rest are page numbers, catchwords, margin notes, titles and table lines.

**Overlap with our Muharaf subset**

AMIDDA's file names keep the official Muharaf split. The Muharaf rows also stand in source order: first the train_ names, then validation_, then test_. We located the two boundaries in each split, which gives the overlap line by line.

| Official Muharaf split | In AMIDDA train | In AMIDDA validation | In AMIDDA test | Official size |
|---|---|---|---|---|
| Train | 17,680 | 2,210 | 2,201 | 22,092 |
| Validation | 852 | 107 | 110 | 1,069 |
| Test | 1,063 | 132 | 139 | 1,334 |

Of the 1,334 official test lines, 1,063 (79.7%) sit in AMIDDA's training split and only 139 in its test split. A model trained on AMIDDA has therefore seen most of the official Muharaf test set. The reverse also holds: 2,201 official training lines are in AMIDDA's test split. One official training line is not in AMIDDA, very likely the line with the empty label.

For KHATT the collection card gives the same warning. A line-by-line match needs the image files, because our KHATT subset was built from another copy with other file names, and matching by text does not work, because all the KHATT labels we read carry the same paragraph. The counts show the overlap instead (section 2.6.3): AMIDDA's KHATT copy very likely holds all 5,635 fixed-paragraph lines of our subset.

**Line length and image width by dataset**

| Dataset | Lines measured | Median length (characters) | Mean length | Longest | Median image width (px) | Width range (px) |
|---|---|---|---|---|---|---|
| RASAM 1 | 5,864 | 70 | 67.1 | 130 | 853 | 22 to 3,029 |
| RASAM 2 | 2,976 | 57 | 60.2 | 136 | 1,257 | 40 to 2,276 |
| TARIMA | 1,828 | 62.5 | 60.0 | 118 | 618 | 48 to 2,738 |
| BAYBARS | 12,483 | 32 | 32.3 | 106 | 915 | 20 to 1,962 |
| ISKANDAR | 4,176 | 44 | 44.0 | 105 | 1,558 | 30 to 3,287 |
| Muharaf | 22,092 | 47 | 49.3 | 181 | 600 | 60 to 1,620 |
| All of AMIDDA, validation | 6,675 | 47 | 49.3 | 176 | 840 | 34 to 3,887 |
| All of AMIDDA, test | 6,684 | 46 | 49.3 | 180 | 843 | 24 to 3,952 |

The first six rows are measured on the training split of each dataset's own release (for Muharaf, the official training split). RASM has no line-level release of its

own, and the viewer of the KHATT copy does not work, so both have no statistics. In our samples, RASM lines are 128 to 3,782 px wide (28 lines) and KHATT lines 697 to 2,435 px wide (63 lines).

Among the six datasets measured, lines are shortest in BAYBARS (median 32 characters) and longest in RASAM 1 (median 70). Image width does not follow text length. The median Muharaf line is 600 px wide, and every Muharaf image we sampled is 60 px high. The median ISKANDAR line has about the same number of characters but is 1,558 px wide. An input size chosen for one dataset will shrink or stretch the others.

**Four labelling conventions**

By dataset size, the `transcription_guidelines` attribute puts the lines into four groups: `rasam` 34,124 lines (51.1%), `muharaf` 24,494 (36.7%), `khatt` 6,673 (10.0%) and `rasm` 1,488 (2.2%). The source card says that only the `rasam` rules were formally written down and checked. In the labels we read, the groups differ in what they write:

- **`rasam`** (the five Calfa datasets): vowel marks are left out unless they are needed. In the RASAM 1 page files, 5.7% of the lines carry one, almost all in MS.ARA.609. The scribe's spelling is kept, for example `شىء` and `عادم`.
- **`rasm`:** nine of the ten labels we read carry the dagger alif, and one is fully vocalised. All ten come from one manuscript.
- **`khatt`:** modern spelling with hamza, full stops and commas.
- **`muharaf`:** full stops, Eastern Arabic digits and tanwin (the double vowel mark at the end of a word) appear, and dialect spellings are kept.

Our rule keeps text as the source gives it, so these differences stay in the labels. A model trained on all eight datasets together gets mixed signals about vowel marks, hamza and punctuation. Our rule is to normalise the text only when scoring.

**Images**

The images sit inside the Parquet rows: 9.45 GB for 66,779 lines, about 140 KB per line. The widths by dataset are in the table above. The heights are not in the dataset statistics, so they are known from samples only: all 41 Muharaf lines we read are 60 pixels high, the 63 KHATT lines we read are 85 to 232 pixels high apart from five tall crops of 665 to 1,180 pixels, and the 28 RASM lines we read are 81 to 926 pixels high. The images of the eight datasets therefore differ a lot in size.

### 2.5.4 Results (before and after, what was done upstream, decisions, limitations that remain)

After preparation, the data of AMIDDA is exactly what the source released. All 20 Parquet files have the same size and the same SHA-256 checksum as the source files. The `data` folder also has the same Git tree hash in the two repositories. We added, removed, renamed and relabelled no line. Only the card in the folder differs slightly from the source card.

**Before and after**

| Property | Source (calfa-ai/AMIDDA) | Published copy |
|---|---|---|
| Place | Its own repository | Folder `AMIDDA/` of our collection |
| Data files | 20 Parquet files, 9.45 GB | The same 20 files, identical |
| Lines | 66,779 | 66,779 |
| Attributes | 9 | 9, with the same names |
| Split | 53,420 / 6,675 / 6,684, by line | Unchanged |
| Labels | Four conventions, see 2.5.3 | Unchanged |
| Card | `README.md`, 7,225 bytes today | A copy with the same sections, 146 bytes larger (7,371 bytes) |
| Warning about Muharaf and KHATT | None | Added to the collection card |
| License note | "other": mixed | "Per source" on the collection card |

**What was done upstream**

The preparation of this subset was done upstream, by Calfa, before we copied it. Line counts and split sizes show part of what was done. For RASAM 1, RASAM 2

and TARIMA, the lines were dropped when Calfa made its line-level releases on Hugging Face. AMIDDA takes those releases as they are.

| Dataset | Compared with | What differs in AMIDDA |
|---|---|---|
| RASAM 1 | Page files on GitHub: 7,343 lines with text | 12 lines and one page fewer, all from MS.ARA.417. The 197 lines without text are left out |
| RASAM 2 | Page files on GitHub: 3,750 lines with text | 30 lines and two pages fewer, from two manuscripts. The 2,800 lines without text are left out |
| TARIMA | Page files on GitHub: 2,365 lines with text | 80 lines and three pages fewer, all from MMSH MS.11. The 308 lines without text are left out |
| BAYBARS | Separate release: 15,567 lines | Same number of lines, split again: the counts per split and per notebook differ |
| ISKANDAR | Separate release: 5,221 lines | Nothing found: same split sizes and same training counts per volume |
| RASM | Competition data: page images with page files | 1,488 line images; who cut them is not documented |

| Dataset | Compared with | What differs in AMIDDA |
|---|---|---|
| KHATT | Hugging Face copy: 6,673 lines | Split again 80/10/10 instead of the copy's own split; `text` renamed `transcription` |
| Muharaf | Public release: 24,495 lines | One training line fewer; split again 80/10/10 instead of the official split; `text` renamed `transcription` |

AMIDDA also brings everything into one schema of nine attributes. The five Calfa releases already had seven of them, so only `dataset` and `transcription_guidelines` are new there. KHATT and Muharaf had only an image and a text.

**Decisions**

| Decision | Reason | Alternative |
|---|---|---|
| Publish AMIDDA as the source's Parquet files, without converting it to `images.zip` and CSV files | The amount of data is very large (9.45 GB), so we could not convert it, and what we do not want can be filtered out during training (the dataset column names the source of each line) | Convert it like the five other subsets and keep `dataset` as a third column, so that one loader reads everything |
| Keep the nine source attributes and their names | To be able to filter the data when needed: the attributes (dataset, script, region_type and others) describe each line | Rename `transcription` to `text` and reduce `region_type` to a few values |

| Decision | Reason | Alternative |
|---|---|---|
| Keep the source split | Project rule: official splits are used where they exist. The second half of the rule, grouping by document, applies only to splits we make ourselves | Split again by document, or use AMIDDA for training only and report no score on its test split |
| Keep the Muharaf and KHATT copies inside, with a warning | The card warns that training on AMIDDA and testing on our Muharaf or KHATT subsets, or the reverse, leaks lines. It tells users to pick one, or to filter on `dataset` | Publish only the six other datasets (35,612 lines), so that no combination can leak |
| Record the license as "Per source" | Each dataset keeps its own license, as on the source card | List the eight licenses on the card and settle the two unclear ones, RASM and KHATT |

**Limitations that remain**

- **Overlap with our Muharaf subset, and probably with KHATT.** Copies of Muharaf and KHATT are 47% of AMIDDA, and 1,063 of the 1,334 official Muharaf test lines are in AMIDDA's training split. AMIDDA must not be mixed with our Muharaf or KHATT subsets without filtering on `dataset`.
- **Split by line.** Validation and test share documents and pages with training, so their scores are likely to overstate how well a model reads unseen manuscripts.
- **KHATT text repeats.** All 48 KHATT labels we read carry one fixed paragraph. If the rest are the same, a model can memorise this paragraph, and a low error on these lines does not prove that it can read the handwriting.
- **Unequal weight.** Three BAYBARS notebooks hold 44% of the lines that have a document name, and 22 documents have fewer than 100 lines.

- **Four labelling conventions.** Vowel marks, hamza and punctuation differ by source, and `region_type` has at least 25 spellings.
- **Margin and page-number lines are mixed in.** About 6% of the measured lines with a region label are not main text. They can be filtered with `region_type`.
- **Image sizes differ widely.** Measured line widths run from 20 to 3,952 px, and every sampled Muharaf image is only 60 px high.
- **Mixed licenses, two unclear.** Apache 2.0, Etalab 2.0 and CC BY-NC-SA 4.0 are stated at the sources. RASM is marked "not redistributed" on the AMIDDA card, and KHATT is for research use only at KFUPM.
- **Different layout and row order.** The loader of the five converted subsets cannot read AMIDDA. Its rows are grouped by dataset. The dataset viewer measures only the first 26,712 training lines.
- **Gaps in the source card.** The split method, the way lines were cut and the Urdu and Persian named on the card are not documented.

## 2.6 KHATT

### 2.6.1 Description (original state, place in the collection)

KHATT is modern handwriting, not manuscripts: 11,398 text lines written by 838 people. Our preparation had three steps: we rebuilt its labels from Latin codes, made our own test set from writers kept out of training, and corrected the file type in its file names. About half of its lines (49.4%) are pieces of one fixed paragraph that the writers copied.

**Original state**

KHATT (KFUPM Handwritten Arabic TexT) comes from King Fahd University of Petroleum and Minerals (KFUPM) in Dhahran. A team led by Professor Sabri Mahmoud built it, with help from TU Dortmund and TU Braunschweig in Germany. Our decoder notebook reads a copy on Kaggle, not KFUPM's own download, so the table describes both.

| Property | KHATT v1.0 at KFUPM | Kaggle copy we started from |
|---|---|---|
| Publisher | KFUPM | Nizar Charrada, nizarcharrada/khattarabic ("KHATT-Arabic") |
| Version and date | Version 1.0, copyright 2014 | Two versions, both on 6 September 2022; version 2 "added_csv" |
| Writers | 1,000 | 838 in the training and validation files (our card) |
| Content | 2,000 paragraph images of a similar text that covers all Arabic letter shapes, 2,000 of unique text, and free paragraphs; line images for the first two kinds | Line images and CSV files; the part of the file list we could read shows no paragraph images |
| Images | Scanned at 200, 300 and 600 dpi | JPG line images in a folder named Train_deskewed |

| | | |
|---|---|---|
| Labels | Ground truth checked by hand, with a Latin version | Latin codes, one per CSV column (our card and notebook) |
| Splits | Separate training (70%), validation (15%) and test (15%) sets | Train.csv with 9,497 lines and Validation.csv with 1,901 lines (notebook), and a test split (our card) |
| License | Free for academic, research and similar non-commercial use after an agreement; no passing on to third parties | "Unknown" |
| Size | Not stated | 344 MB |

Sources: the KHATT website, the KFUPM agreement, the Kaggle page, our card and the decoder notebook.

The file names show who wrote each line and where it comes from. A name like AHTD3A0008_Para2_1.jpg gives the writer (0008), the paragraph (2) and the line in it (1). In the lines we read, paragraphs 1 and 4 are always the same fixed paragraph, and paragraphs 2 and 3 are different texts for each writer. KFUPM's 2,000 similar-text paragraphs for 1,000 writers fit two copies of the fixed paragraph per writer.

Three things are missing in the copy we started from:

- **Arabic labels.** Each line is a row of Latin codes such as ba, ta or sp, so the Arabic text has to be rebuilt.
- **Writer IDs in the test split.** Our card says the test split "exposes no writer ID", so it cannot be checked for writers who are also in training.
- **The right file type.** The CSV files name the images .tif, but the files are .jpg.

The paper to cite is Mahmoud and others, "KHATT: An Open Arabic Offline Handwritten Text Database", Pattern Recognition, 2014 (doi 10.1016/j.patcog.2013.08.009).

**In the collection.** KHATT is an optional subset. Our card keeps it "for pre-training or comparison" and advises: "Leave it out if you are measuring manuscript HTR." It is published as 11,398 line images with decoded Arabic labels: 8,094 lines for training, 1,901 for validation and 1,403 for test. The validation set is the official one, and the test set is 104 writers that we held out of the upstream training file.

### 2.6.2 Features (files, attributes, null and invalid values)

The published copy has two attributes per line, image_path and text, the same layout as our other converted subsets. No value is empty, but the labels follow their own conventions.

**Files**

| Path | Size | Content |
|---|---|---|
| KHATT/images.zip | 308 MB | The JPG line images; about 27 KB per line (zip size divided by the number of lines) |
| KHATT/train.csv | 1.09 MB | 8,094 lines |
| KHATT/val.csv | 0.26 MB | 1,901 lines |
| KHATT/test.csv | 0.19 MB | 1,403 lines |
| KHATT/khatt-label-decoder.ipynb | 16 KB | The notebook that turns the Latin codes into Arabic |

Among the five converted subsets, KHATT is the only one with JPG images, and it ships the notebook that decodes its labels. The choice of test writers is not in the notebook.

**Attributes**

| Attribute | Type | Meaning | Example |
|---|---|---|---|
| image_path | Text | File name of the line image: writer, paragraph and line | AHTD3A0058_Para2_2.jpg |
| text | Text | The decoded Arabic transcription | كلها (بلور مسال) , السبورات آلية يمكن تحويلها إلى شاشات للعرض . |

- **image_path carries information.** The four digits after AHTD3A are the writer, ParaN is the paragraph, and the last number is the line in that paragraph. There is no writer or paragraph column, so the file name is the only place where the writer and the paragraph are recorded.
- **text is rebuilt, not copied.** Each label is the decoding of a row of Latin codes. Section 2.6.3 describes the decoding.

**Null and invalid values**

| Check | Train | Validation | Test |
|---|---|---|---|
| Empty image_path or text | 0 | 0 | 0 |
| Shortest label (characters) | 1 | 4 | 4 |

No value is empty. Four points need care:

- **Punctuation and digits are Latin forms.** The code table turns commas into ,, question marks into ? and numbers into Western digits such as 1992. So the labels cannot contain an Arabic comma (،) or question mark (؟), even though the images show writers using the Arabic comma (see Images).
- **Spaces are kept as in the source.** Labels have double and triple spaces, and spaces before full stops, for example «نائم  بكلمات» and «للحج .».
- # **most likely stands for a crossed-out word.** It comes from the code scr, and appears in labels such as «ذهب نوح مظفر ضرغام # بصحبة».
- **The writers' own mistakes stay in.** The same fixed sentence appears as «يتكلم» and «يتكسم», or «مظفر» and «مظفر».

### 2.6.3 Analysis (split sizes and writers, fixed and unique text, overlap with AMIDDA's KHATT, line length, decoding the labels, images: format and size, what the images show, images against labels)

KHATT is split by writer, but about half of its text is one paragraph that appears in every split.

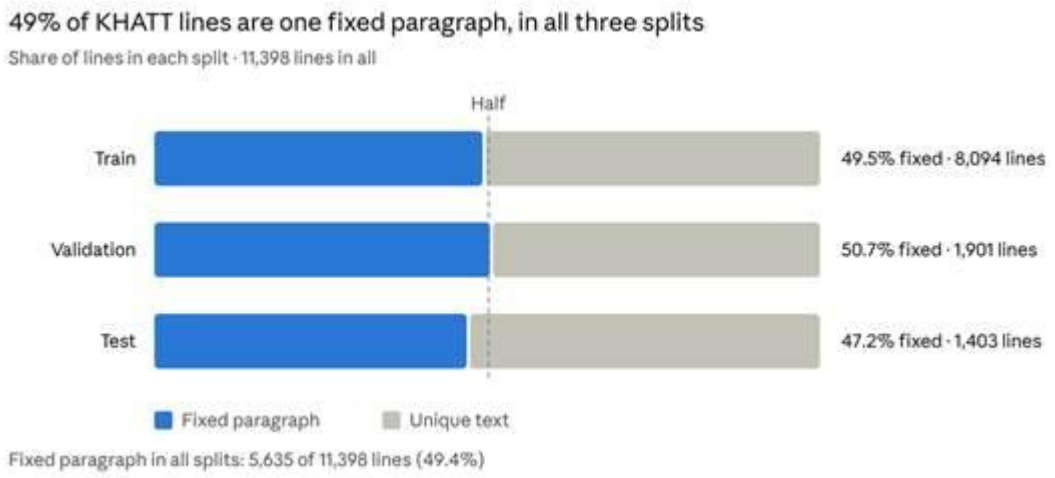
**Split sizes and writers**

| Split | Lines | Writers | Lines per writer | Origin |
|---|---|---|---|---|
| Train | 8,094 | 587 | 13.8 | Upstream training split without 104 writers |
| Validation | 1,901 | 147 | 12.9 | Upstream validation split, unchanged |

| | | | | |
|---|---|---|---|---|
| Test | 1,403 | 104 | 13.5 | 104 writers held out from upstream training (seed 3) |
| All | 11,398 | 838 | 13.6 | |

The writer counts 691, 104 and 838 come from our card; the 587 and 147 follow from them. The test writers are 15.1% of the 691 upstream training writers and carry 14.8% of its lines. The final split is 71% / 17% / 12%, and the upstream test split is not used. We did not check the splits for shared writers ourselves.

**Fixed and unique text**




Text type read from the paragraph number in each image_path (1 and 4 fixed, 2 and 3 unique), counted by row position in train.csv, val.csv and test.csv · 11,398 lines

We counted the lines of each paragraph from the row order of the three CSV files, which are grouped by paragraph. This gives 5,635 lines of the fixed paragraph (paragraphs 1 and 4) and 5,763 lines of unique text (paragraphs 2 and 3). In the training split, paragraphs 1 to 4 hold 1,920, 2,163, 1,921 and 2,090 lines.

Splitting by writer puts each person's handwriting in only one split, but the same text is still in every split. The training split holds 4,010 lines of the fixed paragraph from 587 writers, so its sentences appear there hundreds of times. A model trained on our training split has therefore already seen the text of 662 of the 1,403 test lines. A model that knows Arabic well can simply memorise this text. Only the 741 unique-text test lines measure reading of new text by new writers.

Not every line of a paragraph is present. For example, writer 0008 has lines 1, 5 and 6 of paragraph 1 and lines 3, 4 and 8 of paragraph 4. Writer 1169 has no line 3 in paragraph 2. On average a training writer has 6.8 fixed-paragraph lines. The copy does not say how the lines were chosen.

**Overlap with AMIDDA's KHATT**

The counts show the overlap. The upstream training file has 4,672 fixed-paragraph lines (4,010 in our train and 662 in our test), and the validation file has 963. AMIDDA took its KHATT from a Hugging Face copy with exactly 4,672 training and 963 validation lines, and every AMIDDA KHATT line we read is the fixed paragraph. So that copy is very likely the fixed-text part of the same release. If so, it holds all 5,635 of our fixed lines under other file names. It also holds 1,038 fixed lines from its own test split, but none of our unique-text lines. A model trained on AMIDDA would then already have seen most of our fixed test lines.

**Line length**

| Split | Shortest | Median | Mean | Longest |
|---|---|---|---|---|
| Train | 1 | 63 | 61.3 | 131 |
| Validation | 4 | 63 | 61.7 | 117 |
| Test | 4 | 64 | 61.7 | 105 |

Lengths are in characters. Lines are long and of similar length: two thirds of the training lines (66.1%) have 57 to 84 characters, and only 183 (2.3%) have fewer than 15. The three splits hold about 700,000 characters.

**Decoding the labels**

The upstream files hold 583,275 tokens in the training file and 117,232 in the validation file, about 61 per line. The notebook maps them with a table of 67 codes:

- 36 letters, including the hamza forms, ta marbuta and alef maqsura
- 10 digits
- 16 punctuation marks and symbols, plus the space
- 4 lam-alef forms

The first 63 codes come from the table in the KHATT readme; the 4 lam-alef codes were found in the data and added. One more token, :, is skipped. It appears 4,825

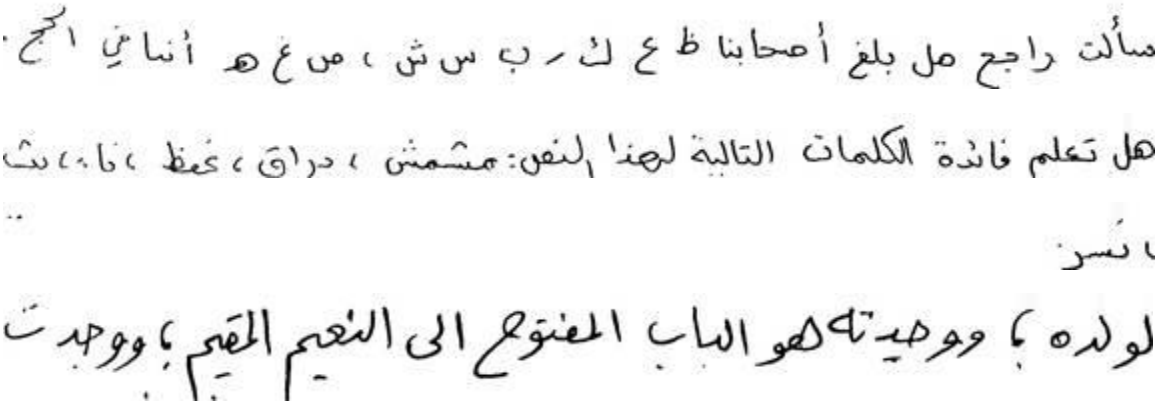
times in the upstream training file and 938 times in the validation file. These are exactly the numbers of unique-text lines in the two files, so it probably marks those lines and is not a letter. With it set aside, every code resolved, and no character was guessed. The table has no codes for vowel marks, so no label has any.

The decoded labels differ from AMIDDA's copy in the fixed paragraph. In the lines we read, AMIDDA's labels put an Arabic comma between the letter groups («ظ ع ك ، ث خ ض») and write the lone letter as «هـ». In every sampled line of ours, those commas are missing (two spaces instead) and the letter is «ه». So the Latin codes drop the Arabic comma. The images show that writer 0058 wrote these commas and the «هـ», so AMIDDA's labels are closer to the handwriting here (see Images).

**Images**

The images are JPG files of about 27 KB per line (the size of images.zip divided by the number of lines). The upstream folder name, Train_deskewed, suggests that the uploader straightened the lines before publishing them; neither the Kaggle page nor our card says so.

We looked at 12 training images that the team shared, and matched each one to its row in train.csv by its text. They are all six lines of one writer's fixed paragraph (AHTD3A0058, rows 4,084 to 4,089) and all six lines of another writer's unique-text paragraph (AHTD3A0008, rows 13 to 18). Two writers out of 587 is a small sample, so the points below are examples, not totals.



| Sample | Row and file | Label in train.csv | What the image shows |
|---|---|---|---|

| | | | |
|---|---|---|---|
| 1 | 4,087,<br>AHTD3A0058_Para1_4.jpg | سألت راجح هل بلغ أصحابنا ظ ع ك ب س ش ص غ ه أننا في الحج . | The writer skipped «ث خ ض», and so does the label. The marks after «ك» and «ش» are not in the label, and the writer's «هـ» is written «ه» |
| 2 | 4,088,<br>AHTD3A0058_Para1_5.jpg | هل تعلم فائدة الكلمات التالية لهذا النص: مشمش دراق غيظ ناء بث | Four commas between the last five words. The label has two or three spaces in their place |
| 3 | 4,089,<br>AHTD3A0058_Para1_6.jpg | نسر . | One word at the right end and two dots at the far left. Ink covers less than 8% of the width |
| 4 | 15,<br>AHTD3A0008_Para2_3.jpg | لولده , ووجدته هو الباب المفتوح الى النعيم المقيم , ووجدت | Two Arabic comma marks, each with a dot below like «؛». The label has a Latin «,» in both places |

**Format and size**

| Property | The 12 images |
|---|---|
| File type | JPG at quality about 75, with no EXIF data |
| Colour | Stored with three colour channels, but every pixel is grey (red = green = blue) |
| Tones | Almost black and white: 89% to 99.5% of pixels are white or nearly white, and 0.4% to 9.4% are black or nearly black. The few grey pixels, at most 1.8%, sit along stroke edges |
| Resolution | Not stored. The files carry no dpi value |

| | |
|---|---|
| Width | 447 to 2,217 pixels; 9 of the 12 are 1,964 to 2,217 |
| Height | 104 to 213 pixels, median 158 |
| File size | 5.3 to 40.2 KB, mean 28.2 KB, close to the 27 KB average of the whole images.zip |

**What the images show**

- **Lines are level.** Our slope estimate is under 1 degree for 8 of the 12 lines and at most 3 degrees for the rest, two of which hold only one or two words. This fits the folder name Train_deskewed.
- **Crops are tight.** Ink touches the bottom edge in all 12 images and the top edge in 3. Some marks are cut, for example the tail of the last «م» of «المقيم» (row 15) and the tanween above «ًطاهرا» (row 18). Row 4,084 also has a short double line in its bottom-left corner that does not look like part of the text.
- **Writing size changes, even within one writer.** Writer 0058 fitted 70 characters into 1,009 pixels on the first line (row 4,084), but used about 2,200 pixels for 60 to 74 characters on each of the next four. Strokes are about 5 pixels thick for writer 0058 and about 7 for writer 0008.
- **Short lines leave white space.** The last line of a paragraph can be one or two words: row 4,089 (sample 3) has ink on less than 8% of its width, and row 18 is only 447 pixels wide.
- **Faint streaks.** Most of the 12 images have short, light grey vertical streaks at the top or bottom edge. They are too light to count as ink, and we do not know where they come from.

**Images against labels**

- **Labels follow what the writer wrote.** Writer 0058 left out «رايق» (row 4,084), «حاج» (row 4,085) and «ث خ ض» (row 4,087), and the labels leave them out too. Writer 0083's labels have all three (rows 4,090, 4,091 and 4,093), so the labels are not copies of the printed paragraph.
- **Commas in the fixed paragraph are missing.** Writer 0058 put marks after «ك» and «ش» in row 4,087 and four commas in row 4,088, but the labels have only extra spaces there. None of the 100 fixed-paragraph labels we checked (rows 4,084 to 4,133 and 6,004 to 6,053) has a comma. AMIDDA's copy writes these commas as «،» and the lone letter as «هـ», as writer 0058 did.

- **Commas in the unique text are Latin.** Writer 0008 wrote seven Arabic comma marks in rows 13 to 17, some with a dot below like «؛». The labels have a Latin «,» in each of the seven places, and the spaces around it change from line to line.
- **Vowel marks are not labelled.** The tanween on «نظرًا» (row 13) and «ظاهرًا» (row 18) is in the images but not in the labels, as expected from a code table with no vowel codes.

So one handwritten mark gets three different labels: «؛» in AMIDDA, «,» in KHATT's unique text, and nothing in KHATT's fixed paragraph. Mapping «,» to «؛» would fix the unique text, but the fixed paragraph would still miss its commas.

### 2.6.4 Results (before and after, decisions, limitations that remain)

After preparation, KHATT has 11,398 lines from 838 writers, with Arabic labels, a test set of writers not used in training, and the right file type in its file names. Every line of the two upstream files we used is kept; the upstream test split is not used.

**Before and after**

| Property | Kaggle copy | Published copy |
|---|---|---|
| Splits used | Train 9,497 and validation 1,901; the test split is left out | Train 8,094, validation 1,901, test 1,403 |
| Test set | Upstream test, without writer IDs | 104 upstream training writers, held out with seed 3 |
| Labels | Rows of Latin codes, one per column | Arabic text in one `text` column |
| File names in the CSV | `.tif` | `.jpg`, as the files are |

| | | |
|---|---|---|
| Layout | Train.csv, Validation.csv, image folders such as Train_deskewed, and a test split | images.zip, train.csv, val.csv, test.csv and the decoder notebook |
| Lines | 11,398 in training and validation | 11,398 |

**Decisions**

The alternatives include the choices AMIDDA made for its own KHATT copy (section 2.5).

| Decision | Reason | Alternative |
|---|---|---|
| Include KHATT as an optional subset | Modern handwriting, not manuscripts; kept "for pre-training or comparison" | Leave it out, or keep only its unique-text lines |
| Start from the Kaggle copy | It was the copy we found on Kaggle, together with a decoder for its labels | Ask KFUPM for KHATT v1.0 under its agreement, or take the Arabic-labelled copy that AMIDDA used, which seems to hold only the fixed paragraph |
| Leave out the upstream test split and hold out 104 of the 691 training writers (seed 3) as test | The upstream test split exposes no writer ID, so writers shared with training could not be ruled out | Keep the upstream test split, most likely KHATT's own test set, so that scores can be compared with other work on it. AMIDDA instead re-split its lines 80/10/10 at random, which can put the same writer in training and test |
| Keep the official validation split | Official, and writer-disjoint from training | Hold out validation writers from training in the same way as the test writers |

| | | |
|---|---|---|
| Decode the labels with 67 codes and skip ; | The upstream labels are codes, not Arabic; every code resolves | Use KFUPM's Arabic ground truth, or AMIDDA's labels, which write the Arabic comma in the fixed paragraph |
| Rename .tif to .jpg in the CSV files | The files are JPG | Leave the CSV files as they were and fix the names when loading |

**Limitations that remain**

- **About half the text is one paragraph.** 5,635 lines (49.4%) are the fixed paragraph, in all three splits. On these lines, a test score may show how well a model remembers a known text, not only how well it reads.
- **Overlap with AMIDDA.** AMIDDA's KHATT copy very likely holds all our fixed lines under other names. Do not mix the two without removing the khatt rows from AMIDDA.
- **Our own test set.** Its scores cannot be compared directly with results on KHATT's own test split.
- **Labels follow their own conventions.** They have Latin punctuation, Western digits, # (most likely for crossed-out words), extra spaces and what look like the writers' own mistakes, and they miss the commas that writers put in the fixed paragraph. Text should be normalised before scoring.
- **Lines missing inside paragraphs.** Many paragraphs have only some of their lines, and the copy does not say how they were chosen.
- **Images probably changed upstream.** The folder name suggests straightened JPG versions, not KFUPM's original scans, and the Kaggle copy does not document how they were made. The 12 we checked are level, almost black and white, and saved as JPG at quality about 75.
- **License.** KFUPM's terms allow academic, research and similar non-commercial use only, and do not allow passing the database to third parties. The Kaggle copy says "Unknown", and our collection publishes the images. This needs checking with KFUPM.

## 2.7 Muharaf

### 2.7.1 Description (original state, place in the collection)

Muharaf holds 24,495 line images of historic handwritten Arabic, each with its transcription. The documents are personal letters, diaries, poems, church records and legal records, mostly from the late 19th to the mid-20th century. Our copy is the authors' own line release, unchanged: the same lines, in the same order, with the official split of 22,092 training, 1,069 validation and 1,334 test lines. The only preparation was to store the images as PNG files in a zip and the labels in CSV files, like the other converted subsets.

**Original state**

Before any preprocessing, the authors publish the public part of Muharaf in two forms. We used the line release.

| Property | Page release (Zenodo) | Line release (Hugging Face) |
|---|---|---|
| Where | Zenodo record 11492215, version 1.0, 7 June 2024 | aamijar/muharaf-public, created 9 November 2024, last changed 24 January 2025 |
| What it holds | 1,216 page images with line polygons, transcriptions and page-element tags in JSON and PAGE-XML, plus line images, page summaries and the authors' training files (8 zip files, 10.6 GB) | 24,495 line images with their text (Parquet, 1.32 GB) |
| Fields per line | Polygon, text and page-element tags, in one file per page | image and text only |
| Split | Three random page splits of 1,100 / 50 / 66 pages, in the trial folders | Train 22,092, validation 1,069, test 1,334 lines |
| License shown | CC BY-NC-SA 2.0 Generic | CC BY-NC-SA 4.0 |

Sources: the paper, the authors' GitHub repository, the Zenodo record and the Hugging Face page.

The whole dataset has 1,644 page images and 36,311 text lines from 50 collections in two archives: the Phoenix Center for Lebanese Studies at the Holy Spirit University of Kaslik (USEK) and the Khayrallah Center for Lebanese Diaspora Studies at NC State University. The public part has 1,216 pages and 24,495 lines. The other 428 pages (11,816 lines) form a restricted part, given on request under a license that forbids passing them on. The documents date from the early 19th to the early 21st century, and most are written in Ruq'ah. The paper says about 21 page images are in Ottoman Turkish.

Experts in Lebanese Arabic transcribed the pages and a historian checked them; they also corrected machine drafts. Crossed-out text is marked but not transcribed. The line images come from the authors' line-straightening code: each line is cut along its polygon, straightened and scaled to 60 pixels high (see Images).

Four things are not documented:

- How the line release was split, and which of the paper's three random page splits it follows. Its line counts fit a split of 1,100 / 50 / 66 pages, with about 20 lines per page in each part.
- Which page and collection each line comes from. The line release has only the image and the text.
- The transcription rules, for example what a fatha on «س», the dots «...» or Latin codes such as «YS-4D-047» stand for (see Analysis).
- Which license applies. The Zenodo page shows CC BY-NC-SA 2.0, while the paper, GitHub and Hugging Face say CC BY-NC-SA 4.0.

The paper to cite is Saeed, Chan, Mijar, Moukarzel, Habchi, Younes, Elias, Wong and Khater, "Muharaf: Manuscripts of Handwritten Arabic Dataset for Cursive Text Recognition", NeurIPS 2024 Datasets and Benchmarks Track (arXiv 2406.09630).

**In the collection.** Muharaf is published as 24,495 line images with the official split of the authors' line release: 22,092 lines for training, 1,069 for validation and 1,334 for test.

### 2.7.2 Features (files, attributes, null and invalid values)

The published subset has two attributes, image_path and text, like the other converted subsets.

**Files**

| File | Size | Content |
|---|---|---|
| images.zip | 1,349,208,609 bytes | 24,495 PNG line images in images/train/, images/val/ and images/test/ |
| train.csv | 2,249,242 bytes | 22,092 rows of image_path,text |
| val.csv | 115,133 bytes | 1,069 rows |
| test.csv | 141,383 bytes | 1,334 rows |
| dataset_dict.json | 43 bytes | Only the split names: {"splits": ["train", "validation", "test"]} |

dataset_dict.json is the file the Hugging Face datasets library writes when it saves a dataset to disk. It looks left over from the conversion; the loader reads only the CSV files.

**Attributes**

| Attribute | Type | Meaning | Example |
|---|---|---|---|
| image_path | Text, 10 characters | File name of the line image: its row number in the split, in six digits, plus .png | 000000.png |
| text | Text | The Arabic transcription, as the source gives it | لبنان في ١٩٦٠/١٠/١٧ |

- image_path starts again at 000000.png in each split, so the same name exists in train, validation and test. A file is found only together with its split folder, for example images/test/000000.png.
- text is the source text unchanged, with its marks, punctuation and digits.

**Null and invalid values**

| Check | Train | Validation | Test |
|---|---|---|---|
| Empty text | 1 | 0 | 0 |
| Empty image_path | 0 | 0 | 0 |
| Shortest label (characters) | 1 | 1 | 1 |

One training label is empty; nothing else is missing. Four points need care:

- **One empty label.** One training row has no text. It is kept in the file.
- **Line breaks inside a few labels.** The card says a handful of labels contain a line break inside the text. We saw none in 300 sampled lines. A reader that splits on line breaks would cut these rows in two.
- **One-character labels.** Every split has labels of one character, such as page numbers like «٢».
- **No page or document column.** Lines cannot be grouped by page, document or collection, because neither the source nor our copy records them.

### 2.7.3 Analysis (split sizes, same lines as the source, line length, what the labels contain, overlap with AMIDDA, images)

Muharaf is the source's line release copied line for line. Its split is most likely by page, its labels keep each writer's and transcriber's habits, and 80% of its test lines also sit in AMIDDA's training split.

**Split sizes**

| Split | Lines | Share | Pages (paper) | Lines per page |
|---|---|---|---|---|
| Train | 22,092 | 90.2% | 1,100 | 20.1 |
| Validation | 1,069 | 4.4% | 50 | 21.4 |
| Test | 1,334 | 5.4% | 66 | 20.2 |
| All | 24,495 | 100% | 1,216 | 20.1 |

The page counts come from the paper's split of the public pages. The line release does not give its page counts, but its line counts fit them, so its split is most likely by page. Validation and test are small, 116 pages together. Because the pages were

drawn at random from 50 collections, pages from the same collection, often by the same writer, are very likely on both sides of the split.

**Same lines as the source**

We compared our copy with aamijar/muharaf-public in the dataset viewer. The row counts, the text-length statistics of all three splits (shortest, median, mean, longest and the 10-bin histograms) and the first 12 rows of each split are the same, in the same order. Both have the one empty training label. So no line was removed, added or edited; only the storage changed.

**Line length**

| Split | Shortest | Median | Mean | Longest |
|---|---|---|---|---|
| Train | 1 | 47 | 49.3 | 181 |
| Validation | 1 | 52 | 52.7 | 180 |
| Test | 1 | 48 | 52.3 | 177 |

Lengths are in characters, spaces included. Short lines are common: 3,745 training lines (17.0%) have fewer than 20 characters, such as dates, signatures and page numbers. Long lines are rare: 332 (1.5%) have 134 characters or more. Validation and test lines are a little longer on average. The labels hold about 1.2 million characters in all.

**What the labels contain**

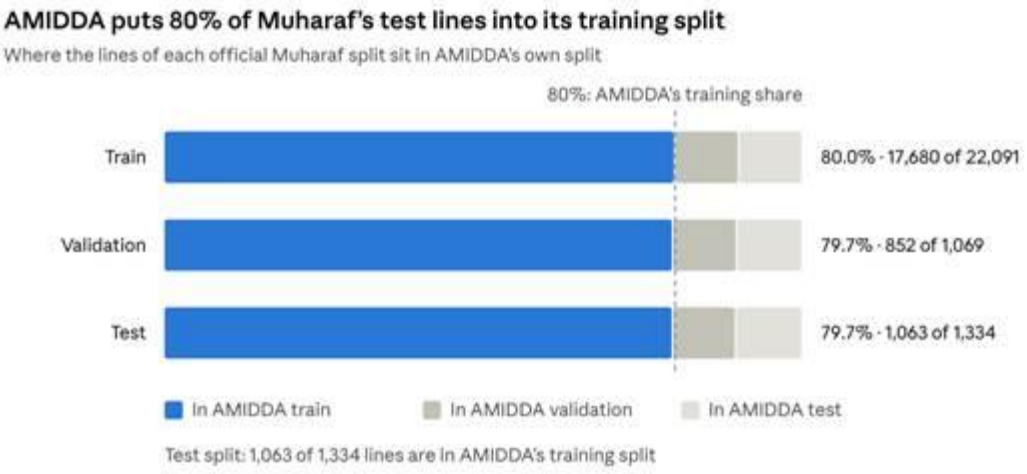
We read 300 lines in seven places across the three splits (212 training, 50 validation and 38 test lines). Lines read in one place often come from one document, so the shares below describe these documents, not the whole set.

| Feature | Lines in the sample | Example |
|---|---|---|
| Any vowel mark, shadda or tanween | 97 (32.3%) | «جاءَني», «ستّ» |
| A fatha on «س» | 45 (15.0%), 38 of them with no other mark | «سَلامي», «يوسَف» |
| Arabic-Indic digits | 44 (14.7%) | «في ١١/٩/١٩٨٥» |

| | | |
|---|---|---|
| Only a number | 10 | «٢», «(٦)» |
| Latin letters | 5 | «YS-4D-047», «.U. S. A» |
| Dots «..» or «...» | 27, 22 of them in one travel diary | «البعض من .. اهل ... الين معنا» |

- **The fatha on «س» looks like a habit, not a vowel.** 66 of the 81 fathas in the sample sit on «س». In some documents almost every «س» carries one, in others none does. It explains much of the card's "about 35% of lines have diacritics": 38 of the 97 marked lines in the sample have no other mark.
- **Tanween is typed in two orders.** It comes after the alef 18 times («جدًا», as in «جدًا») and before it 18 times («جنيهًا», as in «جنيهًا»), even in neighbouring lines of one letter. The two look the same on screen but count as different characters in a score.
- **Commas and digits are Arabic.** The sample has 42 Arabic commas «،» and one Latin comma. Western digits appear in only 5 lines, inside reference codes and one date.
- **Spelling stays as written:** colloquial words such as «منشان» and «شو», old spellings such as «وفآء».
- **Dots seem to stand for words that could not be read.** In a travel diary from 1893 they appear in 22 of 50 sampled lines, often in mid-sentence.
- **Not all text is Arabic.** Some lines are reference codes in Latin letters.

**Overlap with AMIDDA**




AMIDDA rows grouped by the official split named in their image_name, calfa-ai/AMIDDA dataset viewer (section 2.5) · 24,494 lines

AMIDDA (section 2.5) has its own copy of these lines and splits them again, 80/10/10 at random. Each of our splits is therefore spread over AMIDDA's three splits in the same way. Only one training line, most likely the empty one, is missing from AMIDDA. A model trained on AMIDDA cannot be scored fairly on our Muharaf test set, and a model trained on our Muharaf training split cannot be scored fairly on AMIDDA's test split.

**Images**

Every line image is 60 pixels high. The authors made them with the line-straightening code of their recognition system (SFR): a line is cut along its polygon, straightened along its baseline and scaled to 60 pixels (paper, section 5.4). The code works in squares of 60 × 60 pixels, so widths come in steps of 60: all 36 widths we read in the source's viewer are multiples of 60.

| Split | Narrowest | Median | Mean | Widest |
|---|---|---|---|---|
| Train | 60 | 600 | 573 | 1,620 |
| Validation | 60 | 600 | 564 | 1,260 |
| Test | 60 | 600 | 576 | 1,260 |

Widths are in pixels, from the source's dataset viewer; our copy has the same rows. images.zip holds about 55 KB per line, and its size is close to that of the source's Parquet files (1,349 MB against 1,317 MB), so the images were most likely exported as they were. These are not plain crops of the scans: a curved line becomes straight, large and small handwriting end up the same height, and the page images are not part of the subset.

### 2.7.4 Results (before and after, decisions, limitations that remain)

After preparation, Muharaf has the same 24,495 lines as the authors' line release, with the official split and the source text unchanged. Only the storage changed.

**Before and after**

| Property | Line release (aamijar/muharaf-public) | Published copy |
|---|---|---|

| | | |
|---|---|---|
| Format | Parquet files, each row holding the image and its text | images.zip with PNG files, plus train.csv, val.csv and test.csv |
| Lines | Train 22,092, validation 1,069, test 1,334 | The same lines, in the same order |
| Image reference | The image is stored in the row | image_path, the row number in its split (000000.png) |
| Text | As transcribed, with one empty training label | Unchanged; the empty label is kept |
| Size | 1,317,284,681 bytes | images.zip 1,349,208,609 bytes |
| Extra files | None | dataset_dict.json with the split names |

**Decisions**

| Decision | Reason | Alternative |
|---|---|---|
| Take the authors' line release, not the page release | "Line-level first. The strongest available training data is line-level" (proposal) | Cut lines from the Zenodo page images and polygons, which keeps page and collection IDs and allows other crops |
| Keep the official split | "Use official splits where available" (proposal) | Split again by page and collection, as the proposal asks for any new split; this needs the page release |
| Store the images as PNG files in a zip and the labels as image_path,text CSV files | One layout for all converted subsets, so they can be used together (card) | Keep the Parquet files, as was done for AMIDDA |

| | | |
|---|---|---|
| Keep every line and its text as the source gives it, including the empty label | Project rule: keep the original text and normalise only when scoring (card, proposal) | Drop the empty row, and make the tanween order and the fatha on «س» consistent |
| Warn against mixing Muharaf with AMIDDA | AMIDDA holds its own copy of these lines, split independently, so mixing them leaks lines across the split (card) | Remove the muharaf rows from AMIDDA, or split them again to follow the official split |

**Limitations**

- **Most likely split by page at random.** If so, pages from the same collection, often by the same writer, are in both training and test, and the test score shows how well a model reads new pages, not new writers.
- **No page or document ID.** Lines cannot be grouped by page, document or collection, so the split cannot be checked or rebuilt by document from our copy.
- **Small validation and test sets.** They hold 1,069 and 1,334 lines, most likely from 50 and 66 pages.
- **Label habits differ by document.** The fatha on «س», two orders of tanween, dots for words that were not read, Latin codes and spelling as written all count as errors in a strict score. Text should be normalised before scoring.
- **Straightened 60-pixel images.** The images come from the authors' line-straightening code, not straight from the scans, and the page images are not included.
- **Overlap with AMIDDA.** 79.7% of our test lines are in AMIDDA's training split.
- **Restricted part not included.** 428 pages (11,816 lines) are given only on request, under a license that forbids passing them on.
- **License and personal data.** The license is CC BY-NC-SA 4.0 (non-commercial, share-alike), with 2.0 on the Zenodo page. Some records are recent: a 1985 church letter in the training split gives a named man's birth date and marriage history. The authors' datasheet says the documents do not concern living people; this should be checked.

# 3. Summary Across the Subsets

This section puts the seven subsets side by side. It adds no new measurements: every figure comes from section 2.

## 3.1 The seven subsets side by side

| Subset | Size | Split | Labels |
|---|---|---|---|
| AHDQ | 642 pages | Our own split, 90 / 5 / 5, stratified by style | Fully diacritised text of a whole page |
| Omar-al-Saleh | 18,057 lines | Official train and test; validation is 5% of train, drawn by line; blind test left out | Mostly bare text; about 3% of the lines have diacritics; Eastern and Western digits |
| AraMs-28k-HTR | 27,969 lines | Official, by book (9 / 2 / 3) | Letters and spaces only: no diacritics, digits or punctuation |
| Muharaf | 24,495 lines | Official, most likely by page at random | As written: about a third of the lines we read have vowel marks; Arabic digits and punctuation |
| KHATT | 11,398 lines | Official validation; our test set holds out 104 writers | Decoded from Latin codes: Latin punctuation and digits, no vowel marks; 49.4% are pieces of one fixed paragraph |

| Subset | Size | Split | Labels |
|---|---|---|---|
| OpenITI-Makhzan | 23,079 lines on 1,497 pages | Three splits by document (98 / 54 / 56) | Diacritics generally not transcribed; seven languages |
| AMIDDA | 66,779 lines | By line, 80 / 10 / 10 inside each of eight datasets | Four labelling conventions |

## 3.2 Main findings

- **Some subsets are narrow.** Omar-al-Saleh has one writer. Muharaf is most likely split at random by page, so pages of the same collection probably stand on both sides. In AraMs-28k-HTR, 84% of the training split comes from books where only perfectly matched lines were kept.
- **The splits are not equally independent.** AraMs-28k-HTR is split by book, and the KHATT test set holds out whole writers. The validation set of Omar-al-Saleh was drawn by line, Muharaf is split by page (most likely at random), and AMIDDA by line, so on those sets a model is likely to meet documents, collections or hands that it also saw in training.
- **Label conventions differ.** AHDQ is fully diacritised, AraMs-28k-HTR has no marks, digits or punctuation, KHATT has Latin punctuation and no vowel marks, and OpenITI-Makhzan generally leaves diacritics out. A strict score therefore mixes reading errors with differences of convention.
- **AMIDDA overlaps two other subsets.** 79.7% of the Muharaf test lines (1,063 of 1,334) are in the training split of AMIDDA, and AMIDDA's KHATT copy very likely holds all 5,635 fixed-paragraph lines of our KHATT subset. AMIDDA must not be mixed with the Muharaf or KHATT subsets without filtering on dataset.
- **Some text repeats or is concentrated.** 49.4% of the KHATT lines are pieces of one fixed paragraph. Six shelfmarks hold 50.8% of the OpenITI-Makhzan training lines. Three BAYBARS notebooks hold 44% of the AMIDDA lines that have a document name.
- **A public test set can be contaminated.** The labels of the Omar-al-Saleh test split are public, so a system can have been trained on them.

- **Very short lines weigh differently.** 22.7% of the Omar-al-Saleh training lines have fewer than 15 characters, mostly calendar items, against 1.2% of the AraMs-28k-HTR lines with fewer than 14.

## 3.3 Licenses and redistribution

The table gives the license of each subset as stated at its source, with the points that are still open.

| Subset | License | Open point |
|---|---|---|
| AHDQ | MIT | None. |
| Omar-al-Saleh | CC BY 4.0, after accepting the terms of a gated dataset | The source is gated. It should be checked that its terms allow our copy to be public. The source asks to be cited through Hamoud et al. (2026). |
| AraMs-28k-HTR | CC BY-NC-SA 4.0 | Non-commercial and share-alike, so any use of the combined collection is non-commercial. |
| OpenITI-Makhzan | CC BY-NC-SA 4.0 | Our dataset card says the license is not verified, while the Zenodo record states it. The card is out of date. |
| Muharaf | CC BY-NC-SA 4.0 (2.0 on the Zenodo page) | The two versions disagree. The restricted part (428 pages) is not included. Some records are recent, and the authors' datasheet says the documents do not concern living people, which should be checked. |

| Subset | License | Open point |
|---|---|---|
| KHATT | KFUPM: academic, research and similar non-commercial use, and no passing on to third parties. The Kaggle copy says "Unknown" | Our collection publishes the images. This needs checking with KFUPM. |
| AMIDDA | Mixed: Apache 2.0 (RASAM 1, RASAM 2, TARIMA), Etalab 2.0 (BAYBARS, ISKANDAR), CC BY-NC-SA 4.0 (Muharaf), "not redistributed" (RASM), KHATT as above | RASM is marked "not redistributed" on the AMIDDA card, although its 1,488 lines are in the files. The licenses of RASM and KHATT are unclear. |

Three sources are share-alike and non-commercial (AraMs-28k-HTR, OpenITI-Makhzan and Muharaf), and KHATT is limited to research use, so the collection as a whole should be treated as non-commercial research data.

## 3.4 Splits and how to reproduce them

We made three of the splits ourselves: the split of AHDQ, the validation set of Omar-al-Saleh and the test set of KHATT. The table shows who made each split and what is needed to repeat it. KHATT ships the notebook that decodes its labels

(khatt-label-decoder.ipynb); the other preparation steps are described in the Results part of each section.

| Subset | Made by | Rule | Seed and record |
|---|---|---|---|
| AHDQ | Us | 90 / 5 / 5 by page, stratified by style | train.csv, val.csv, test.csv |
| Omar-al-Saleh | The source; us for validation | Official train and test; validation is 5% of the source train split, drawn by line | Seed 0; val.csv |
| AraMs-28k-HTR | The authors | By book: 9 train, 2 validation, 3 test | split_used.json |
| Muharaf | The authors | Official split of the line release | The three CSV files, in source order |
| KHATT | The source; us for the test set | Official validation split; the test set is 104 of the 691 upstream training writers, held out | Seed 3; test.csv |
| OpenITI-Makhzan | Not described in the data paper | Three splits of 98, 54 and 56 documents, used as shipped | train.csv, val.csv, test.csv |
| AMIDDA | Calfa, the source | 80 / 10 / 10 by line inside each of the eight datasets | The upstream Parquet files, unchanged |

## 3.5 Limits of this report

- **Samples.** Several findings rest on samples, and the text says so: 75 Omar-al-Saleh labels, 300 Muharaf lines, 12 KHATT images from two writers and 4 AraMs-28k-HTR line images.

- **Partial statistics.** For AMIDDA, the dataset viewer measures only the first 26,712 of the 53,420 training lines, so the training counts for KHATT and Muharaf are worked out from the sizes of those two datasets.
- **Checks that need the files.** The seven Omar-al-Saleh lines missing against the source, the spaces and line breaks in its labels, whether its file names repeat across the three splits, the number of joined words in AraMs-28k-HTR labels, whether the KHATT splits share a writer, and a match of AMIDDA's KHATT copy with ours by image.
- **Image sizes.** Pixel sizes are given only where this report measured them: the Muharaf and AMIDDA widths, the OpenITI-Makhzan line boxes and the samples above. They were not measured for AHDQ and Omar-al-Saleh.

# 4. References


1. Hamoud, H., et al. (2026). NAKBA NLP 2026: Shared Task on Arabic Handwritten Manuscript Understanding. Proceedings of NakbaNLP @ LREC 2026. https://aclanthology.org/events/nakbanlp-2026/
2. U4RASD, Doha Institute (2026). Omar Al-Saleh Manuscripts - Segments. Hugging Face Datasets. https://huggingface.co/datasets/U4RASD/omar-al-saleh-manuscripts-segments
3. Saeed, Chan, Mijar, Moukarzel, Habchi, Younes, Elias, Wong and Khater (2024). Muharaf: Manuscripts of Handwritten Arabic Dataset for Cursive Text Recognition. NeurIPS 2024 Datasets and Benchmarks Track. arXiv:2406.09630. Data: Zenodo record 11492215; line release aamijar/muharaf-public on Hugging Face.
4. XIDMA_AI (2024). AHDQ: Arabic Handwriting Dataset from Quran. Kaggle. https://www.kaggle.com/datasets/abdoukhadrembacke/ahdq-arabic-handwriting-dataset-from-quran
5. Guechaoui, M., Zellagui, M. D., Chaib, S. and Dhelim, S. (2026). AraMS-28k: The Largest Publicly Released Line-Level Dataset of Historical Arabic Manuscripts with Margin and Insertion-Anchor Annotations. arXiv:2608.26921. Data: Zenodo record 22095333. Code: https://github.com/ArchaText/AraMS-28k-Dataset
6. Allen, J. P., Mullan, J., Nigst, L., et al. (2026). OpenITI MAKHZAN: An Open Annotated Dataset of Arabic, Persian, Ottoman Turkish, and Urdu Print and Manuscript Data. Journal of Open Humanities Data, 12(1), article 69. https://doi.org/10.5334/johd.465. Data: Zenodo record 19861912, version 2026.1.2.
7. Vidal-Gorene, Lucas, Salah and Decours-Perez (2026). Is a Generic Dataset and Foundation VLM for Arabic HTR Worth It? Lessons from AMIDDA. Preprint hal-05693582, presented at HIP 2026 @ ICDAR 2026. Data: https://huggingface.co/datasets/calfa-ai/AMIDDA
8. Calfa. RASAM 1, RASAM 2, TARIMA, BAYBARS and ISKANDAR. Hugging Face (calfa-ai/RASAM-1, calfa-ai/RASAM-2, calfa-ai/tarima, calfa-ai/baybars, calfa-ai/iskandar) and GitHub (calfa-co/rasam-dataset, calfa-co/tarima).
9. British Library and PRImA Research Lab (2019). RASM2019: Competition on Recognition of Historical Arabic Scientific Manuscripts, ICDAR 2019. http://www.primaresearch.org/RASM2019/
10. Mahmoud and others (2014). KHATT: An Open Arabic Offline Handwritten Text Database. Pattern Recognition. https://doi.org/10.1016/j.patcog.2013.08.009. Copies used: Charrada, N. (2022), KHATT-Arabic, Kaggle (nizarcharrada/khattarabic); johnlockejrr/KHATT_v1.0_dataset on Hugging Face.
11. TheSeniorTeam. Arabic Manuscript Collection. Hugging Face Datasets, revision 809d854 (12 September 2026). https://huggingface.co/datasets/TheSeniorTeam/Arabic_Manuscript_Collection_Dataset
12. Project proposal of this senior project: Senior Project Documentation.