Stanford CS224v Course

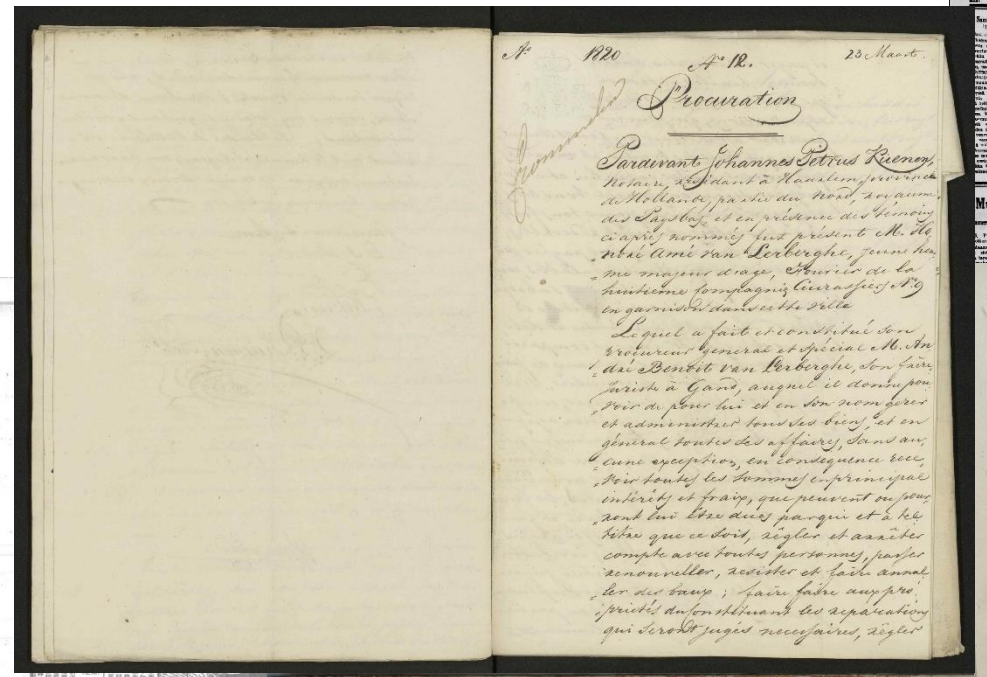
Conversational Virtual Assistants with Deep Learning

Lecture 10

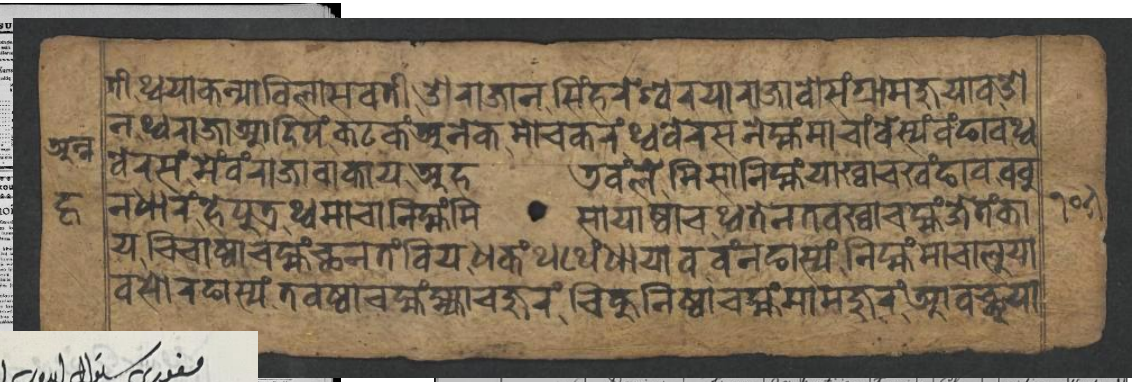
Vision-Language Models to Make Historical Documents Accessible

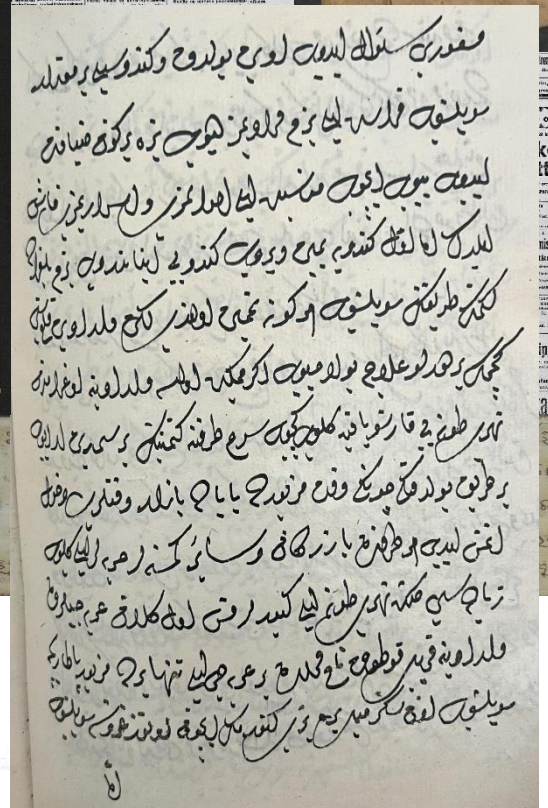
Sina Semnani & Monica Lam

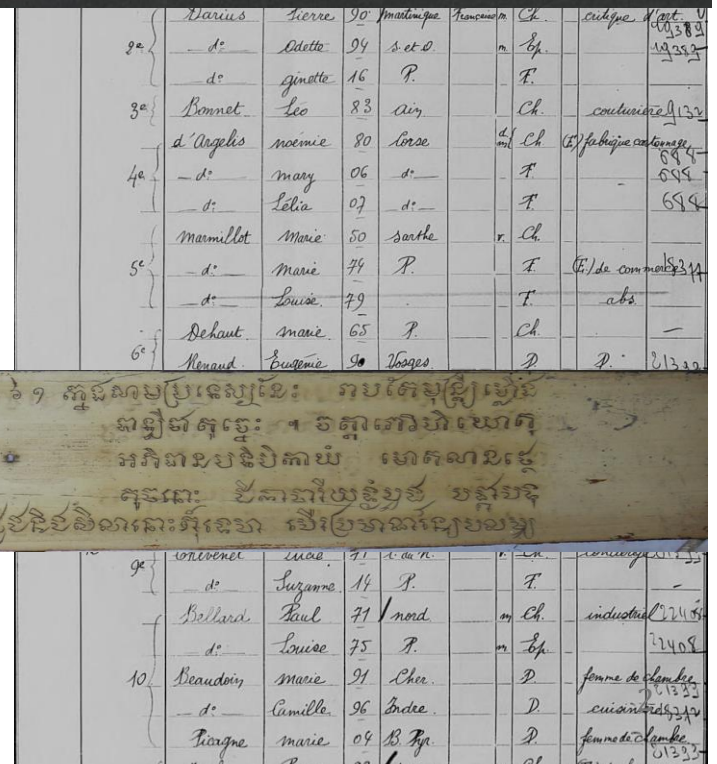
Thousands of Years of Cultural Heritage

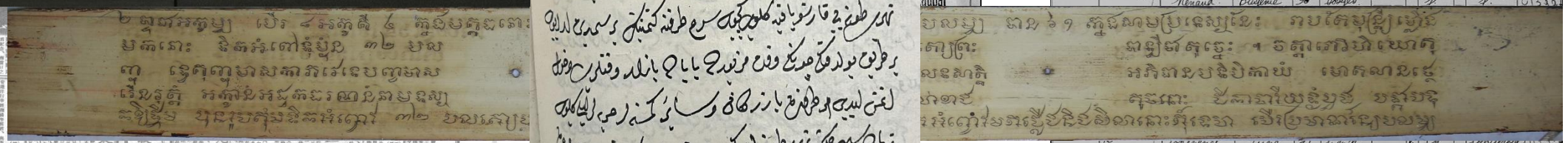


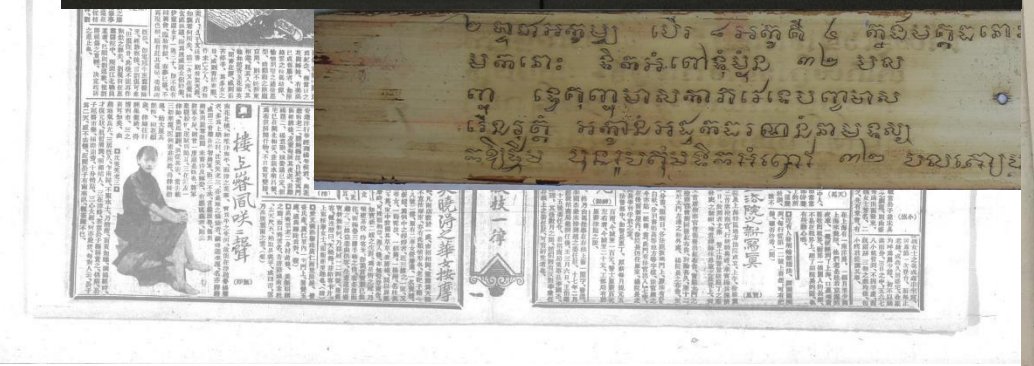




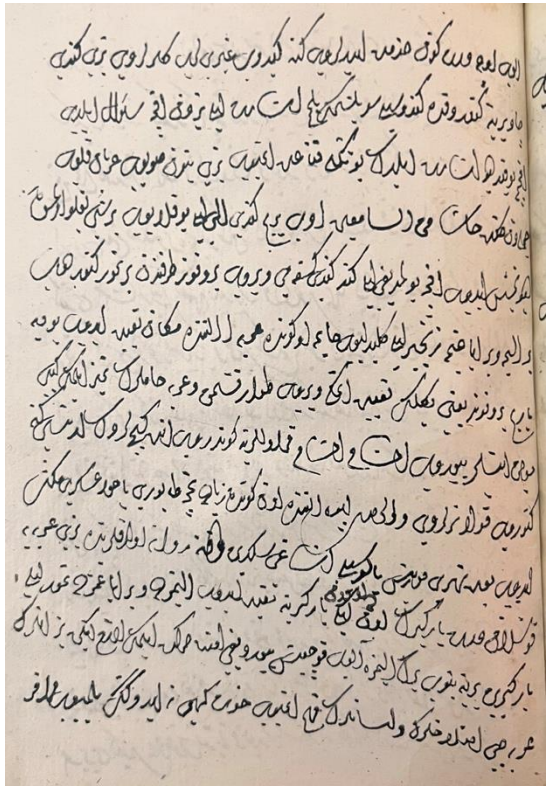








It Can Take Years Just to Read One Language



THE UNIVERSITY OF CHICAGO | DEPARTMENT OF MIDDLE EASTERN STUDIES
Division of the Arts & Humanities

Knowledge of **at least two years of modern Turkish is required** to study Ottoman. **Knowledge of Arabic and/or Persian is a great asset**. **First-year Ottoman Turkish offers an introduction** to the usage of the Arabic script in Ottoman and relevant facts of Arabic and Persian grammar. Readings are in **print**, generally with **examples from the 18th to early 20th centuries**. … **Intensive Ottoman summer course at** the ANAMED Center of Koç **University in Istanbul**.

Osman Aga of Temesvar, *Die Autobiographie des Dolmetschers Osman Aga aus Temeschwar*, ed. R. F. Kreutel, The E. J. W. Gibb Memorial Trust, Cambridge, 1980.

Many Historical Document Digitization Projects

But without text extraction, they are unsearchable and unreadable

LIBRARY OF CONGRESS

173 million items



The New York Times, October 31, 1915

The Stars and Stripes (Paris, France), May 10, 1918

New York Journal, January 3, 1897

The Minidoka Irrigator (Hunt, Idaho), January 15, 1944

Clipping from World War History, Nov. 14, 1918

BRITISH LIBRARY

The Endangered Archives Programme

Supported by Arcadia

Over 16 million images



Library of Congress, *Newspapers Collection Overview*, Library of Congress, 2025.
British Library, *Endangered Archives Programme*, British Library, 2025.

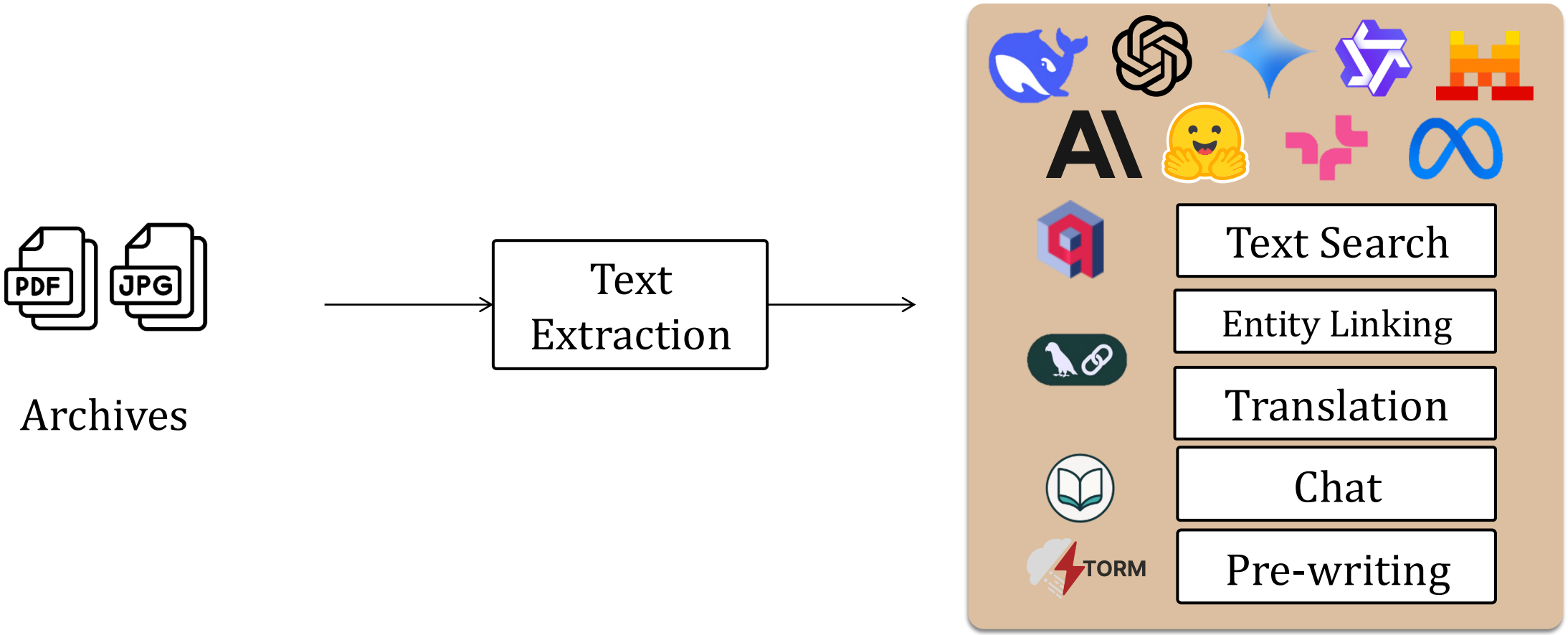
Transformation of mainstream narratives with historic evidence



Some examples from our colleagues

1	1842 – 1948	Hong Kong	Chinese and English newspapers since it became a British colony
2	1859 – 1861	Italy	Italian founding papers of unification
3	1884 – 1960s	West Africa	English newspapers on life not captured by history written by colonists
4	1895 – 1952	South Africa	Erased indigenous heritage: Khoi and San place names
5	1909 – 1979	California	Eugenic violence: forced sterilization of people with disabilities
6	1950s – 1970s	Caribbean	Operation Legacy: Destruction of British colonial records
7	1986	Ukraine	Documents from Chernobyl (recently destroyed by drones)

Text Extraction:
The Gateway to AI-Driven Knowledge Access

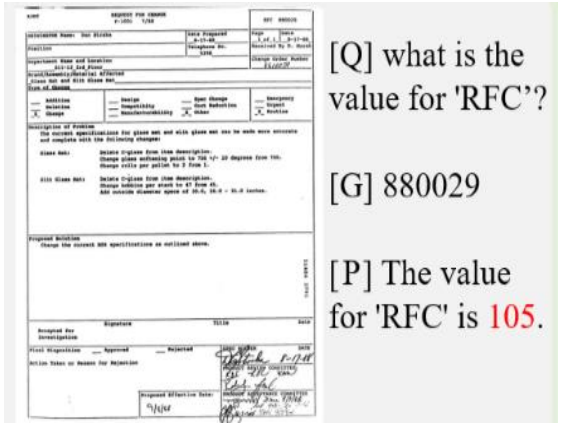


Previous Work

Existing Document Datasets



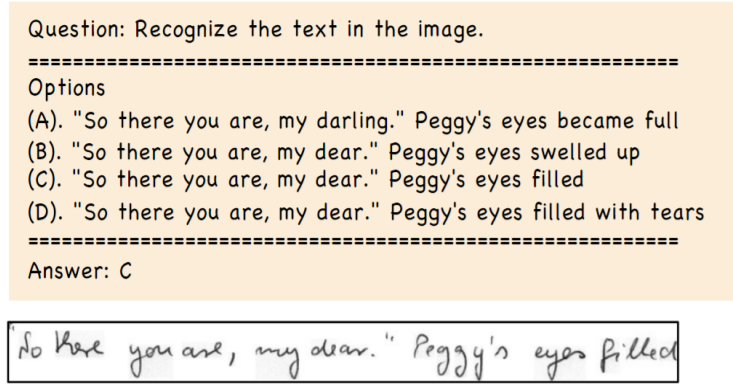
PubLayNet



OCRBench



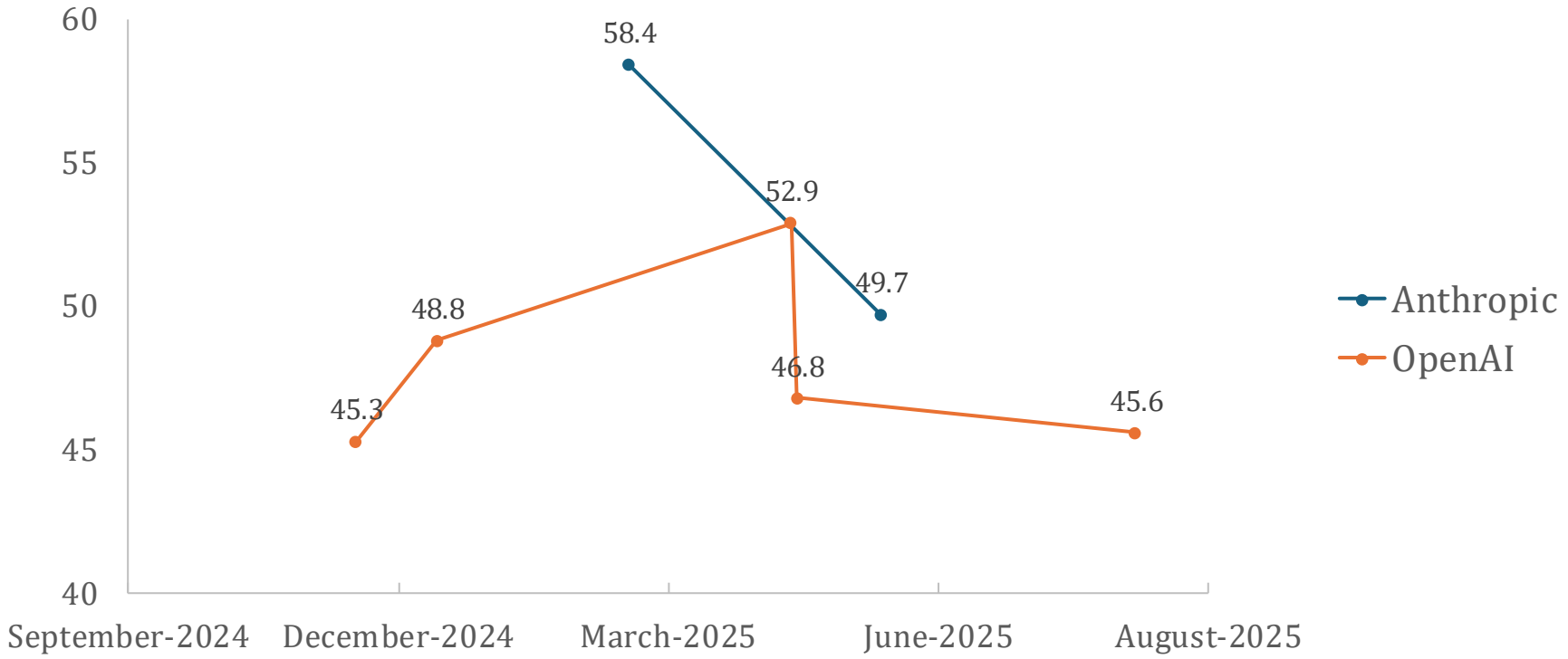
CC-OCR



MMT-Bench

Zhong et al., *PubLayNet: largest dataset ever for document layout analysis,* ICDAR, 2019.
Liu et al., *OCRBench: On the Hidden Mystery of OCR in Large Multimodal Models,* arXiv, 2024.
Yang et al., *CC-OCR: A Comprehensive and Challenging OCR Benchmark for Evaluating Large Multimodal Models in Literacy*, arXiv, 2024.
Ying et al., *MMT-Bench: A Comprehensive Multimodal Benchmark for Evaluating Large Vision-Language Models Towards Multitask AGI*, ICML, 2024.

Performance of Flagship Models on Churro-DS



LLMs Regress on Tasks Not Tracked by Model Developers

Text Extraction from Historical Corpora Is Challenging

Classic approach

- Manually annotate:
  - Text region boundaries
  - Text inside each region
  - Reading order between text regions
- Train a custom neural model
- Repeat for every new corpus

Major bottleneck

- Many historical datasets lack fine-grained annotations
- Many languages have no annotated datasets at all



Summary

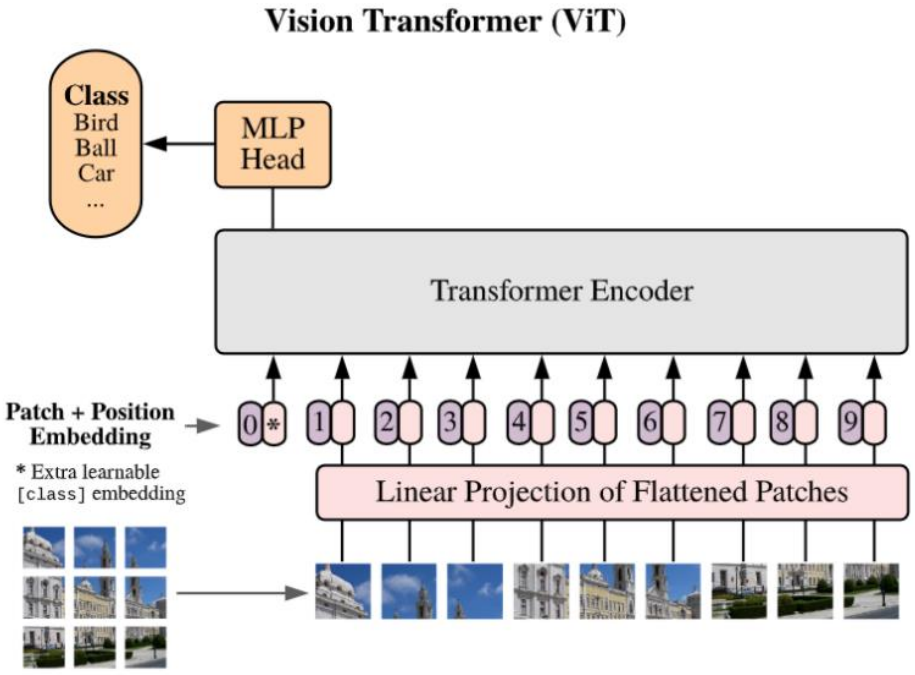
- Tons of “digitized” material – they are just images
- Many years of work to understand old text
- AI can make all this material easily accessible:
  - Read, search, translate, summarize, analyze, discover new insights
- Existing AI-based OCR tools need fine-grain annotations
  - Not enough training data

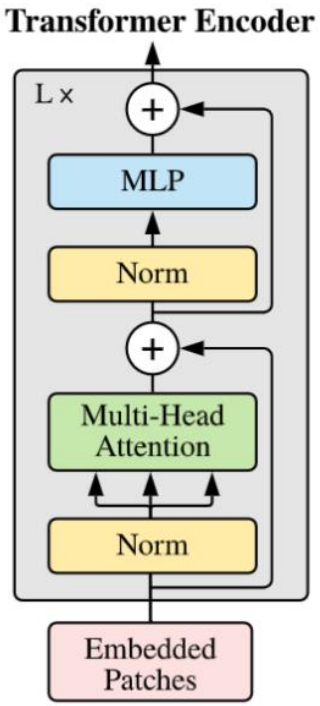
Key Idea

Source of valuable data:

Scholarly publications on history have page-level transcriptions!

Recent Breakthroughs in VLMs





Vision transformer [Dosovitskiy et al., 2021]

- 224 x 224 pixel inputs

Dynamic resolution [Dehghani et al., 2023]

- Qwen 2 VL [Wang et al., 2024]
- A pretrained VLM with dynamic resolution

Shift towards page-level OCR models:

- **CHURRO VLM [Semnani, 2025]**
- Qwen 2.5 VL [Bai et al., 2025]
- OlmOCR [Poznanski et al., 2025]

Dosovitskiy et al., An Image is Worth 16X16 Words: Transformers for Image Recognition at Scale, ICLR 2021.
Dehghani et al., Patch n' Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution, NeurIPS 2023.
Wang et al., Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution, arXiv 2024.
Bai et al., Qwen2.5-VL Technical Report, arXiv 2025
Poznanski et al., olmOCR: Unlocking Trillions of Tokens in PDFs with Vision Language Models, arXiv 2025.

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

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

In EMNLP 2025

CHURRO: Collection of Historical Unified Resources for Robust OCR

Lecture Goal



CHURRO: A Grass-Root Movement to **Make World History Readable**

Original Document

↓

Vision Language Model (VLM)

↓

Whole document in HDML
Historical Document Markup Language

2. Churro DS: An Open, Academic Historical Corpus

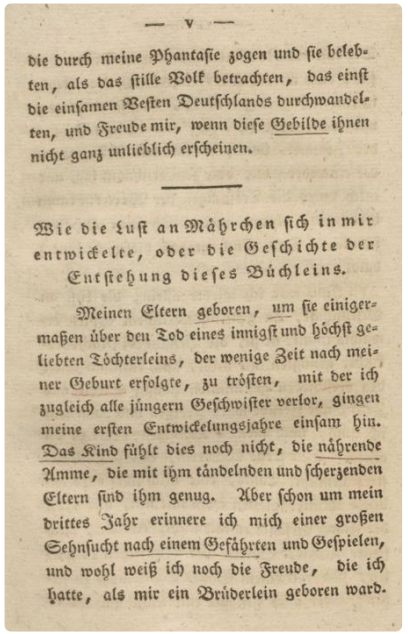
3. Churro VLM: Train an Open-Weight VLM

1. Design of HDML

Remove Fine-Grained Annotation Requirements

- Use Vision Language Models (VLMs) for full-page OCR
- No manual annotation of text regions required

Example 1



— v —

die durch meine Phantasie zogen und sie belebten, als das stille Volk betrachten, das einst die einsamen Westen Deutschlands durchwandelten, und Freude mir, wenn diese Gebilde ihnen nicht ganz unlieblich erscheinen.

Wie die Lust an Mährchen sich in mir entwickelte, oder die Geschichte der Entstehung dieses Büchleins.

Meinen Eltern geboren, um sie einigermaßen über den Tod eines innigst und höchst geliebten Töchterleins, der wenige Zeit nach meiner Geburt erfolgte, zu trösten, mit der ich zugleich alle jüngern Geschwister verlor, gingen meine ersten Entwickelungsjahre einsam hin. Das Kind fühlt dies noch nicht, die nährende Amme, die mit ihm tändelnden und scherzenden Eltern sind ihm genug. Aber schon um mein drittes Jahr erinnere ich mich einer großen Sehnsucht nach einem Gefährten und Gespielen, und wohl weiß ich noch die Freude, die ich hatte, als mir ein Brüderlein geboren ward.

VLM



V

die durch meine Phantasie zogen und sie belebten, als das stille Volk betrachten, das einst die einsamen Vesten Deutschlands durchwandelten, und Freude mir, wenn diese Gebilde ihnen nicht ganz unlieblich erscheinen.

Wie die Lust an Mährchen sich in mir entwickelte, oder die Geschichte der Entstehung dieses Büchleins.

Meinen Eltern geboren, um sie einigermaßen über den Tod eines innigst und höchst geliebten Töchterleins, der wenige Zeit nach meiner Geburt erfolgte, zu trösten, mit der ich zugleich alle jüngern Geschwister verlor, gingen meine ersten Entwickelungsjahre einsam hin. Das Kind fühlt dies noch nicht, die nährende Amme, die mit ihm tändelnden und scherzenden Eltern sind ihm genug. Aber schon um mein drittes Jahr erinnere ich mich einer großen Sehnsucht nach einem Gefährten und Gespielen, und wohl weiß ich noch die Freude, die ich hatte, als mir ein Brüderlein geboren ward.

Example 2

Einzelne Zeitungsseite,
Illustrierte Kronen-Zeitung vom 1. Juli 1933, Seite 10.



Illustrierte Kronen-Zeitung. Samstag, 1. Juli 1933.

ROTPUNKT-ZAHNPASTA

Volkssport

Fußball.

Hiden geht nach Paris.

Jimmy Hogan hat ihn engagiert.

Zwei Erstochene — vier Jahre Kerker.

Admira wirbt für Oesterreich.

Der Bericht eines Fremdenlegionärs aus Algier.

Zum Mitropa-Cup-Spiel Austria gegen Slavia.

Schwerathletik.

Mannschaftsmeisterschaft von Wien im Stemmen.

Wettspielberichte.

Full-page Extraction with Meta Data

Full Extracted Text: here

```
−<HistoricalDocument>
 −<Metadata>
    <Language>German</Language>
    <WritingDirection>ltr</WritingDirection>
   −<PhysicalDescription>
      Einzelne Zeitungsseite, Illustrierte Kronen-Zeitung vom 1. Juli 1933, Seite 10.
    </PhysicalDescription>
   −<Description>
      Nachrichten und Sportberichte aus der Illustrierten Kronen-Zeitung.
    </Description>
    <Script>Latin</Script>
  </Metadata>
 −<Page>
   −<Header>
    −<Heading type="other">
       <Line>Nr. 12.012. Seite 10.</Line>
      </Heading>
    −<Heading type="main">
       <Line>Illuſtrierte Kronen-Zeitung.</Line>
      </Heading>
    −<Heading type="other">
       <Line>Samstag, 1. Juli 1933.</Line>
      </Heading>
    </Header>
   −<Body>
    −<Paragraph>
       <Line>Angekl.: Ich bin im Jahre 1925 im All−</Line>
       <Line>gemein Krankenhaus vor Aerzten geprüft wor−</Line>
       <Line>den. In acht von zehn Fällen habe ich Krankheiten</Line>
       <Line>nur an dem Geruch des Harnes er−</Line>
       <Line>kannt.</Line>
      </Paragraph>
    −<Paragraph>
       <Line>Sachverſtändiger Prof. Dr. Wa−</Line>
       <Line>ſitzky: Ich erſehe aus einem Beſcheid, daß Sie</Line>
       <Line>in einem Fall eine Rückenmarksver−</Line>
       <Line>krümmung feſtgeſtellt haben. Sie erkannten</Line>
       <Line>ſogar, daß dieſes Leiden mit einem vor mehreren</Line>
       <Line>Jahren erfolgten Sturz oder Schlag zuſammen−</Line>
       <Line>hängt. Es iſt doch unmöglich, ein ſolches</Line>
       <Line>Leiden mit der Naſe feſtzuſtellen.</Line>
      </Paragraph>
```

Representation of the Text Extraction Result (OCR-Oriented)

- **Proposed Specifically for OCR**:
  - **ALTO**: Analyzed Layout and Text Object XML Schema by Library of Congress
  - **PAGE**: Page Analysis and Ground Truth Elements
  - **DocTags** Markup by IBM research
  - QwenVL **HTML** by Qwen Team, Alibaba Group
  - Limitations:
    - Represent the location of elements on the page without specifying their semantics or relationship
    - Have tons of optional syntax that are not used in practice
    - Do not account for “Gaps”, “Marginal notes”… that can impact the meaning of the content

Representation of the Text Extraction Result (Scholar Oriented)

- TEI (Text Encoding Initiative)
  - Widely used by libraries and archives, but not OCR systems
  - 500+ concepts
  - Multiple ways to encode the same page
  - Permits arbitrarily deep nesting
  - No native support for elements like Header, Footer, Margin

A New HDML (Historical Document Markup Language)

Design Goal: Suitable for training page-level text extraction systems

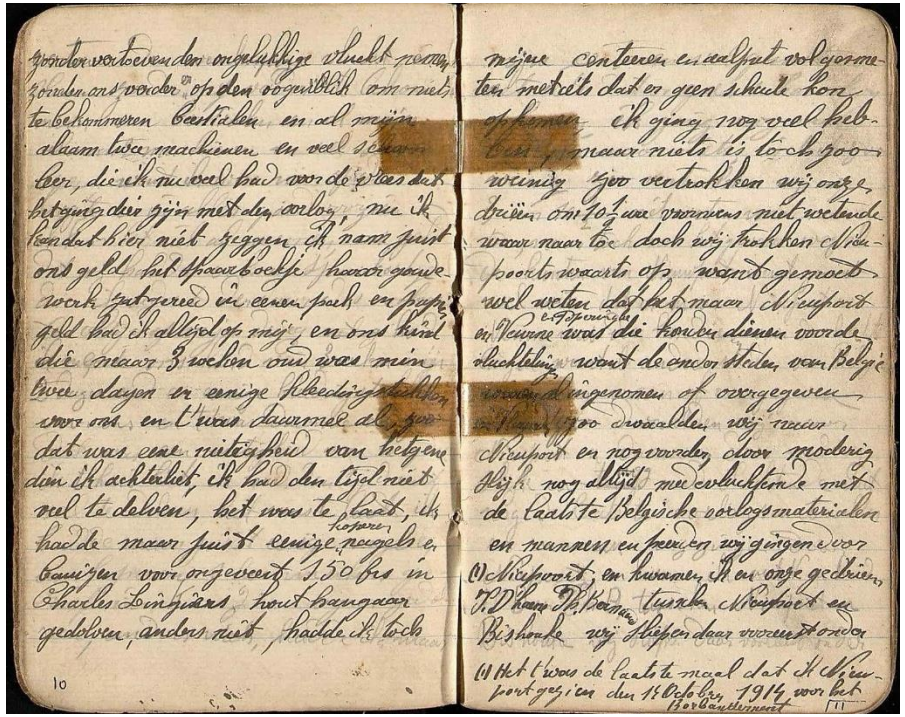
- Represent the whole document and nothing but the document
- Capture key information for downstream AI tasks
- Information available in many corpora -- No manual annotation needed
- Unique representation for training efficiency

Design: Refined from TEI

- A subset of TEI (45 concepts): Header, PageNumber, MarginalNote, Paragraph ...
- Provides a rough layout order -- no physical coordinate information

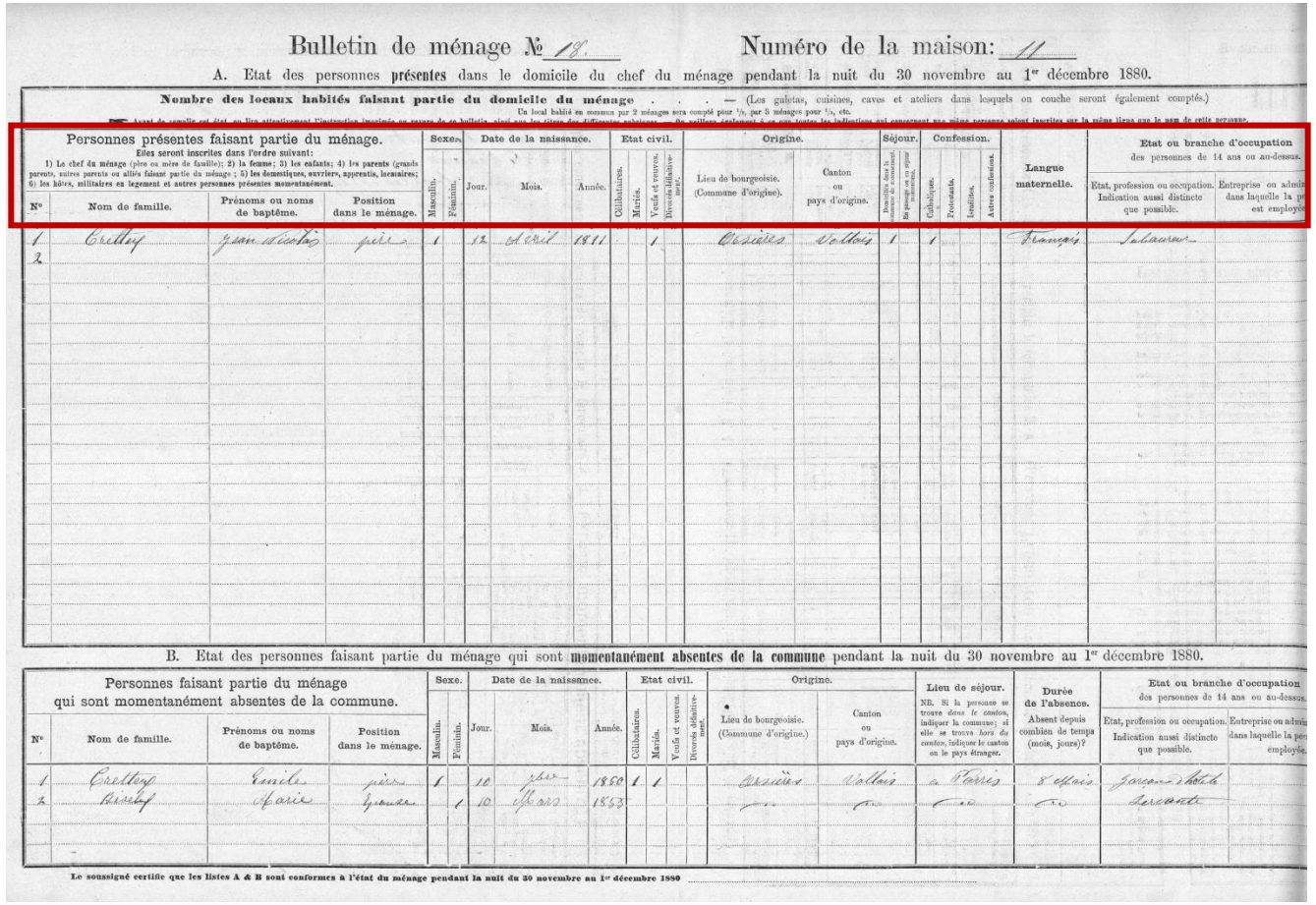
HDML: an abstract representation that enables automation
– a classical example of CS contributions

Metadata



```
<Metadata>
    <Language>Dutch</Language>
    <Script>Latin</Script>
    <WritingDirection>ltr</WritingDirection>
    <PhysicalDescription>
      Two pages from a handwritten diary, with adhesive tape causing some obstruction.
    </PhysicalDescription>
    <Description>
      Diary entries of Ernest Clarysse from October 1914,
      detailing his flight and observations during World War I.
    </Description>
  </Metadata>
```

Table



Bulletin de ménage № 18. Numéro de la maison: 11

A. Etat des personnes **présentes** dans le domicile du chef du ménage pendant la nuit du 30 novembre au 1er décembre 1880.

Nombre des locaux habités faisant partie du domicile du ménage . . . — (Les galetas, cuisines, caves et ateliers dans lesquels on couche seront également comptés.)

Un local habité en commun par 2 ménages sera compté pour 1/2, par 3 ménages pour 1/3, etc.

Personnes présentes faisant partie du ménage. Elles seront inscrites dans l'ordre suivant: 1) Le chef du ménage (père ou mère de famille); 2) la femme; 3) les enfants; 4) les parents (grands parents, autres parents ou alliés faisant partie du ménage; 5) les domestiques, ouvriers, apprentis, locataires; 6) les hôtes, militaires en logement et autres personnes présentes momentanément.

N°	Nom de famille.	Prénoms ou noms de baptême.	Position dans le ménage.	Sexe. Masculin.	Sexe. Féminin.	Date de la naissance. Jour.	Mois.	Année.	Etat civil. Célibataires.	Mariés.	Veufs et veuves.	Divorcés définitivement.	Origine. Lieu de bourgeoisie. (Commune d'origine).	Canton ou pays d'origine.	Séjour. Domicilié dans la commune de recensement.	Séjour. En passage ou en séjour momentané.	Confession. Catholiques.	Protestants.	Israélites.	Autres confessions.	Langue maternelle.	Etat ou branche d'occupation des personnes de 14 ans ou au-dessus. Etat, profession ou occupation. Indication aussi distincte que possible.	Entreprise ou admin dans laquelle la pe est employée.
1	Cretteux	Jean Nicolas	père	1		12	Avril	1811			1		Orsières	Vallais	1		1				Français	Sabaurou	
2																							

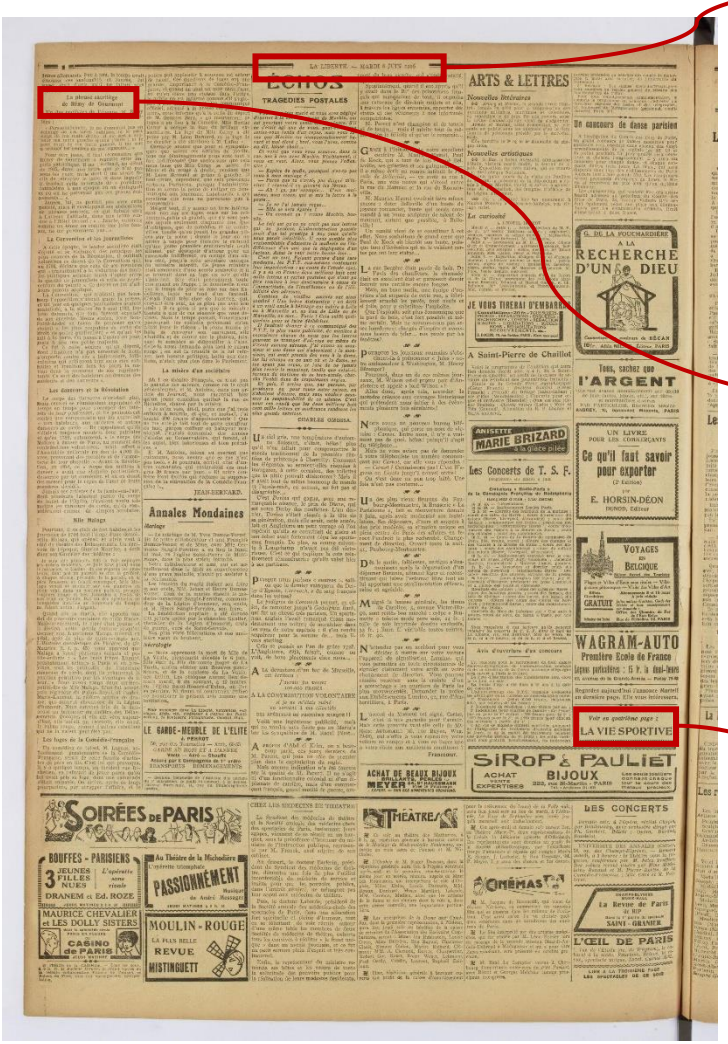
B. Etat des personnes faisant partie du ménage qui sont **momentanément absentes de la commune** pendant la nuit du 30 novembre au 1er décembre 1880.

N°	Personnes faisant partie du ménage qui sont momentanément absentes de la commune. Nom de famille.	Prénoms ou noms de baptême.	Position dans le ménage.	Sexe. Masculin.	Sexe. Féminin.	Date de la naissance. Jour.	Mois.	Année.	Etat civil. Célibataires.	Mariés.	Veufs et veuves.	Divorcés définitivement.	Origine. Lieu de bourgeoisie. (Commune d'origine.)	Canton ou pays d'origine.	Lieu de séjour. NB. Si la personne se trouve *dans le canton*, indiquer la commune; si elle se trouve *hors du canton*, indiquer le canton ou le pays étranger.	Durée de l'absence. Absent depuis combien de temps (mois, jours)?	Etat ou branche d'occupation des personnes de 14 ans ou au-dessus. Etat, profession ou occupation. Indication aussi distincte que possible.	Entreprise ou admin dans laquelle la per employée.
1	Cretteux	Emile	père	1		10	7bre	1850	1	1			Orsières	Vallais	à Paris	8 mois	Garçon d'hôtel	
2	Hiselef	Marie	épouse		1	10	Mars	1853					—	—	—	—	Servante	

Le soussigné certifie que les listes A & B sont conformes à l'état du ménage pendant la nuit du 30 novembre au 1er décembre 1880

```
<Table>
    <TableRow>
        <TableCell role="header" rowspan="2">
            <Line>Nº</Line>
        </TableCell>
        <TableCell role="header" rowspan="2">
            <Line>Nom de famille.</Line>
        </TableCell>
        <TableCell role="header" rowspan="2">
            <Line>Prénoms ou noms</Line>
            <Line>de baptême.</Line>
        </TableCell>
        <TableCell role="header" rowspan="2">
            <Line>Position</Line>
            <Line>dans le ménage.</Line>
        </TableCell>
        <TableCell role="header" colspan="2">
            <Line>Sexe.</Line>
            </TableCell>
        <TableCell role="header" colspan="3">
            <Line>Date de la naissance.</Line>
        </TableCell>
        <TableCell role="header" colspan="4">
            <Line>Etat civil.</Line>
        </TableCell>
        <TableCell role="header" colspan="2">
            <Line>Origine.</Line>
        </TableCell>
        <TableCell role="header" colspan="2">
            <Line>Séjour.</Line>
        </TableCell>
        <TableCell role="header" colspan="4">
            <Line>Confession.</Line>
        </TableCell>
        <TableCell role="header" rowspan="2">
            <Line>Langue</Line>
            <Line>maternelle.</Line>
        </TableCell>
        <TableCell role="header" colspan="2">
            <Line>Etat ou branche d'occupation</Line>
            <Line>des personnes de 14 ans ou au-dessus.</Line>
        </TableCell>
    </TableRow>
     ...
</Table>
```

Headings

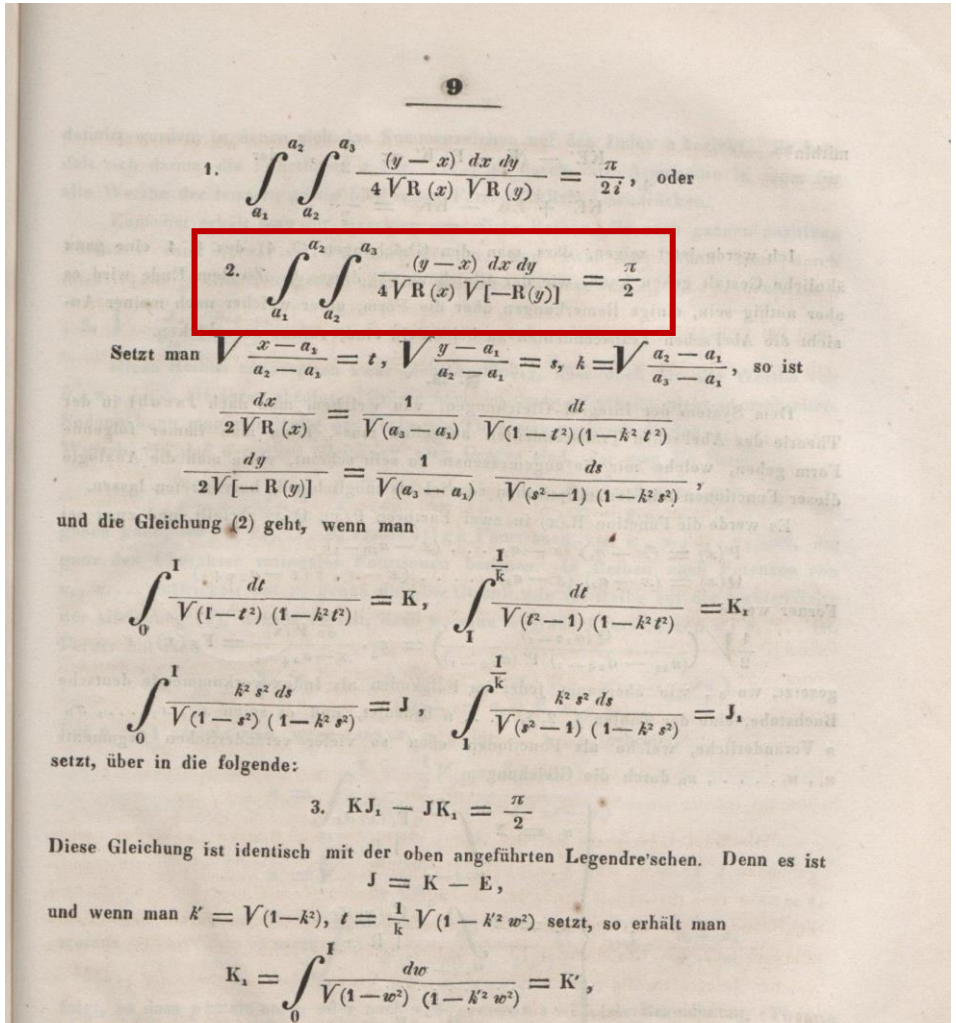



```
<Heading type="main">
    <Line>LA LIBERTÉ — MARDI 8 JUIN 1926</Line>
</Heading>
```

```
<Heading type="sub">
    <Line>La phrase sacrilège</Line>
    <Line>de Rémy de Gourmont</Line>
</Heading>
```

```
<Heading type="other">
    <Line>Voir en quatrième page :</Line>
    <Line>LA VIE SPORTIVE</Line>
</Heading>
```

Math!



9

1. $$\int_{a_1}^{a_2}\int_{a_2}^{a_3} \frac{(y-x)\,dx\,dy}{4\sqrt{R(x)}\,\sqrt{R(y)}} = \frac{\pi}{2i}, \text{ oder}$$

2. $$\int_{a_1}^{a_2}\int_{a_2}^{a_3} \frac{(y-x)\,dx\,dy}{4\sqrt{R(x)}\,\sqrt{[-R(y)]}} = \frac{\pi}{2}$$

Setzt man $\sqrt{\frac{x-a_1}{a_2-a_1}} = t$, $\sqrt{\frac{y-a_1}{a_2-a_1}} = s$, $k = \sqrt{\frac{a_2-a_1}{a_3-a_1}}$, so ist

$$\frac{dx}{2\sqrt{R(x)}} = \frac{1}{\sqrt{(a_3-a_1)}} \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}}$$

$$\frac{dy}{2\sqrt{[-R(y)]}} = \frac{1}{\sqrt{(a_3-a_1)}} \frac{ds}{\sqrt{(s^2-1)(1-k^2s^2)}},$$

und die Gleichung (2) geht, wenn man

$$\int_0^1 \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}} = K, \quad \int_1^{\frac{1}{k}} \frac{dt}{\sqrt{(t^2-1)(1-k^2t^2)}} = K_1$$

$$\int_0^1 \frac{k^2s^2\,ds}{\sqrt{(1-s^2)(1-k^2s^2)}} = J, \quad \int_1^{\frac{1}{k}} \frac{k^2s^2\,ds}{\sqrt{(s^2-1)(1-k^2s^2)}} = J_1$$

setzt, über in die folgende:

3. $KJ_1 - JK_1 = \frac{\pi}{2}$

Diese Gleichung ist identisch mit der oben angeführten Legendre'schen. Denn es ist

$$J = K - E,$$

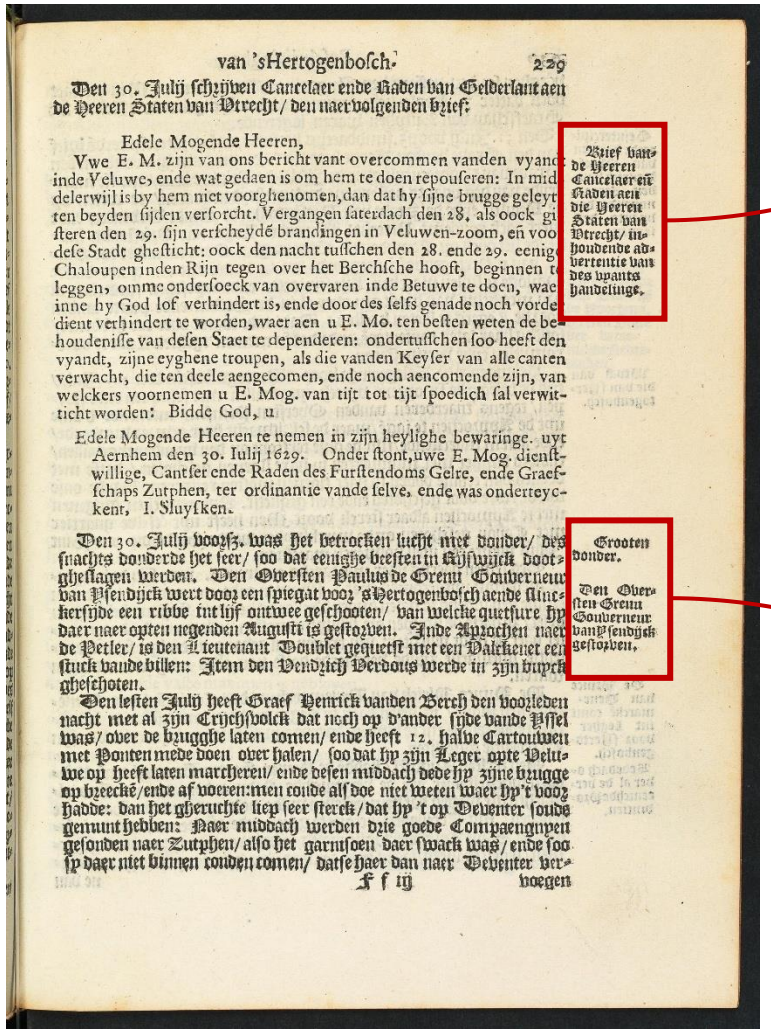
und wenn man $k' = \sqrt{(1-k^2)}$, $t = \frac{1}{k}\sqrt{(1-k'^2w^2)}$ setzt, so erhält man

$$K_1 = \int_0^1 \frac{dw}{\sqrt{(1-w^2)(1-k'^2w^2)}} = K',$$

```
<Line>
    2.
    <Formula notation="latex">
        \int\limits_{a_1}^{a_2} \int\limits_{a_2}^{a_3}
        \frac{(y-x) dx dy}{4 \sqrt{R(x)}
        \sqrt{[-R(y)]}} = \frac{\pi}{2}
    </Formula>

</Line>
```

Marginal Notes



van 'sHertogenbosch. 229

Den 30. Julij schrijven Cancelaer ende Raden van Gelderlant aen de Heeren Staten van Vtrecht/ den naervolgenden brief:

Edele Mogende Heeren,

Vwe E. M. zijn van ons bericht vant overcommen vanden vyand inde Veluwe, ende wat gedaen is om hem te doen repouseren: In middelerwijl is by hem niet voorghenomen, dan dat hy sijne brugge geleyt ten beyden sijden versorcht. Vergangen saterdach den 28. als oock gisteren den 29. sijn verscheydē brandingen in Veluwen-zoom, eñ voor dese Stadt ghesticht: oock den nacht tusschen den 28. ende 29. eenige Chaloupen inden Rijn tegen over het Berchsche hooft, beginnen te leggen, omme ondersoeck van overvaren inde Betuwe te doen, waer inne hy God lof verhindert is, ende door des selfs genade noch vorder dient verhindert te worden, waer aen u E. Mo. ten besten weten de behoudenisse van desen Staet te dependeren: ondertusschen soo heeft den vyandt, zijne eyghene troupen, als die vanden Keyser van alle canten verwacht, die ten deele aengecomen, ende noch aencomende zijn, van welckers voornemen u E. Mog. van tijt tot tijt spoedich sal verwitticht worden: Bidde God, u

Edele Mogende Heeren te nemen in zijn heylighe bewaringe. uyt Aernhem den 30. Iulij 1629. Onder stont, uwe E. Mog. dienstwillige, Cantser ende Raden des Furstendoms Gelre, ende Graefschaps Zutphen, ter ordinantie vande selve, ende was onderteyckent, I. Sluysken.

Den 30. Julij voorsz. was het betrocken lucht met donder/ des snachts donderde het seer/ soo dat eenighe beesten in Rijswijck dootghe slagen werden. Den Oversten Paulus de Grenu Gouverneur van Ysendijck wert door een spiegat voor 'sHertogenbosch aende slinckersijde een ribbe int lijf ontwee geschooten/ van welcke quetsure hy daer naer opten negenden Augusti is gestorven. Inde Aprochen naer de Petler/ is den Lieutenant Doublet gequetst met een Valckenet een stuck vande billen: Item den Vendrich Verdous werde in zijn buyck gheschoten.

Den lesten Julij heeft Graef Henrick vanden Berch den voorleden nacht met al zijn Crijchsvolck dat noch op d'ander sijde vande Yssel was/ over de brugghe laten comen/ ende heeft 12. halve Cartouwen met Ponten mede doen over halen/ soo dat hy zijn Leger opte Veluwe op heeft laten marcheren/ ende desen middach dede hy zijne brugge op breeckē/ ende af voeren: men conde als doe niet weten waer hy't voor hadde: dan het gheruchte liep seer sterck/ dat hy 't op Deventer soude gemunt hebben: Naer middach werden drie goede Compaengnyen gesonden naer Zutphen/ also het garnisoen daer swack was/ ende soo sy daer niet binnen conden comen/ datse haer dan naer Deventer ver-

F f iij voegen

Brief vande Heeren Cancelaer eñ Raden aen die Heeren Staten van Vtrecht/ inhoudende advertentie van des vyants handelinge.

Grooten donder.

Den Oversten Grenu Gouverneur van Ysendijck gestorven.

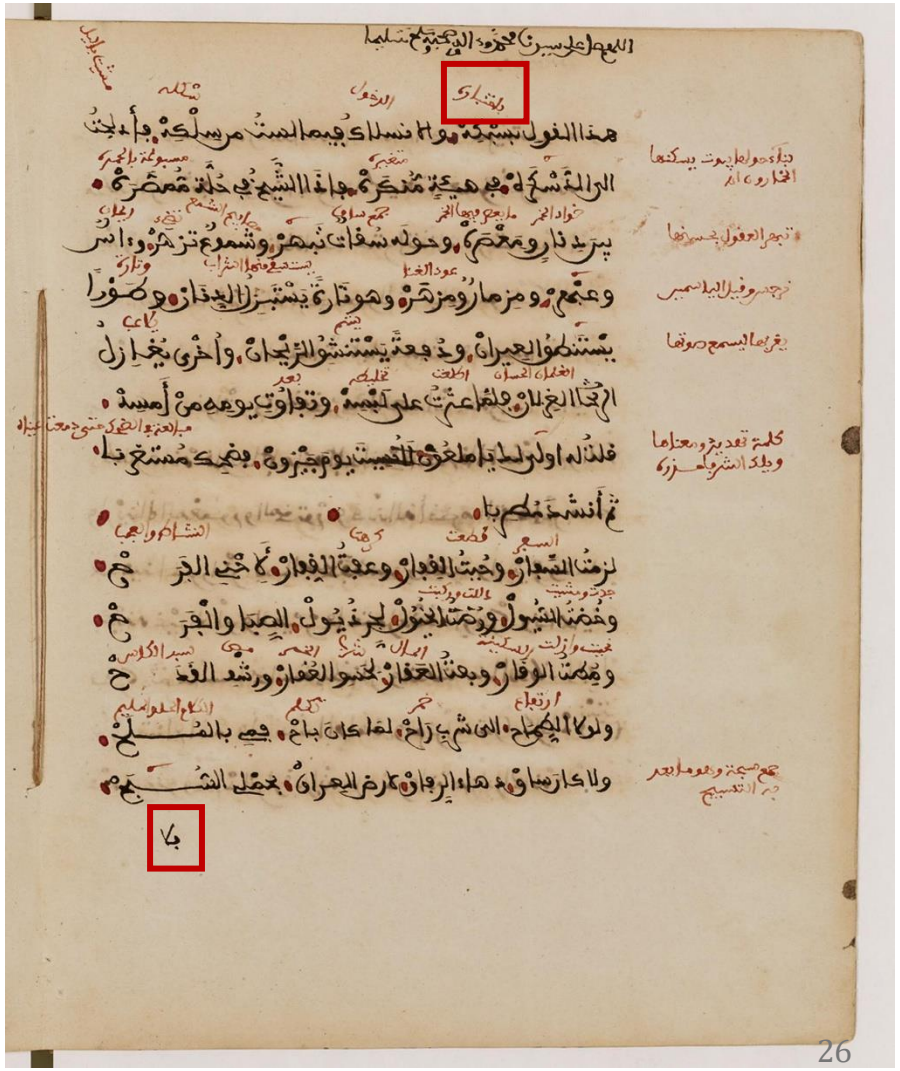
```
<MarginalNote placement="right_margin">
    <Line>Brief van-</Line>
    <Line>de Heeren</Line>
    <Line>Cancelaer eñ</Line>
    <Line>Raden aen</Line>
    <Line>die Heeren</Line>
    <Line>Staten van</Line>
    <Line>Vtrecht/ iñ-</Line>
    <Line>houdende ad-</Line>
    <Line>vertentie van</Line>
    <Line>des vyants</Line>
    <Line>handelinge.</Line>
</MarginalNote>
```

```
<MarginalNote placement="right_margin">
    <Line>Grooten</Line>
    <Line>donder.</Line>
    <Line>Den Over-</Line>
    <Line>ſten Grenu</Line>
    <Line>Gouverneur</Line>
    <Line>van Yſendijck</Line>
    <Line>geſtorven.</Line>
</MarginalNote>
```

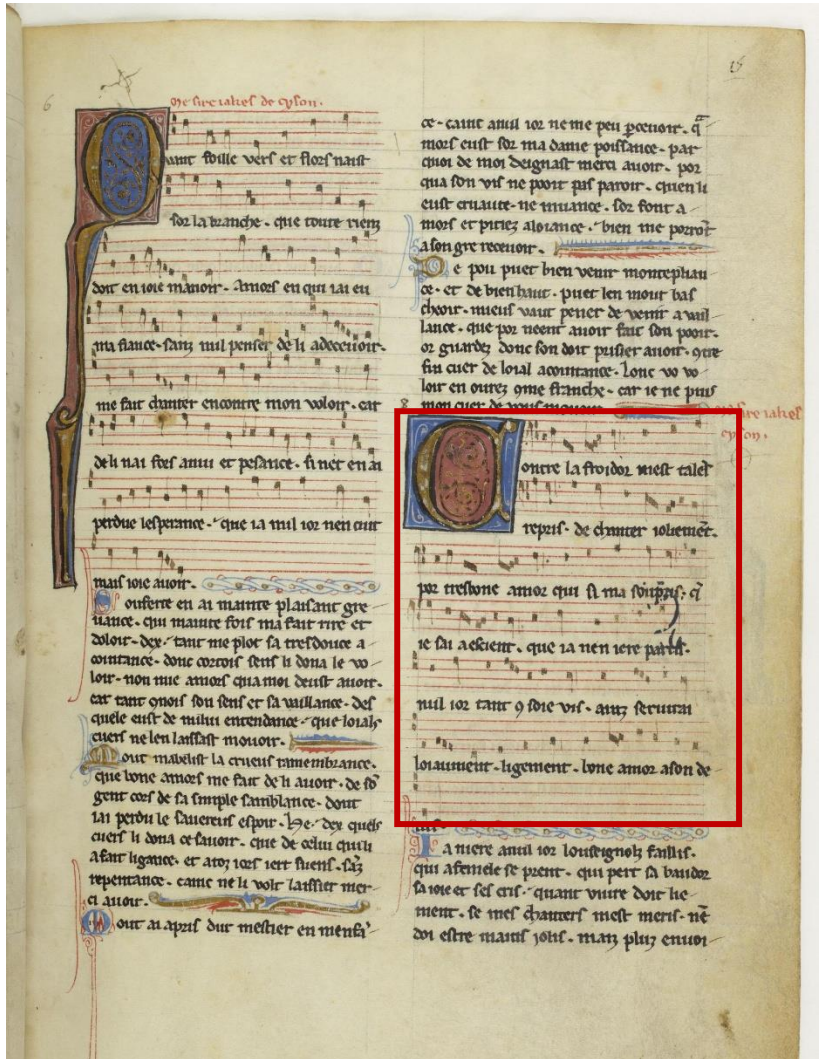
Interlinear Notes and Catchword

```
<Line>
    هذا القول بسبكه والانسلاك فيما لست من سلكه فأدلجت
    <InterlinearNote>
        <Emphasis type="colored">
            باقتار
        </Emphasis>
    </InterlinearNote>
</Line>
```

```
<Footer>
    <CatchWord>بلا</CatchWord>
</Footer>
```



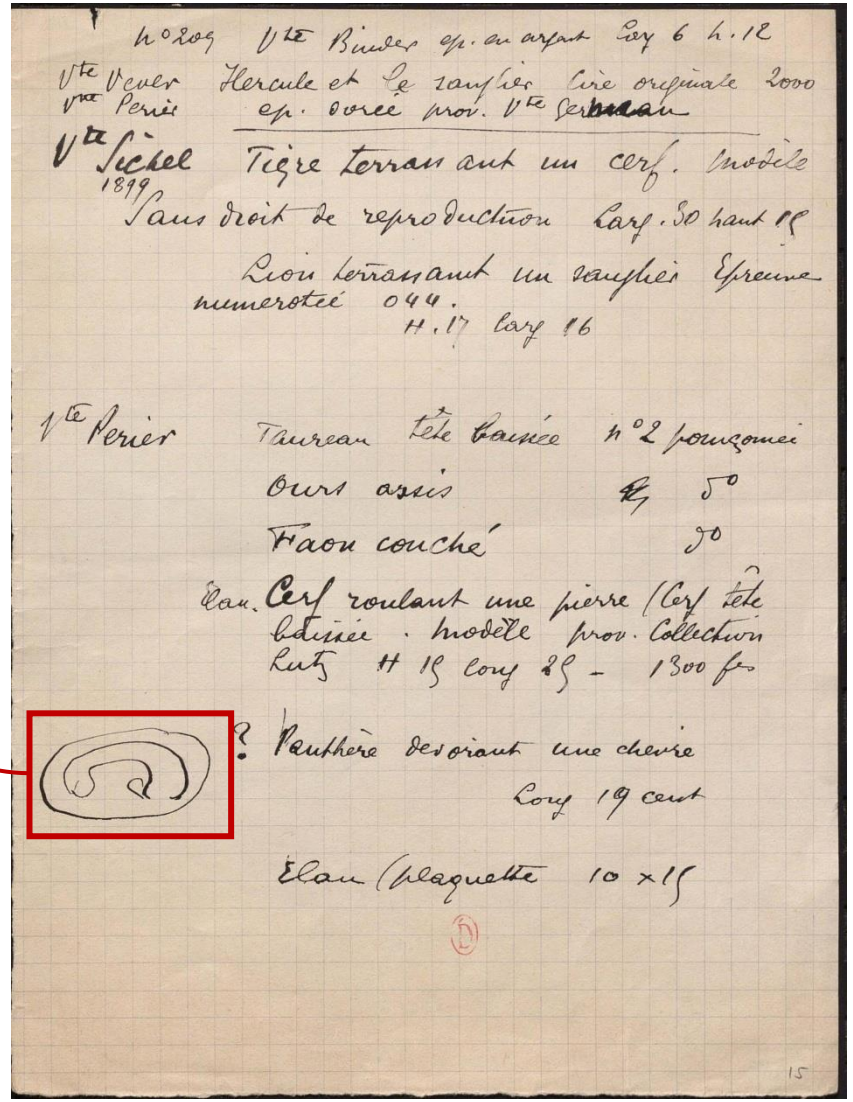
Decorations and Musical Notation



```
<MusicalNotation/>
<Paragraph>
    <Line><Initial type="decorated">C</Initial>ontre la froidor mest talẽs</Line>
</Paragraph>
<MusicalNotation/>
<Paragraph>
    <Line>epris · de chanter ioliemẽt ·</Line>
</Paragraph>
<MusicalNotation/>
<Paragraph>
    <Line>por tresbone amor qui si ma soupris · q̃</Line>
</Paragraph>
<MusicalNotation/>
<Paragraph>
<Line>ie sai aescient · que ia nen iere parti ·</Line>
</Paragraph>
<MusicalNotation/>
<Paragraph>
    <Line>nul ior tant 9 soie vis · ainz seruirai<Above>s</Above></Line>
</Paragraph>
<MusicalNotation/>
<Paragraph>
    <Line>loiaument ·ligement · bone amor ason de</Line>
</Paragraph>
<MusicalNotation/>
```

Figures

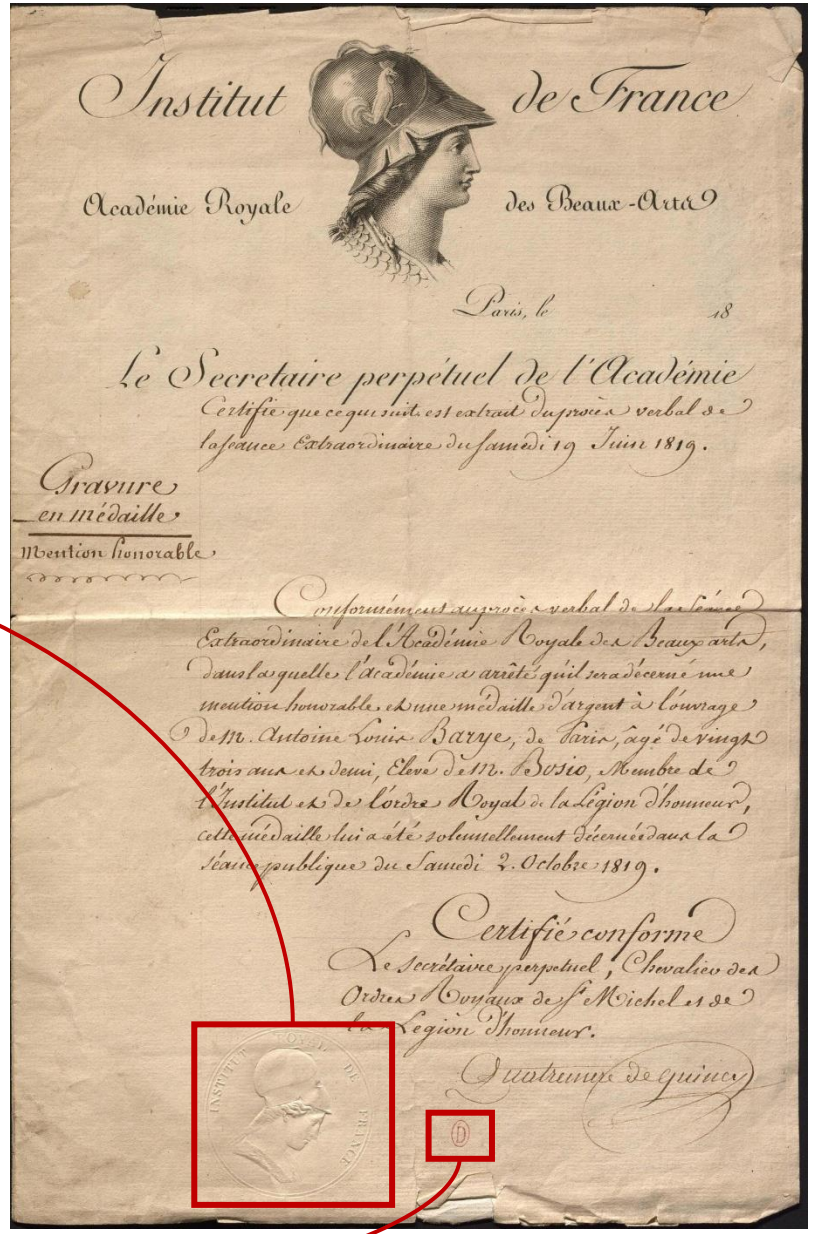
```
<Figure>
    <Description>
        Sketch of a stylized, curled abstract form,
        possibly a decorative motif or a simplified animal representation.
    </Description>
</Figure>
```



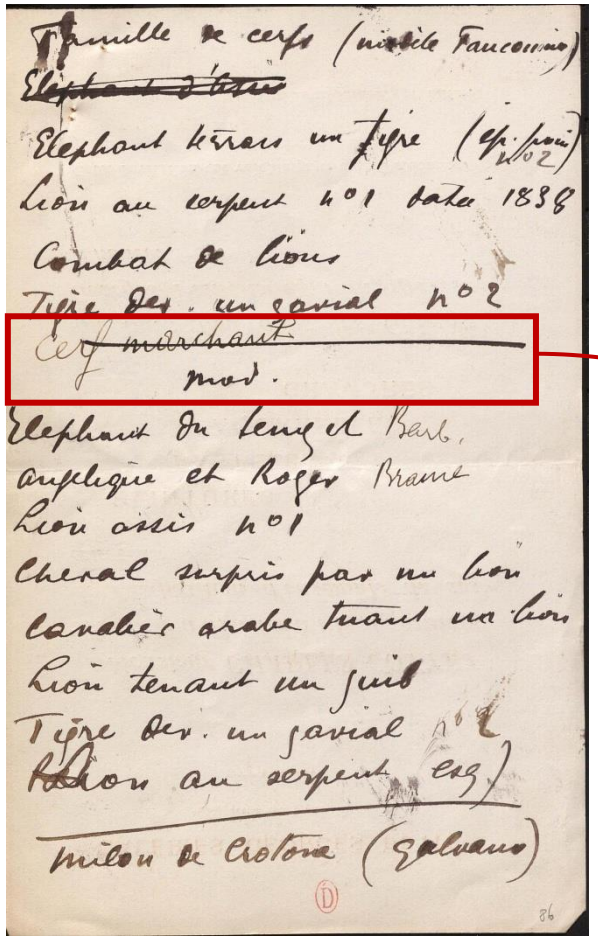
Stamps and Seals

```
<Seal description="Embossed seal: INSTITUT ROYAL DE FRANCE"/>
```

```
<Stamp description="Letter D in a circle, possibly a paraph or control mark."/>
```

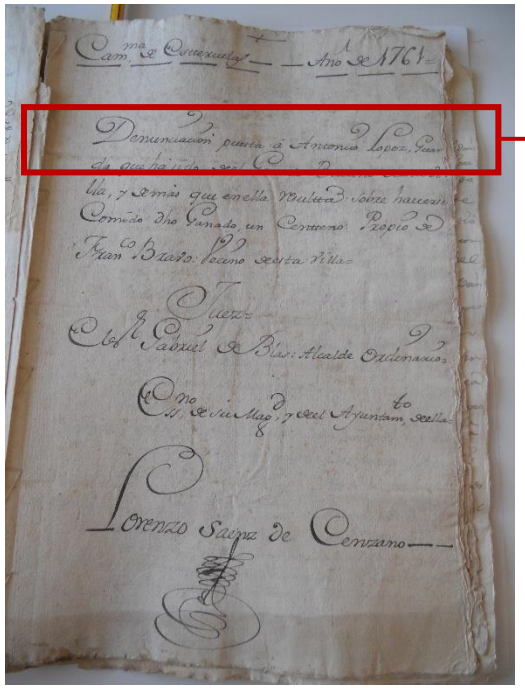


Deletion and Addition

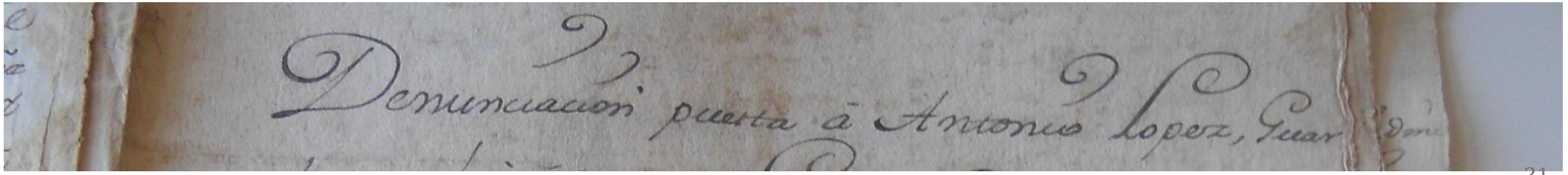



```
<Item>
    <Line><Deletion>Cerf marchant</Deletion></Line>
    <Line><Addition>mod.</Addition></Line>
</Item>
```

Damages and Gaps



```
<Line>Denunciacion puesta à Antonio Lopez, Guar<Gap reason="damaged" extent="da"/></Line>
```



Lecture Goal



CHURRO: A Grass-Root Movement to **Make World History Readable**

Original Document

↓

Vision Language Model (VLM)

↓

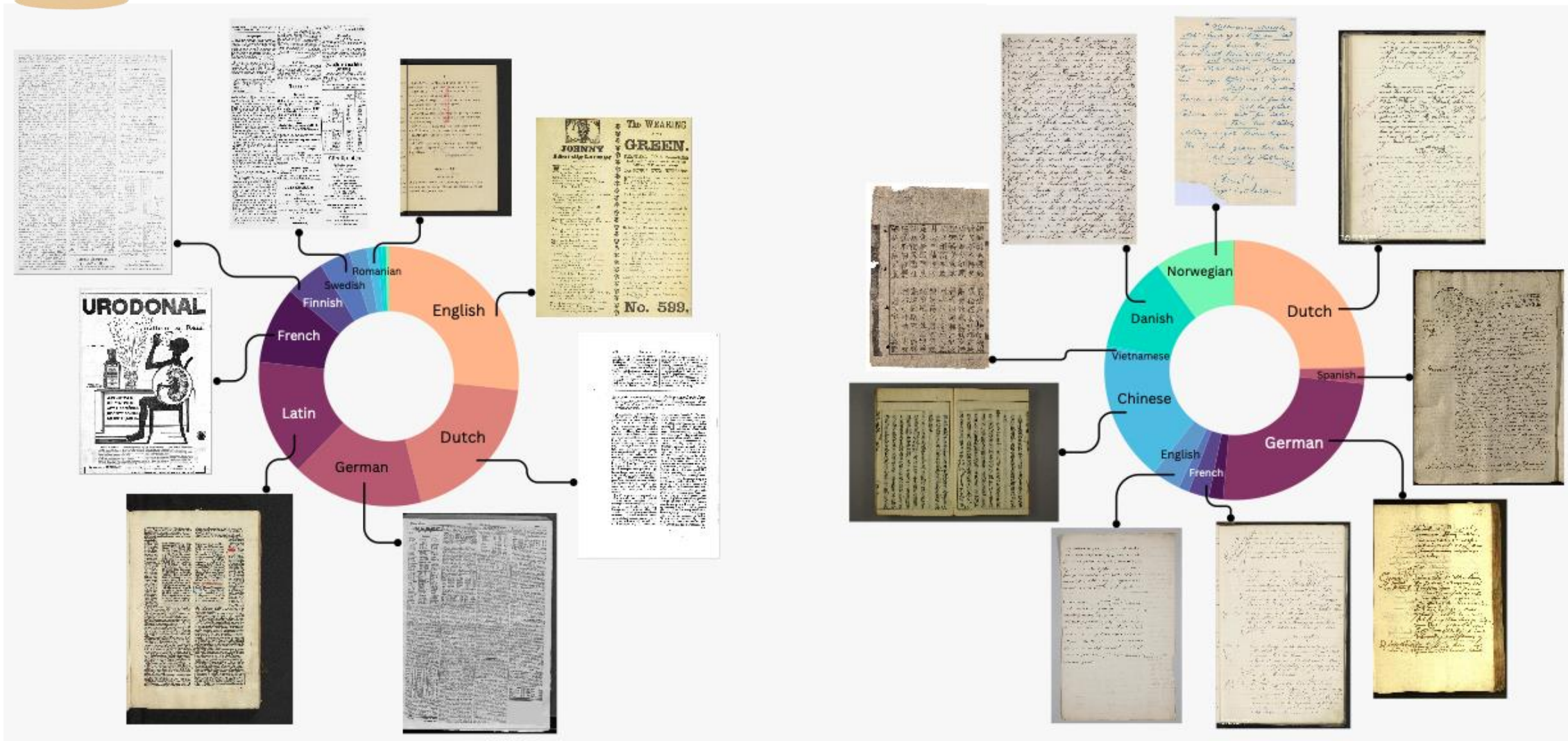
Whole document in HDML
Historical Document Markup Language

2. Churro DS: An Open, Academic Historical Corpus

3. Churro VLM: Train an Open-Weight VLM

1. Design of HDML

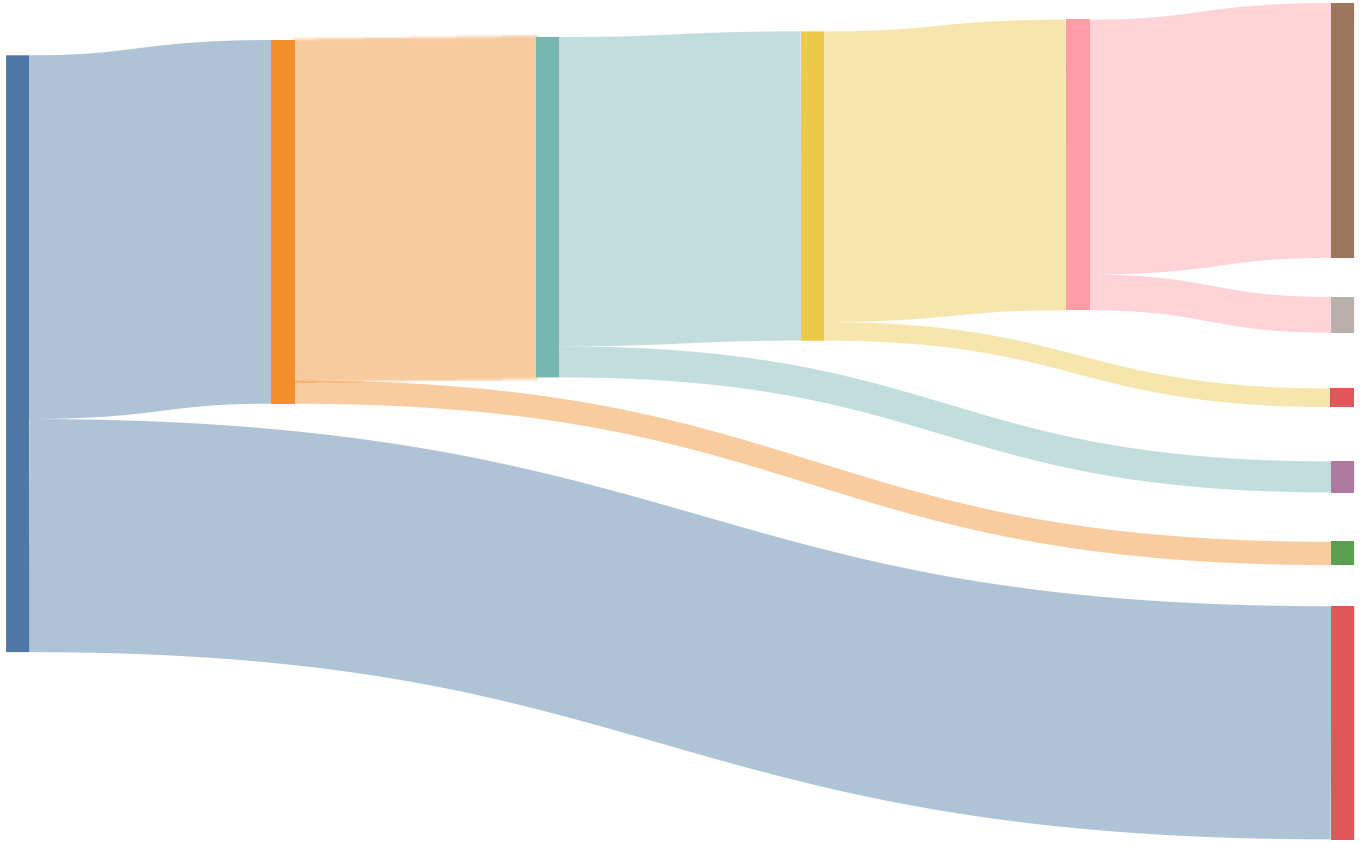
CHURRO Dataset: A Human-Annotated Full-Page Historical Document Dataset



Sources of CHURRO Dataset

Sources

- Zenodo
- GitHub
- Universities
- Libraries
- Books



~2 TB of data
155 corpora
Tens of custom data formats

Restrictive License

Not Convertible to Full-Page

Machine Labels

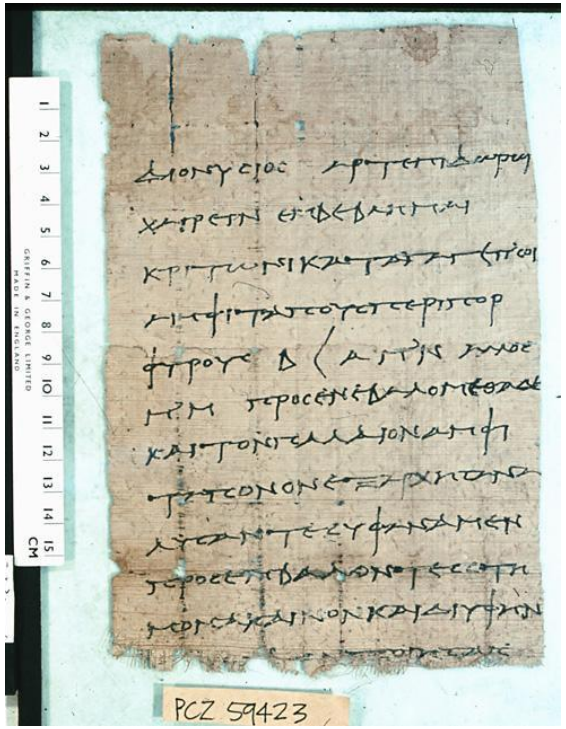
Inaccessible

Not Historical

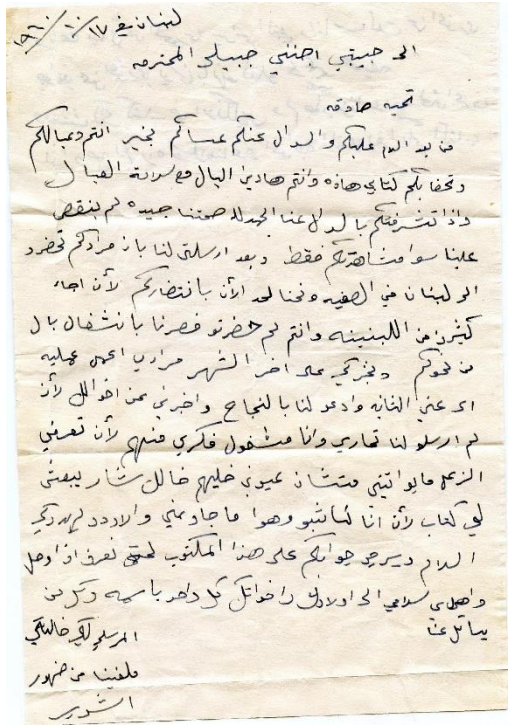
Zenodo: Open repository for EU-funded research outputs from Horizon Europe, Euratom and earlier Framework Programmes.

CHURRO-DS Spans 22 Centuries

Zenon or Zeno (3rd century BC) was a public official in Ptolemaic Egypt around the 250s–230s BC.



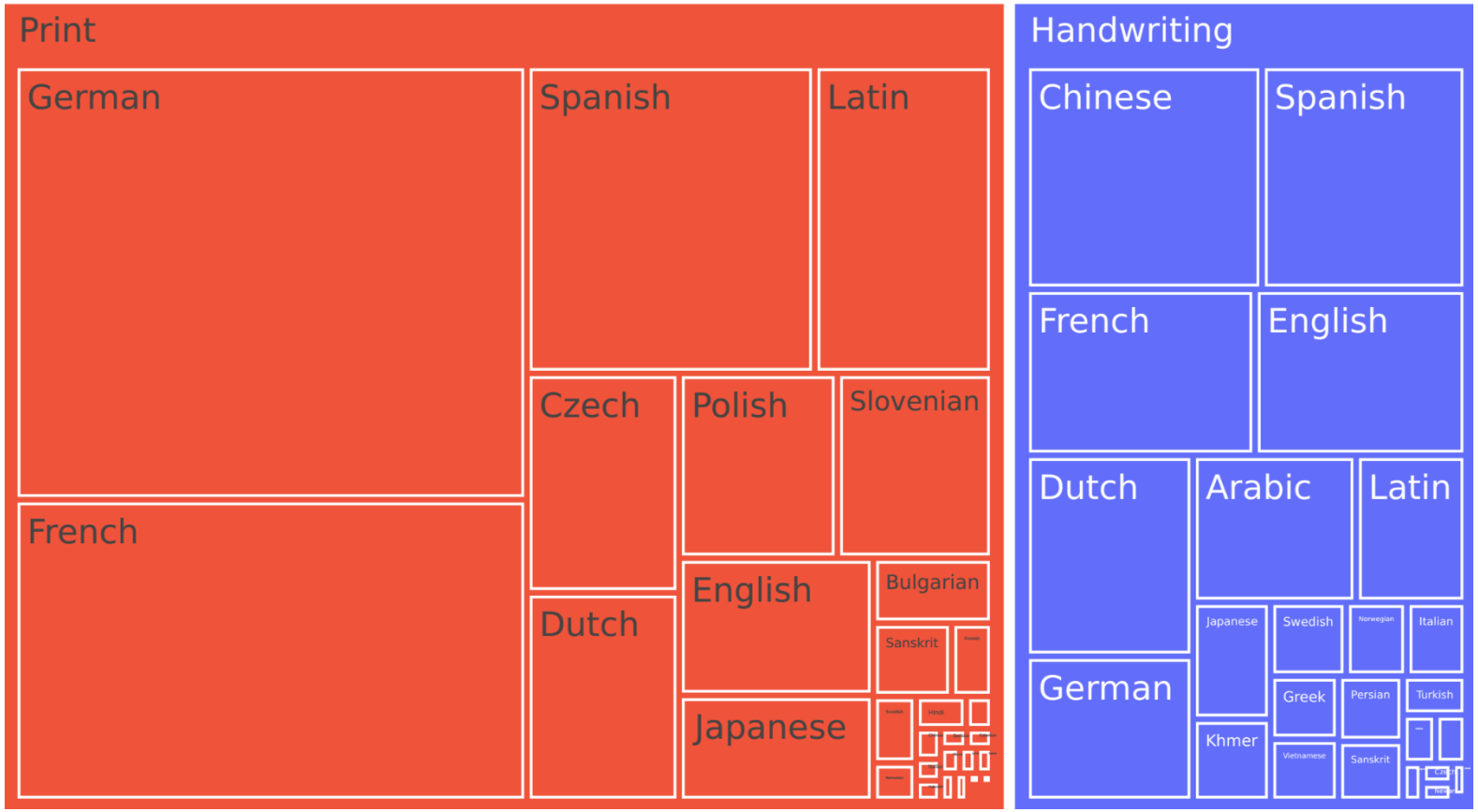
Personal letter written in Arabic. The letter is dated October 17, 1960, from Lebanon. The sender, Malvina from Dhour El Choueir, writes to her niece Ajni Jbeili.



Ground-Truthed Data Set of Zenon Papyri for Handwritten Text Recognition. Isabelle et al. Zenodo. 2022
Muharaf: Manuscripts of Handwritten Arabic Dataset for Cursive Text Recognition. Saeed et al. NeurIPS Datasets. 2024.

CHURRO Contains 46 Language Clusters

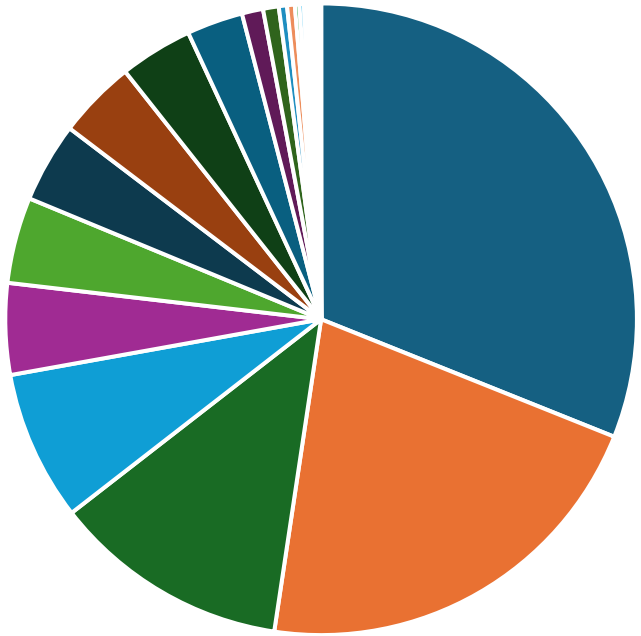
German
Middle High German
Low German
Old High German
Swiss German



ISO 639-3: Language Coding Agency, 2007.

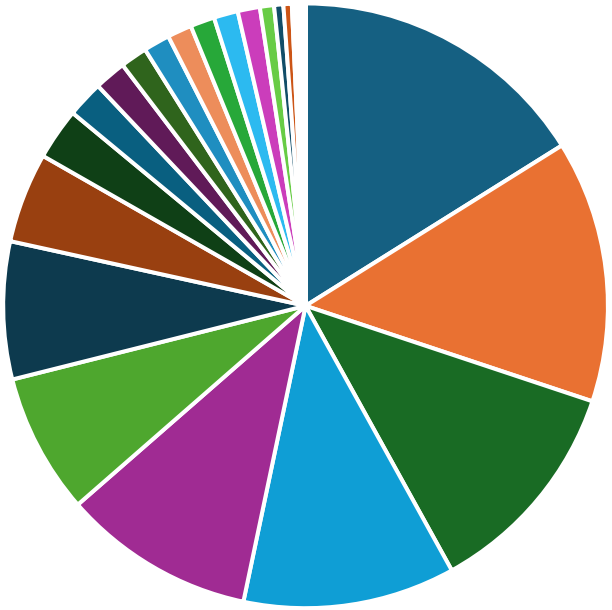
Distribution of Languages

Print (32 Languages)



German French Spanish Latin Czech Dutch
Polish Slovenian English Japanese Bulgarian Sanskrit
Finnish Swedish Romanian Hindi Bangla Chinese
Italian Serbian Malayalam Slovak Estonian Latvian
Sorbian Indonesian Corsican Portuguese Catalan Greek
Russian Slavic Occitan Yiddish Hungarian Malay
Javanese Croatian

Handwriting (28 Languages)



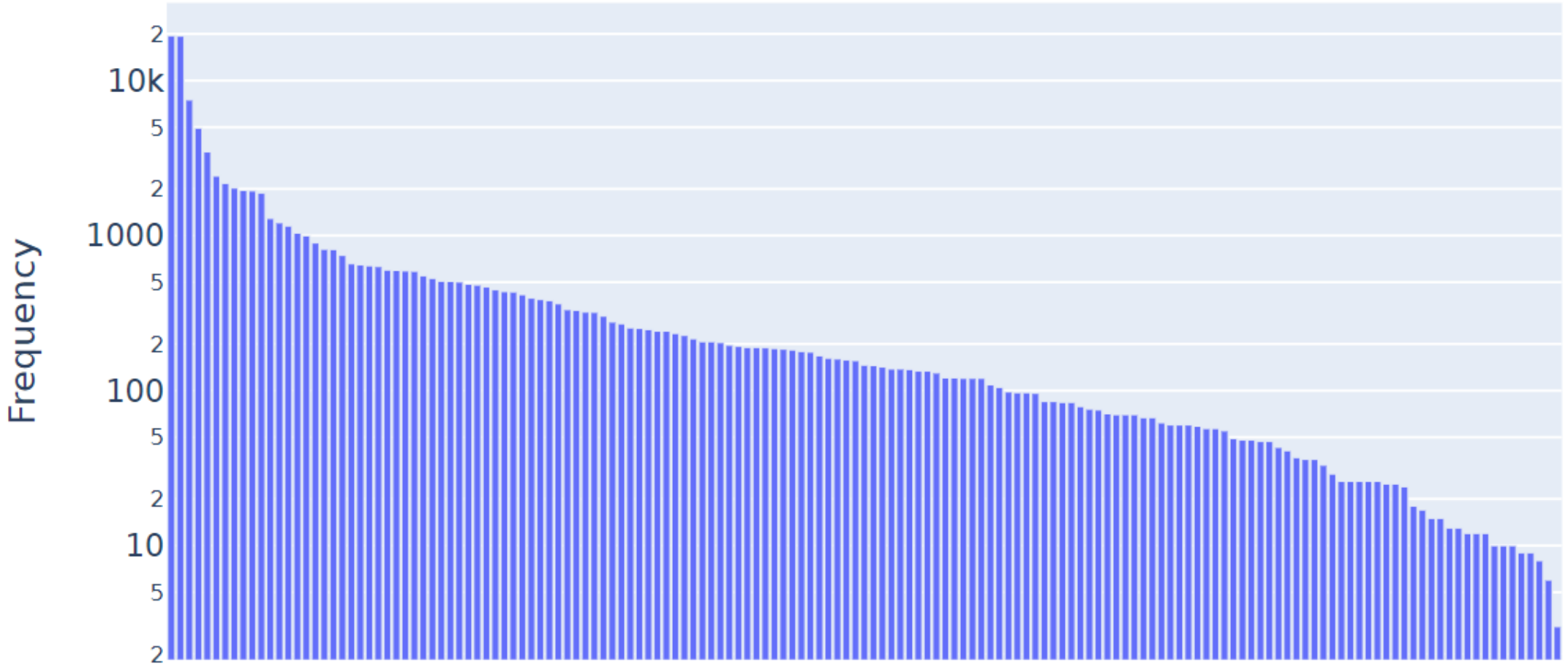
Chinese Spanish French English Dutch German
Arabic Latin Japanese Khmer Swedish Italian
Norwegian Greek Vietnamese Persian Sanskrit Turkish
Hebrew Catalan Portuguese Czech Newari Finnish
Pali Russian Danish Yiddish



CHURRO-DS: Summary

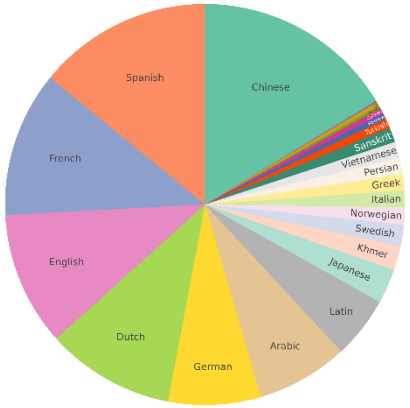
- Unified 150 existing datasets, **100K manuscripts and prints**, 3 writing directions
- **22 centuries**: 3rd century BC to 20th century AD
- Page-level annotation with a **standardized metadata schema in our HDML**
  - **Macrolanguage classification**: ISO 639-3 (Agency, 2007)
  - **Script classification**:
    - ISO 14924 (International Organization for Standardization, 2002)
    - Glottolog
  - **Reading order**: ALTO XML (library of Congress, 2025), PAGE XML
  - **Faithfulness**: Diplomatic transcriptions

Distribution of Corpus Sizes

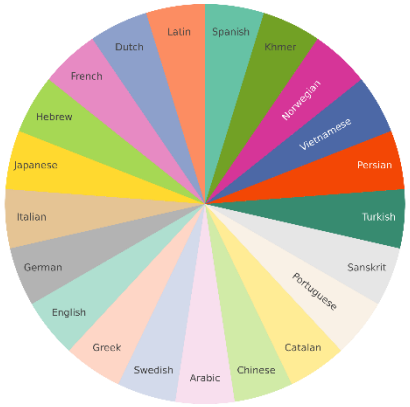


CHURRO Dataset Splits

Handwritten Documents

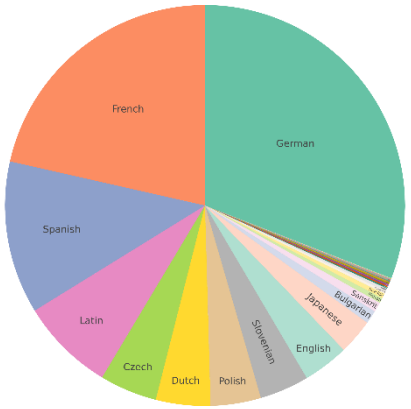


Train: 98,131



Validation: 1,170
Test: 1,170

Printed Documents





Lecture Goal



CHURRO: A Grass-Root Movement to **Make World History Readable**

Original Document

↓

Vision Language Model (VLM)

↓

Whole document in HDML
Historical Document Markup Language

2. Churro DS: An Open, Academic Historical Corpus

3. Churro VLM: Train an Open-Weight VLM

1. Design of HDML



CHURRO VLM: A Vision–Language Model for Reading Historical Documents

Fine-tuned on CHURRO-DS from 3B-parameter Qwen VL 2.5 model

Bai et al., *Qwen2.5-VL Technical Report*, arXiv, 2025.

Experimental Setup

- Benchmark 32 VLMs and OCR systems in a zero-shot setting

- Accuracy Metric: Mormalized Levenstein similarity score

$$1 - (character\ level\ edit\ distance)$$

  - Normalized to lie within $[0, 1]$
  - Evaluated against the gold transcription

Results

System	ben	bul	ces	deu	eng	fin	fra	hin	jpn	lat	nld	pol	ron	san	slv	spa	swe	zho	Avg
Closed VLMs (Zero-Shot)																			
Gemini 2.5 Pro	94.5	93.4	95.6	76.9	93.3	55.3	93.5	93.2	61.7	87.5	97.9	80.3	66.3	89.7	97.4	91.3	81.1	6.6	80.9
Gemini 2.5 Flash	91.2	91.0	94.8	69.3	84.6	32.8	80.9	78.6	47.1	85.0	87.1	73.7	60.2	82.1	97.0	94.4	68.9	8.3	73.7
GPT-4.1 Mini	77.3	90.3	93.8	75.3	88.3	40.1	76.6	82.2	55.3	82.3	82.7	73.1	56.1	71.9	96.3	89.9	79.0	5.3	73.1
Claude Sonnet 3.7	85.2	86.4	94.8	64.7	80.1	25.8	76.1	89.8	29.8	85.0	76.1	71.3	64.1	82.1	95.0	84.8	70.2	1.8	70.2
GPT-5 Mini	78.6	83.8	92.6	67.8	86.4	39.3	83.4	78.2	44.3	81.5	80.5	70.4	43.5	65.5	92.9	90.6	72.7	4.6	69.8
Claude Opus 4.1	82.3	82.9	91.5	58.7	78.3	16.8	77.7	87.5	17.3	84.1	70.1	65.3	57.7	79.5	96.3	85.7	65.2	1.8	66.6
O4 Mini	75.8	81.9	89.2	63.2	82.3	21.2	80.5	73.3	38.8	79.5	74.2	64.0	39.4	59.6	95.1	89.0	69.6	2.4	65.5
GPT-4.1	71.4	78.3	89.5	63.7	81.8	34.5	75.0	76.4	19.7	76.9	76.6	64.8	40.9	60.6	90.6	88.0	66.5	3.9	64.4
O3	76.5	77.8	88.0	58.4	81.6	19.7	80.1	76.3	12.0	79.6	73.1	65.5	33.6	59.8	94.8	89.4	60.0	1.5	62.7
O1	74.7	76.0	87.5	58.9	82.1	24.3	72.2	78.1	15.6	77.5	74.0	64.2	35.3	60.2	90.4	89.0	62.6	2.0	62.5
Claude Sonnet 4	73.3	77.6	86.6	56.1	76.2	15.8	69.9	76.5	17.5	83.1	70.7	63.4	51.0	69.1	93.8	78.9	58.8	2.0	62.2
GPT-5	73.6	77.1	86.9	59.2	81.1	19.9	79.8	61.5	14.2	78.2	71.8	64.2	33.1	53.0	90.3	89.6	58.6	1.3	60.7
GPT-4o	66.7	63.0	82.9	48.8	78.3	15.6	70.2	70.6	11.9	74.2	62.8	54.8	40.1	48.1	89.5	86.4	47.1	2.1	56.3
GPT-4o Mini	50.9	58.7	76.9	48.9	77.1	17.0	66.6	57.8	11.9	69.4	56.9	52.1	43.8	42.1	85.2	85.4	47.9	2.4	52.8
GPT-4.1 Nano	36.2	61.0	82.5	54.4	81.3	20.4	67.8	34.1	14.0	71.9	58.2	58.4	41.4	26.1	86.3	82.3	53.6	3.6	51.9
GPT-5 Nano	28.2	27.9	68.2	41.1	73.0	5.3	65.3	28.5	6.1	56.9	32.2	38.7	27.3	13.0	78.0	69.5	37.2	0.5	38.7

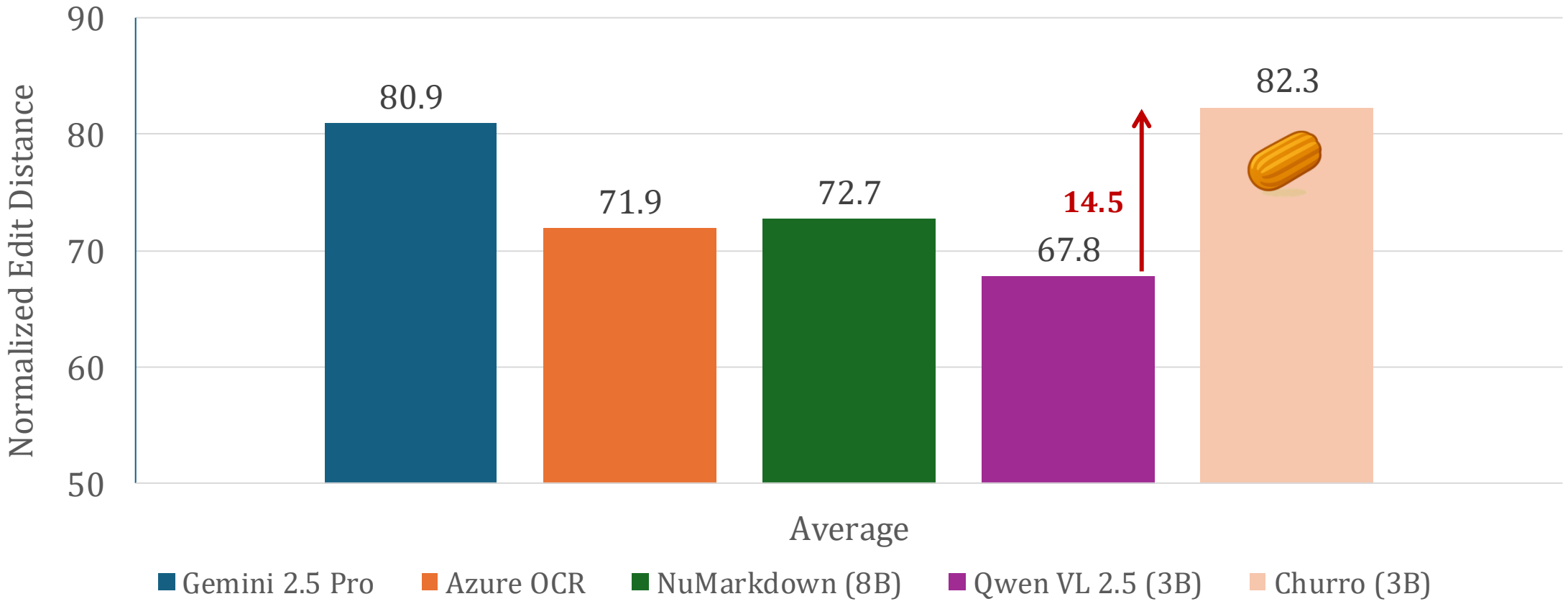
From left to right: Bangla, Bulgarian, Czech, German, English, Finnish, French, Hindi, Japanese, Latin, Dutch, Polish, Romanian, Sanskrit, Slovenian, Spanish, Swedish, Chinese

Results

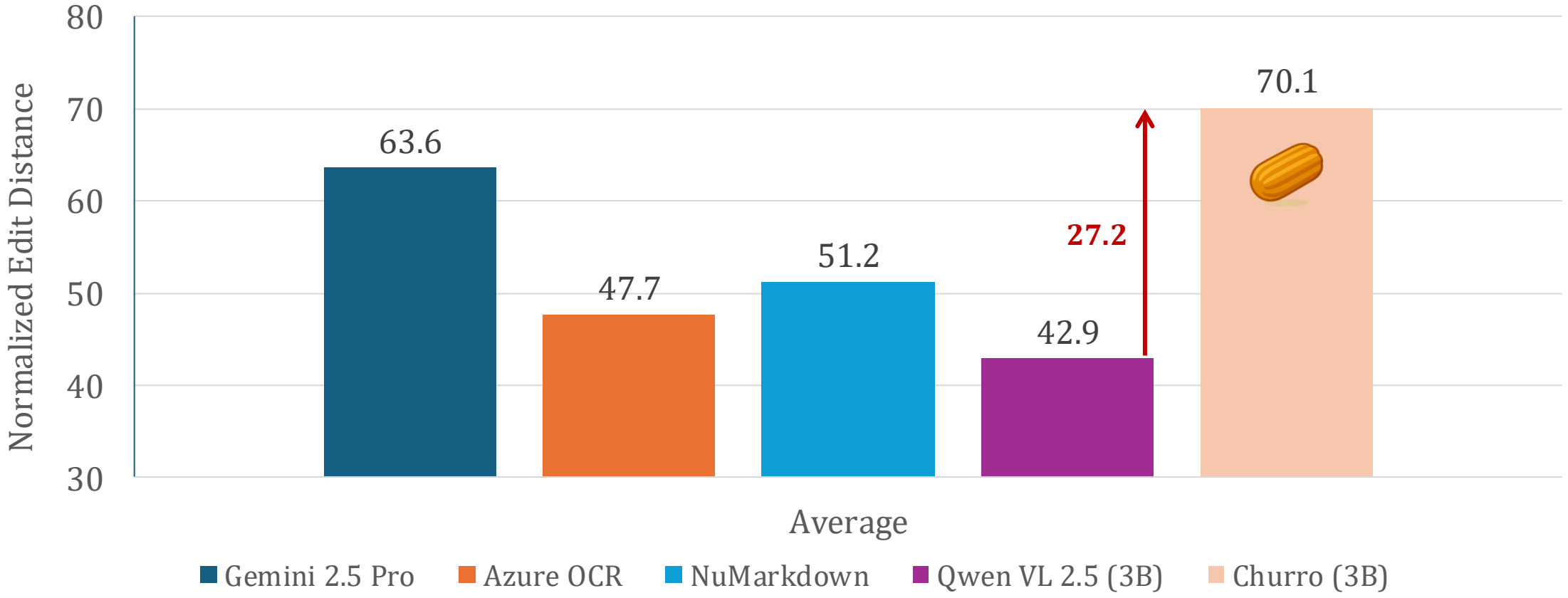
System	ben	bul	ces	deu	eng	fin	fra	hin	jpn	lat	nld	pol	ron	san	slv	spa	swe	zho	Avg
Open-Weight VLMs (Zero-Shot)																			
NuMarkdown (8B)	84.7	87.1	95.7	71.9	83.0	39.0	87.1	75.5	55.2	80.6	84.7	77.4	52.9	73.3	95.6	86.9	77.3	1.1	72.7
olmOCR (8B)	79.2	84.8	90.0	66.4	79.7	28.5	83.7	85.0	42.5	82.5	82.9	72.4	51.4	66.4	96.2	90.7	73.2	0.2	69.8
Nanonets OCR (3B)	75.6	85.5	94.7	69.4	85.7	24.8	82.9	83.1	32.9	79.7	78.1	74.2	57.3	71.4	95.7	91.2	72.2	0.4	69.7
Qwen 2.5 VL (3B)	70.0	83.3	93.6	64.6	82.0	30.3	79.9	77.9	37.1	80.4	75.7	67.7	59.1	72.6	94.8	87.2	64.3	0.5	67.8
RolmOCR (8B)	82.7	79.4	87.5	63.9	88.4	17.0	78.4	85.1	35.1	79.2	73.7	65.5	57.9	67.7	86.1	90.7	70.8	1.2	67.2
Qwen 2.5 VL (72B)	87.9	89.6	83.6	53.3	75.7	10.6	80.2	89.0	47.3	85.3	65.9	57.1	57.4	70.2	95.9	89.5	50.9	3.7	66.3
Gemma 3 (27B)	61.6	62.0	78.3	49.7	70.4	21.0	66.8	74.3	18.3	69.1	50.4	52.7	48.7	61.4	89.4	72.5	44.0	5.4	55.3
MiMo VL (8B)	63.8	72.7	82.4	51.0	71.5	7.2	72.6	51.4	26.8	82.6	50.4	49.6	40.0	44.7	92.4	75.7	50.5	2.1	54.8
Skywork R1V3 (38B)	20.7	35.8	70.6	42.3	73.4	17.6	68.7	18.8	12.2	67.3	47.2	46.8	31.5	17.1	81.6	72.3	37.0	2.7	42.4
InternVL 3.5 (30B)	16.4	10.2	56.5	35.3	76.3	13.1	64.6	16.0	13.2	62.8	42.7	31.5	18.8	11.6	77.3	67.5	27.8	2.8	35.8
R (4B)	19.0	17.2	43.9	28.4	67.7	13.9	62.0	25.2	6.3	48.8	32.0	27.3	24.6	16.9	62.5	62.8	27.0	2.9	32.7
Nemotron Nano VL (8B)	7.7	14.2	35.8	34.3	57.1	6.6	60.0	11.7	0.4	55.9	32.7	33.1	14.8	9.4	59.0	63.4	30.9	1.1	29.3
Phi 4 Multimodal (5B)	2.2	1.7	4.0	5.3	42.9	5.2	11.6	3.7	1.8	11.4	7.0	6.3	5.6	3.5	25.6	16.7	7.4	0.7	9.0
OCR Systems																			
Azure OCR	15.3	91.4	91.7	77.9	86.1	65.0	84.5	72.9	53.6	82.6	91.6	81.1	50.7	80.8	93.9	85.5	79.6	9.4	71.9
Mistral OCR	59.4	88.1	86.5	57.1	79.3	28.6	74.7	82.2	3.3	79.8	78.9	60.9	48.6	82.7	93.8	89.3	58.1	1.3	64.0
VLM Fine-Tuned on CHURRO-DS																			
CHURRO (3B)	70.9	96.1	95.6	82.3	91.0	63.2	89.6	94.6	74.1	92.0	95.7	81.6	79.6	93.1	97.6	91.4	87.2	6.2	82.3
Δ with Qwen 2.5 VL (3B)	+0.9	+12.8	+2.0	+17.7	+9.1	+32.9	+9.7	+16.7	+37.0	+11.6	+20.0	+14.0	+20.5	+20.6	+2.8	+4.2	+22.9	+5.7	+14.5

From left to right: Bangla, Bulgarian, Czech, German, English, Finnish, French, Hindi, Japanese, Latin, Dutch, Polish, Romanian, Sanskrit, Slovenian, Spanish, Swedish, Chinese

Summary of Results for Printed Documents



Summary of Results for Handwritten Documents

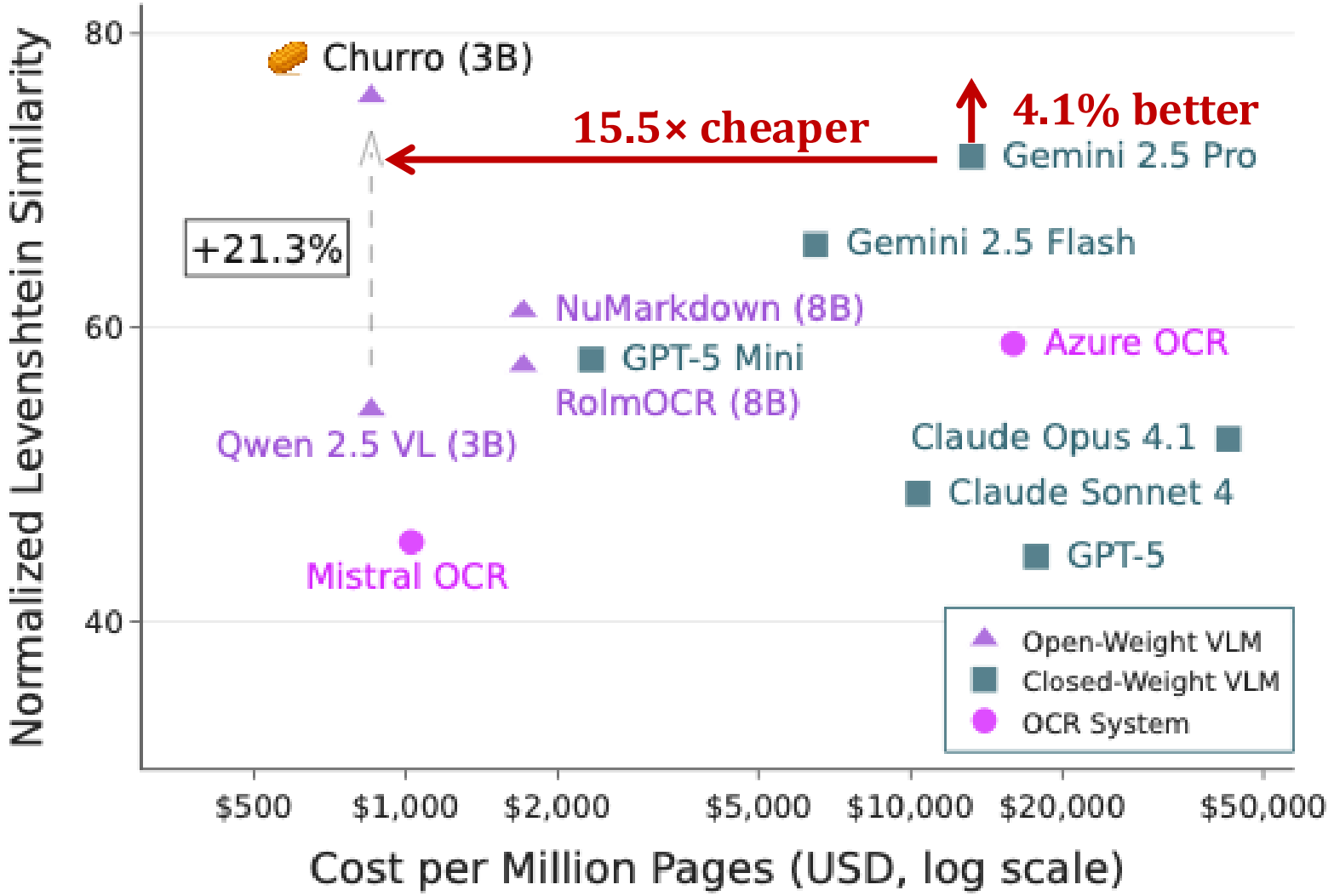


Summary of Per-Language Results

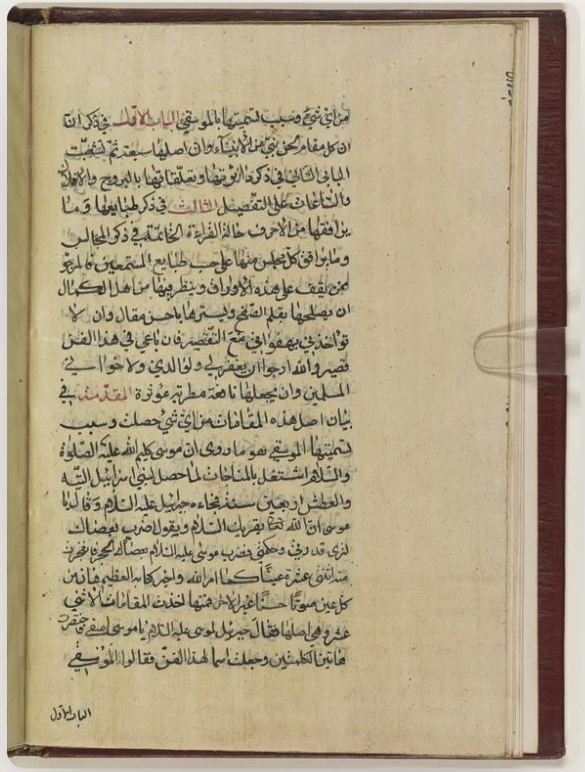
- Handwriting accuracy is 80%+ in:
  - Catalan, Dutch, English, French, German, Italian, Swedish

- Print accuracy is 90%+ in:
  - Bulgarian, Czech, Dutch, English, Hindi, Latin, Sanskrit, Slovenian, Spanish

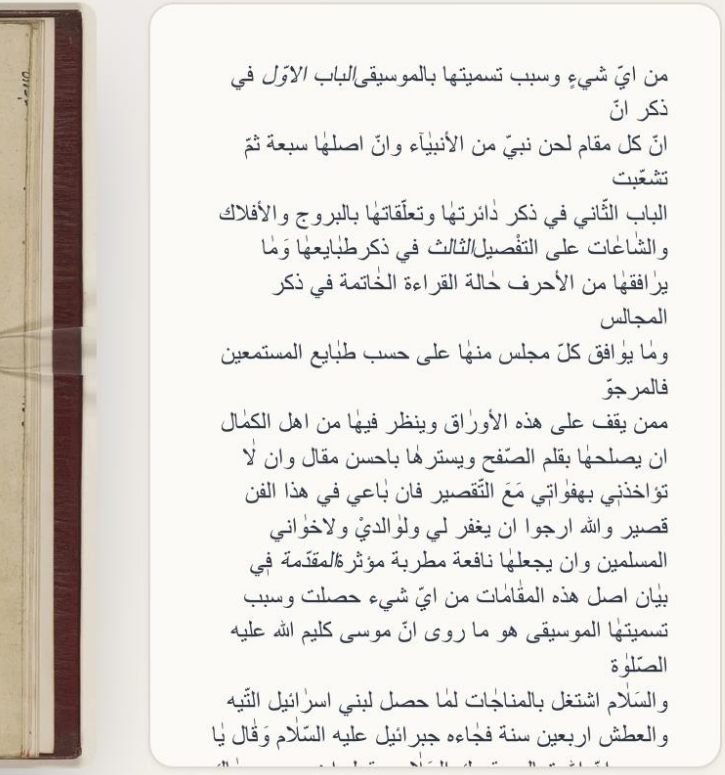
The Accuracy-Cost Trade-off



WHAT SHOULD WE DO NEXT?



من ايّ شيءٍ وسبب تسميتها بالموسيقى*الباب الاوّل* في ذكر انّ
انّ كل مقام لحن نبيّ من الأنبيآء وانّ اصلها سبعة ثمّ تشعّبت
الباب الثّاني في ذكر دائرتها وتعلّقاتها بالبروج والأفلاك
والشاعات على التفصيل*الثالث* في ذكرطبايعها وَما
يرافقها من الأحرف حالة القراءة الخاتمة في ذكر المجالس
وما يوافق كلّ مجلس منها على حسب طبايع المستمعين
فالمرجوّ
ممن يقف على هذه الأوراق وينظر فيها من اهل الكمال
ان يصلحها بقلم الصّفح ويسترها باحسن مقال وان لا
تؤاخذني بهفواتي مَعَ التّقصير فان باعي في هذا الفن
قصير والله ارجوا ان يغفر لي ولوالديّ ولاخواني
المسلمين وان يجعلها نافعة مطربة مؤثرة*المقدّمة* في
بيان اصل هذه المقامات من ايّ شيء حصلت وسبب
تسميتها الموسيقى هو ما روى انّ موسى كليم الله عليه
الصّلوٰة
والسّلام اشتغل بالمناجات لمّا حصل لبني اسرائيل التّيه
والعطش اربعين سنة فجاءه جبرائيل عليه السّلام وَقال يا



History Genie

A growing library of 317+ global primary documents.

Use AI tools to transcribe, compare documents, and ask detailed questions.

Start Exploring

Sign Up

Our state-of-the-art transcription model covers **46 languages and scripts** across **22 centuries**

CREATED BY

Open Virtual Assistant Lab

Stanford University

Learn more about the AI research and the team →

Three College Courses Have Used History Genie to Teach History with Authentic Sources



Trevor Getz

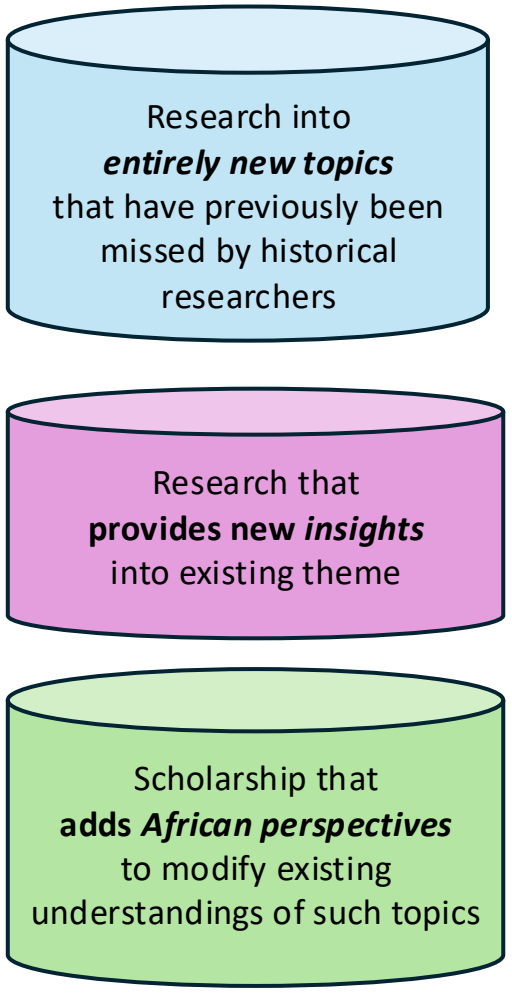
Professor of
African and World History at
San Francisco State University
President
World History Association

Stanford University

- **AFRICAAM 240** Knowledge Outside the Colonial Library

San Francisco State University

- **HISTORY 366** and **Anthropology 609** Approaches to the African Pasts
- **HISTORY 305** AI and Historical Inquiry



Example: The execution of a Sierra Leonean leader named Whobay and subsequent protests.

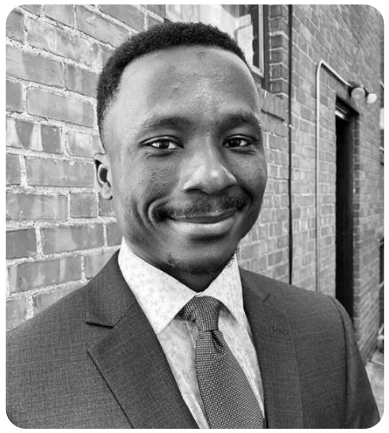
Example: the impact of the industrialization of agriculture in Liberia or attempts to set up independent states on the coast of Ghana.

Example: Nineteenth century Zionism, anti-slavery movements, and the abolition of alcohol.

Historians Have Used History Genie to Discover New Evidence

My aim was to center local sources, but it **was difficult to achieve** that until you shared History Genie with me.

History Genie **allowed me to complement dairies and narratives** by British military men with those authored in the African Times by West Africans who lived along the Atlantic Coast.



Tony Yeboah
Assistant Professor of African Art History
Tulane University



Benjamin Wendorf
Professor of History
Quinsigamond Community College

The History Genie tool was immensely helpful in what otherwise would have been a **lengthy process** of reading entirely through each edition of these newspapers to find relevant material.



A Grass-Root Movement to Make Historical Documents Accessible

- **CHURRO-DS dataset with a new HDML language**
  - 99,491 pages
  - 22 centuries
  - 46 language clusters
  - Handwriting and print
- **CHURRO-VLM: a compact open-weight vision-language model**
  - Comparable to state-of-the-art commercial system (Gemini 2.5 Pro)
  - 15.5x cheaper
- **HistoryGenie: AI assistant for historical documents**

Semnani et al., *CHURRO: Making History Readable with an Open-Weight Large Vision-Language Model for High-Accuracy, Low-Cost Historical Text Recognition*, EMNLP, 2025.