

When AI Meets the Archive: Transforming the *Islam West Africa Collection* with Large Language Models

Frédéric MADORE

Leibniz-Zentrum Moderner Orient

AI and the Changing Landscape of Scholarship

- Generative AI sparks both enthusiasm and anxiety in academia
- LLMs promise efficiency but threaten scholarly foundations
- History faces disruption in methods and professional identity
- Human interpretation remains essential despite automation

Is this the Last Generation of Historians?

OpenAI's Deep Research is a strange and shockingly powerful agentic AI research assistant that offers a clear glimpse of the future. But where do we fit in?



MARK HUMPHRIES
FEB 07, 2025

HIGHER EDUCATION

Everyone Is Cheating Their Way Through College

ChatGPT has unraveled the entire academic project.



By James D. Walsh, *Intelligencer* features writer

MAY 7, 2025

SAVE

ANNALS OF EDUCATION

WHAT HAPPENS AFTER A.I. DESTROYS COLLEGE WRITING?

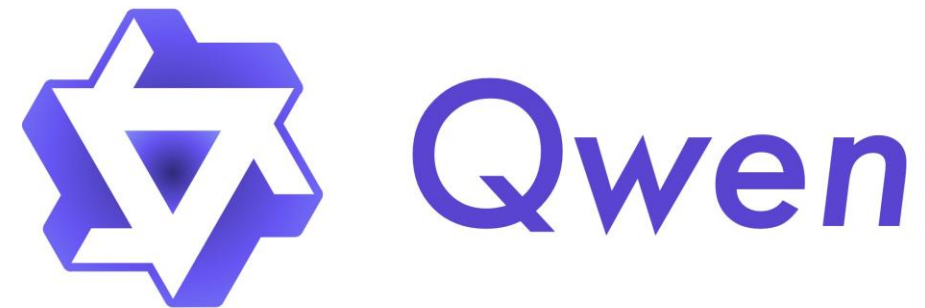
The demise of the English paper will end a long intellectual tradition, but it's also an opportunity to reexamine the purpose of higher education.

By Hua Hsu

June 30, 2025

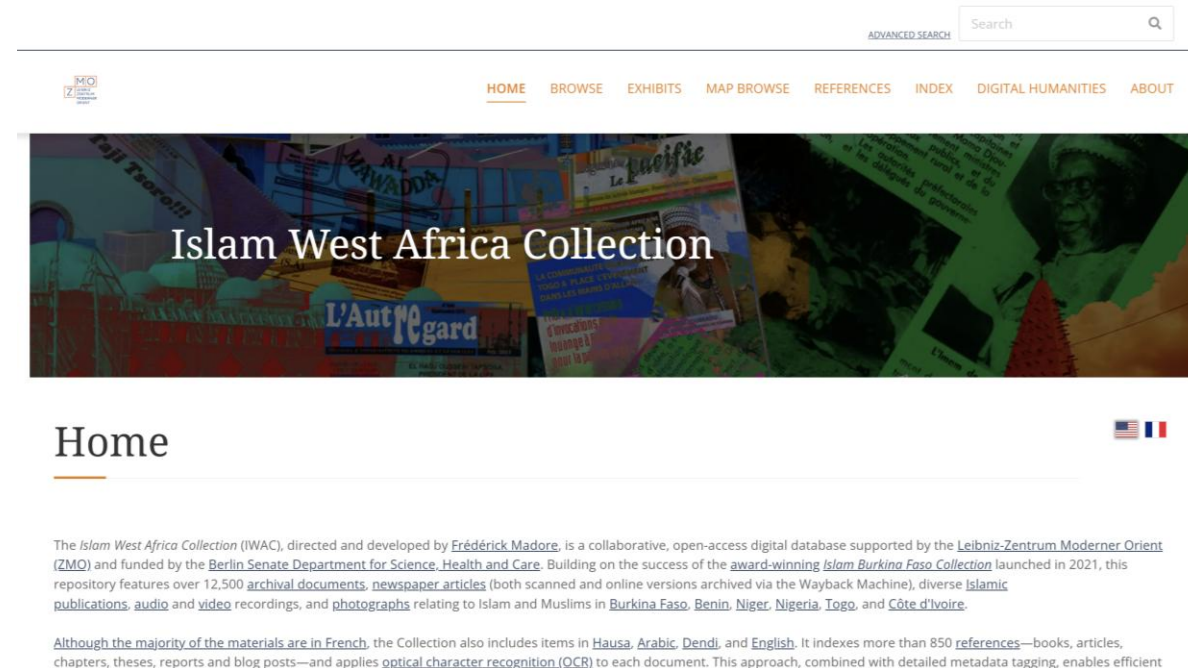
Archives, AI, and the Digital Humanities in Africa

- Archives at critical juncture: efficiency vs. ethical dilemmas
- LLMs expand qualitative research but introduce bias, opacity
- Open-source AI tools broaden global scholarly participation
- Chinese models (DeepSeek, Kimi, Qwen) boost accessibility and use
- Challenges: equitable access, local languages, cultural contexts



Islam West Africa Collection (IWAC) and AI-Driven Solutions

- Open-access archive: 14,000+ documents across six countries
- A scale problem
- Three applications explored:
 1. AI-enhanced document processing (OCR, NER)
 2. AI-powered sentiment analysis of newspapers
 3. LLM chatbot interface with retrieval-augmented generation (RAG)



<https://islam.zmo.de/s/westafrica/>

AI Solutions vs. Sustainability Concerns

- AI offers pragmatic benefits for African digital archives
- Concerns: opacity, bias, computational costs, environmental impact
- “Minimal computing” and “digital sobriety” frameworks matter
- Lightweight open-source models make tools increasingly viable



Origins, Significance & Addressing Gaps

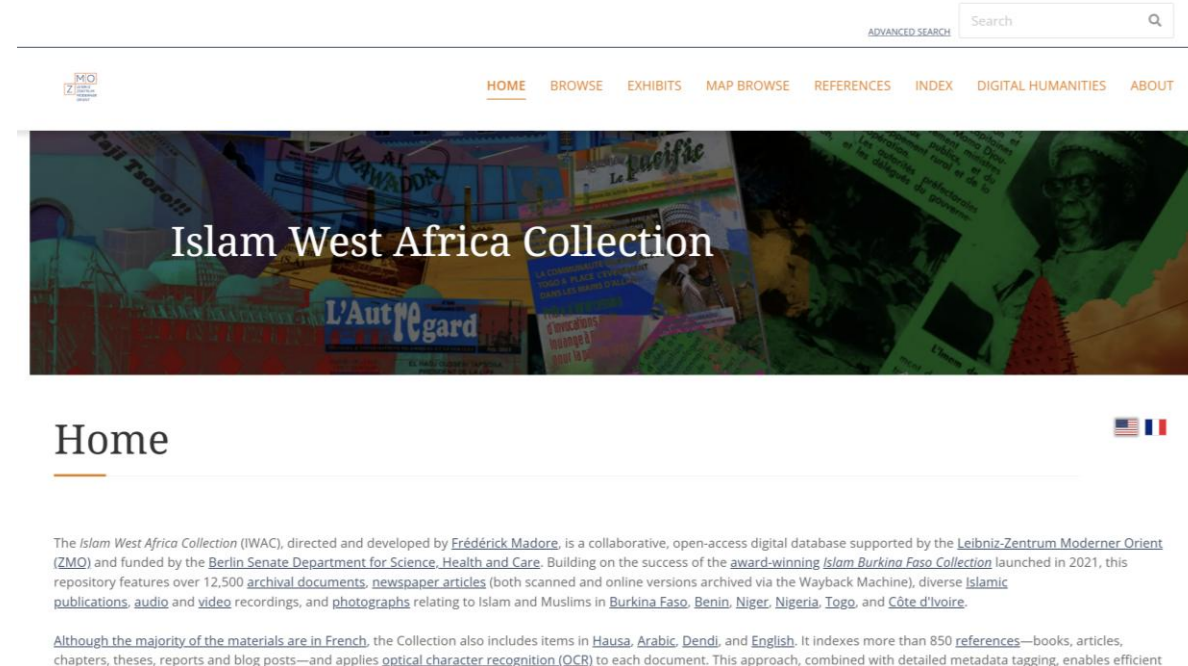
- Grew from extensive fieldwork & archival research since 2013
- Addressing Scarcity: Digitizing rare/ephemeral materials (newspapers 1962-2024, unique Islamic publications)
- Countering Manuscript Bias: Focus on diverse sources beyond traditional manuscript projects (print culture, born-digital)
- Bridging Divides: Makes scattered sources (private collections, 9+ institutions) centrally accessible online
- Open Science Influence: Commitment to wide dissemination and enabling new research



Archives of the newspaper *La Nation* (Benin)

From the *Islam Burkina Faso Collection* to the IWAC

- 14,000 items (archival docs, newspaper articles, Islamic publications, video, photos)
- 850 scholarly references indexed
- Burkina Faso, Benin, Niger, Nigeria, Togo, Côte d'Ivoire
- Primarily French; also includes Hausa, Arabic, Dendi, English
- 4,300 index entries (people, places, topics)
- Publisher agreements and fair use



<https://islam.zmo.de/s/westafrica/>

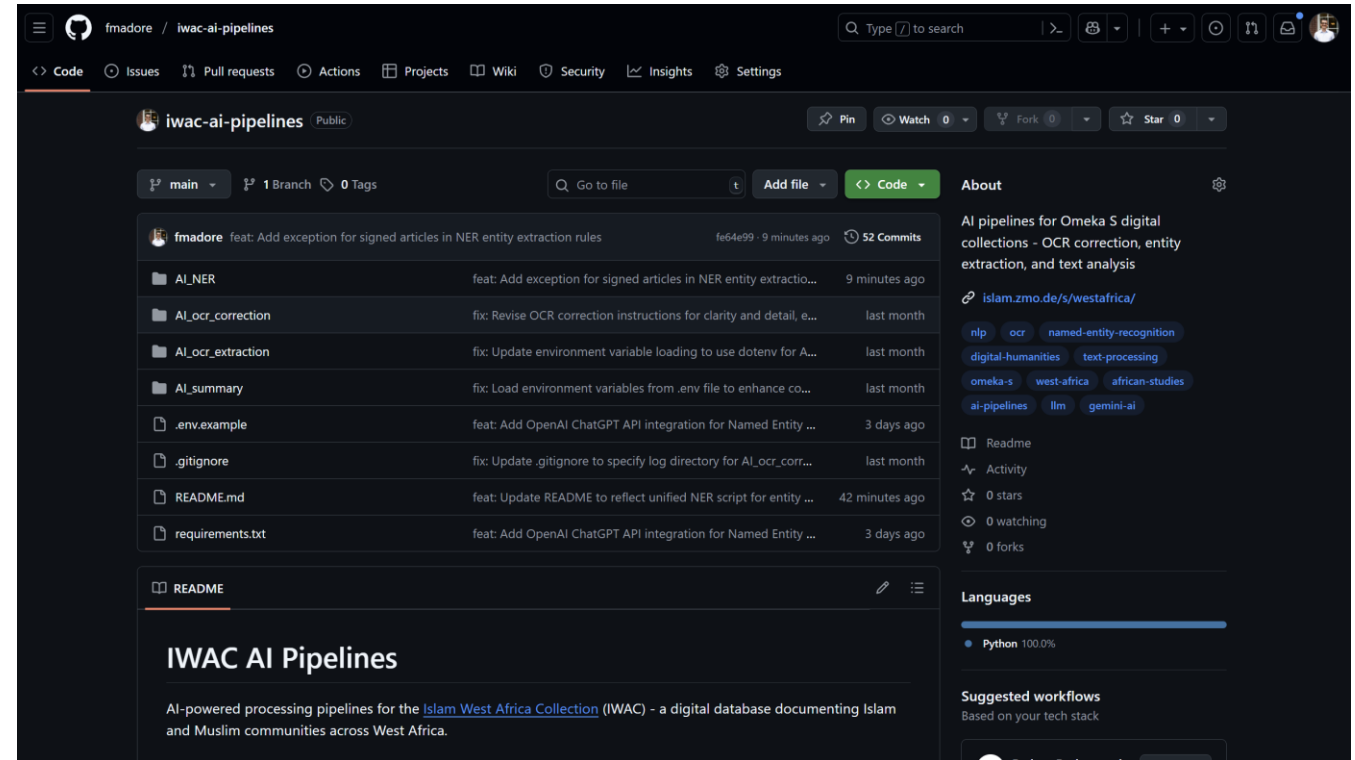
The Scale Problem

- Scale of the database: over 25 million words
- 62,000 pages of single-spaced text in Times New Roman
- **Future growth**
- Thousands of additional unprocessed documents
- Manual OCR correction and metadata unscalable



AI-powered processing pipelines

- Algorithms essential to manage massive corpus
- Traditional analysis cannot scale effectively
- AI workflows sustain growth and usability
- Automates OCR checks, entity tagging, summarisation



<https://github.com/fmadore/iwac-ai-pipelines>

Optical Character Recognition (OCR) and Document Layout Analysis

- Newspapers ≠ books → multi-columns, ads, images, legends → reading order chaos
- One-person manual fixes no longer viable
- Multimodal AI “sees” page layout and reads text → clean segmentation + better OCR
- Two paths
- Post-correct legacy OCR with Python + LLM prompts
- Direct image-to-text with latest vision-language models



Google API Cloud Vision

AI-Powered OCR for Newspapers

- LLMs use context to correct degraded text
- Gemini Flash 2.5 pipeline fixes legacy OCR
- Prompted for structure, diacritics, names, uncertainty flags

- **Accent and spelling corrections**

- Original: [...] après mes **études** au Nigeria. Et depuis **lors**, Cheikh [...]
- Corrected: [...] après mes **études** au Nigeria. Et depuis **lors**, Cheikh [...]

- **Character confusion correction**

- Original: Chouaïb **considerait** tout le monde comme faisant partie de sa famille. **Il** nous [...]
- Corrected: Chouaïb **considérait** tout le monde comme faisant partie de sa famille. **Il** nous [...]

- **Spacing and Hyphenation Corrections**

- Original: la foi, **l'a:mour** du travail [...]
- Corrected: la foi, **l'amour** du travail [...]

- **Other common mistakes**

- rn vs. m: Combination “**rn**” misinterpreted as “**m**”
- O vs. 0: Confusing the letter “**O**” with the number “**0**”
- Celte vs. Cette: Misinterpreting similar-looking letters such as “**l**” and “**t**”



https://github.com/fmadore/iwac-ai-pipelines/blob/main/AI_ocr_correction/ocr_correction_prompt.md

Beyond Correction: Multimodal AI “Sees” and “Reads”

- Analyzes scanned images directly, not just OCR text
- Multimodal models process images, text, audio, video, tables
- Combines visual layout cues with linguistic context
- Enables accurate segmentation of titles, captions, ads, tables
- Requires latest—and more costly—AI models
- Gemini 2.5 Pro

Organisation du hadj

Les imams se battent

L'organisation du Hadj est en train de devenir le grand diviseur commun des musulmans en Côte d'Ivoire. Les uns (musulmans), sont accusés par les autres (musulmans), de voleurs, d'imposteurs, de truands à travers prônes, prêches et sermons. Le triste spectacle auquel se sont livrés les imams, lundi dernier à l'hôtel du district d'Abidjan lors de la réunion bilan du hadj, a mis au grand jour la division de la communauté.

Les guides religieux musulmans s'entre déchirent autour de l'organisation du hadj. C'est la lutte sans merci au niveau de chaque structure pour le contrôle de l'organisation du pèlerinage. Le triste spectacle auquel se sont livrés les imams, lundi dernier à l'hôtel du district d'Abidjan lors de la réunion bilan du hadj 2008 a mis au grand jour la division de la communauté. Les imams ont failli en venir aux mains. Les structures proches du conseil national islamique (Cni) ne veulent plus accepter que le Conseil supérieur des imams (Cosim) soit co-acteur ou co-organisateur du pèlerinage. « Le Cosim n'est pas l'émanation de tous les imams. Nous devons mettre en place une nouvelle structure. Laquelle sera coorganisatrice du hadj avec l'Etat, le porte parole de cette nouvelle structure sera le porte parole de la communauté musulmane ivoirienne », a décidé à l'occasion le président du Cni, Idriss Koudouss.

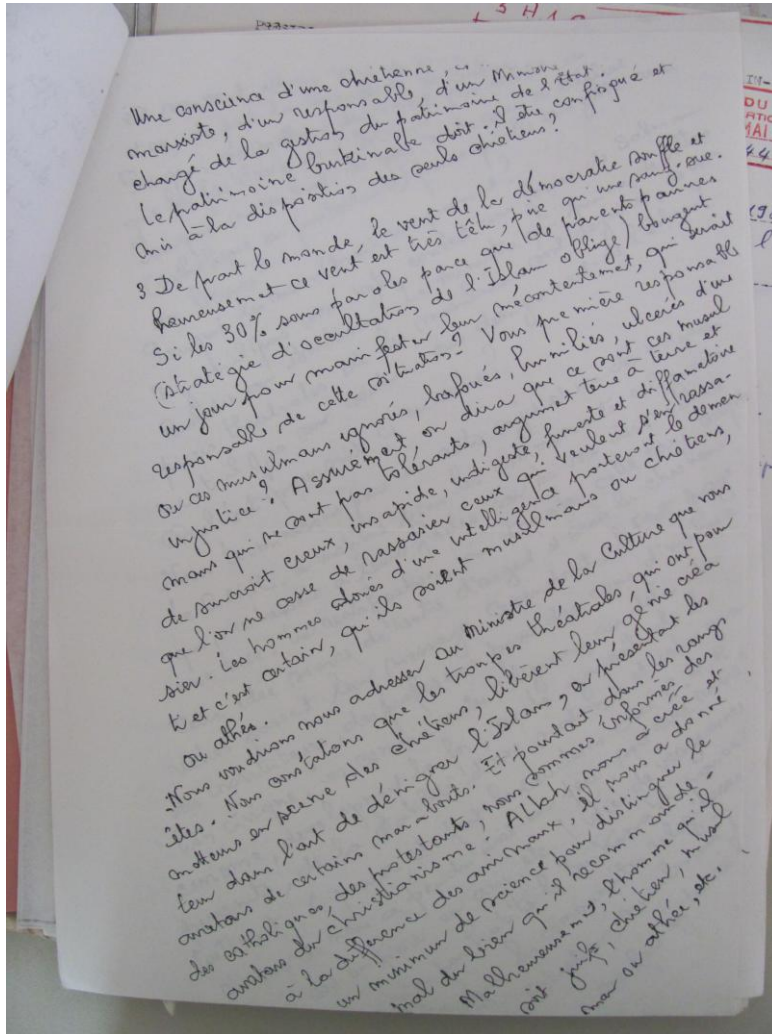
Des propos qui ont mis en exergue la profonde dissension au sein de la communauté. Il a fallu de peu que ceux qui sont censés éclairer la communauté n'en viennent aux mains. La rencontre s'est transformée en une foire aux injures et empoignades à l'encontre du président du Cosim, Cheick abboubakar Fofana. Une semaine, avant, El Hadj Mamadou Dosso, Imam en charges de Recherches et de documentation au Centre d'Education et de Recherches Islamiques-C.E.D.R.I.S avait tiré la sonnette d'alarme. Selon lui, la communauté musulmane dans les mosquées connaît à présent ses maux: « Il y existe des oppositions entre Imams, entre Imams et membres de structures de gestion, entre musulmans arabisants et francisants, entre musulmans ivoiriens et non ivoiriens, toutes, des oppositions créées et entretenues par ceux et celles qui voient en ce haut lieu de cultes, une occasion d'hégémonie associative et de promotion individuelle, et une opportunité d'affaires lucratives », avait-il prévenu.

Pour lui, l'organisation du Hadj reste depuis 1994 l'année de sa cession à la communauté musulmane, le plus grand diviseur commun des musulmans en Côte d'Ivoire. « Les uns (musulmans), sont accusés par les autres (musulmans), de voleurs, d'imposteurs, de truands à travers prônes, prêches et sermons relayés par la presse. « Si vous croyez en Allah et au jugement dernier disait le Prophète Muhammad (saw), tenez des propos décents ou taisez-vous », déplore l'homme de Dieu. Qui s'interroge: « On peut se poser des questions: organise-t-on le hadj au nom d'Allah et pour les musulmans? L'organise-t-on pour avoir de l'argent et/ou de l'ascendance sur ses coreligionnaires? Pendant qu'on s'entredéchire autour de l'organisation du Hadj, une partie des musulmans du pays refusent de célébrer les fêtes de Tabaski et de fin de Ramadan aux dates fixées par des structures musulmanes, au motif que celles-ci seraient inféodées aux pouvoirs publics ». Le hadj est devenu, tout bien calculé, un fonds de commerce pour nombre d'hommes de Dieu. Les imams, après l'échec du hadj 2006 bis ont pris la relève des démarcheurs dans l'organisation du pèlerinage aux côtés de l'Etat. Mais les choses vont toujours de mal en pis car des imams semblent vouloir se remplir eux aussi les poches. « Aujourd'hui, personne ne sait la structuration du prix du hadj que les organisateurs (Imams et Etat) tiennent pour secret. Pour mieux se sucrer sur le dos des pèlerins. Cette guéguerre n'est rien d'autre que la course à l'argent que le pèlerinage génère », accuse un responsable d'association de démarcheurs qui a requis l'anonymat. Le ministre de l'intérieur, Desiré Tagro a donné, pour sa part, une semaine aux musulmans de réfléchir sur la rétrocession du hadj aux musulmans ou la poursuite de l'organisation du pèlerinage par l'Etat. A l'allure où vont les choses, il sera très difficile pour les structures musulmans de trouver une décision commune. Tant les divergences sont profondes.

ABOU TRAORÉ

Gemini 2.5 Pro (experimental)

Even with Handwritten Text Recognition (HTR)



Une conscience d'une chrétienne, marxiste, d'un responsable d'un Ministère chargé de la gestion du patrimoine de l'Etat. Le patrimoine burkinabé doit-il être confisqué et mis à la disposition des seuls chrétiens ?

3 De part le monde, le vent de la démocratie souffle et heureusement et ce vent est très têtu, pire qu'une sangsue. Si les 30% sans paroles parce que de parents musulmans (stratégie d'occultation de l'Islam oblige) bougent un jour pour manifester leur mécontentement, qui serait responsable de cette situation. Vous première responsable. Ou ces musulmans ignorés, bafoués, humiliés, ulcérés d'une injustice, Assurément, on dira que ce sont ces musulmans qui ne sont pas tolérants, argument facile et rapide de surcroit creux, indigent et diffamatoire que l'on ne cesse de ressasser. Les hommes doués d'une intelligence porteront le démenti et c'est certain, qu'ils soient musulmans ou chrétiens, ou athées.

Nous voudrions nous adresser au Ministre de la Culture que vous êtes. Nous constatons que les troupes théâtrales, qui ont pour auteurs en scène des chrétiens, libèrent leur génie créateur dans l'art de dénigrer l'Islam ; en présentant les rangs entiers de certains marabouts. Et pourtant, nous sommes des catholiques, des protestants, nous sommes informés et avons des avatars du christianisme. Allah nous a créé et à la différence des animaux, il nous a donné un minimum de science pour distinguer le mal du bien. Malheureusement, l'homme, qu'il soit juif, chrétien, musulman ou athée, etc.

Opportunities and Risks of AI OCR



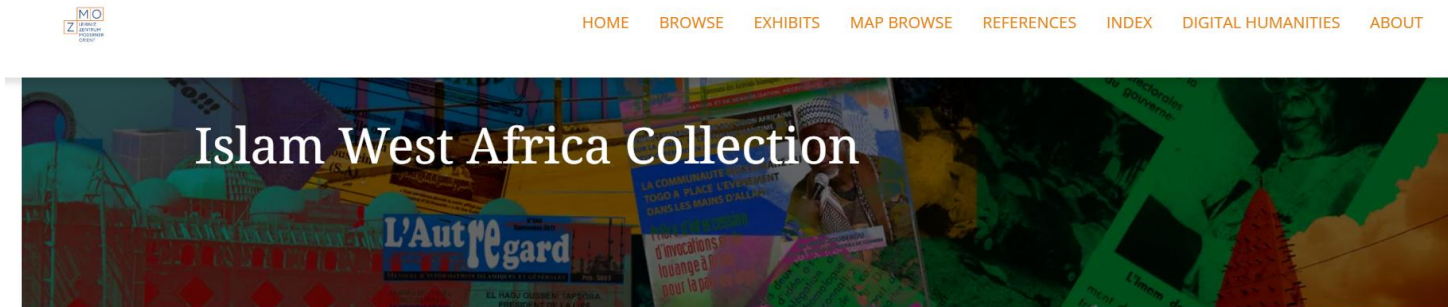
- **Expands access, makes materials analyzable**
- **Risks: hallucination, standardization, hidden changes**
- **AI errors harder to detect than OCR glitches**
- **Safeguards: preserve originals, document prompts/models**
- **Lesson: trade visible flaws for hidden ones**

From Manual Keywords to LLM-Powered Named Entity Recognition (NER)

- Manual keywording too slow, staff/resources limited
- Standard NER misses African names, places
- OCR noise further reduces accuracy
- LLMs use context to identify entities despite errors
- IWAC pipeline: prompts, authority files, OpenRefine reconciliation

Une délégation de six membres, conduite par [M. Fousséni Mama], ministre délégué à la [Présidence], a quitté [Lomé] dimanche à destination de [Moscou] via [Paris]. Cette délégation qui comprend, outre le ministre délégué, [MM. El Hadj Safiou], Vice-Président de l'[Assemblée Nationale], [Mohamed Belly], [Mensah Kassim], [El Hadj Awalou] et l'imam [El Hadj Limou], est invitée par l'[organisation des Musulmans de l'Asie Centrale] à visiter un certain nombre de Républiques autonomes de l'[Union Soviétique]. Dans une brève déclaration faite à la presse avant son départ, [M. Mama] a qualifié ce voyage de mission de prise de contact et d'information : « Les musulmans sont assez nombreux de par le monde et il n'y a rien d'anormal que ceux d'[Union Soviétique] s'intéressent à leurs coréligionnaires du [Togo] ». Il a ensuite déclaré que la délégation visitera le [Kazakstan], l'[Ouzbekhistan], le [Turkménistan], la [Kirghizie] et surtout la grande ville historique d'[Asie Centrale], [Samarcande].

Efficient NER: AI Speed, Human Oversight



Person

Abdou-Salami Abdou Rahim

en [fr](#)

Export

[csv](#) · [json-ld](#) · [tsv](#) · [txt](#)

Title	Abdou-Salami Abdou Rahim
Alternative Title	Abdou Rahim
	Abdoul Rahim Salam
	Abdou Rahim Abdou Salami
	Abdou Rahim-Abdou Salami
	Abdoul Salam Abdoulahim
	Abdul Salam Abdul Rahim
	Abdul Salami Rahim
	Abdou-Salami
	Abdou Salami Abdou Rahimi

- 950 articles processed in 3,5 hours for only \$2,23 (ChatGPT-5 mini)

```
AI_NER > output > item_set_31882_processed_openai_unreconciled_subject_and_topic.csv
1  Unreconciled Value,Count
2  Islam,570
3  Religion,133
4  Communauté,120
5  Fête,76
6  Fraternité,58
7  Transport,58
8  Voyage,57
9  Musulmans,56
10 Air Afrique,56
11 Sacrifice,53
12 Construction,53
13 Dialogue,46
14 Formation,43
15 Organisation,43
16 Financement,43
17 Conseil Économique et Social,43
18 Assemblée Nationale,37
19 Ismaël,34
20 Lamine Fadika,34
21 Abdoulaye Koné,31
22 Décès,27
23 Prix,26
24 Ministère de L'Intérieur,25
25 Parti,24
26 Mouton,24
27 Fraternité Matin,23
28 Calendrier,23
29 Association,22
30 Hébergement,22
31 Président Houphouët-boigny,22
32 Pèlerins,21
33 Inhumation,21
34 Lune,21
35 Associations,20
36 Inauguration,20
37 Grande Mosquée D'Adjamé,20
38 Travail,19
39 Yaya Ouattara,19
40
```


Taming Digital Abundance with LLM-Powered Distant Reading

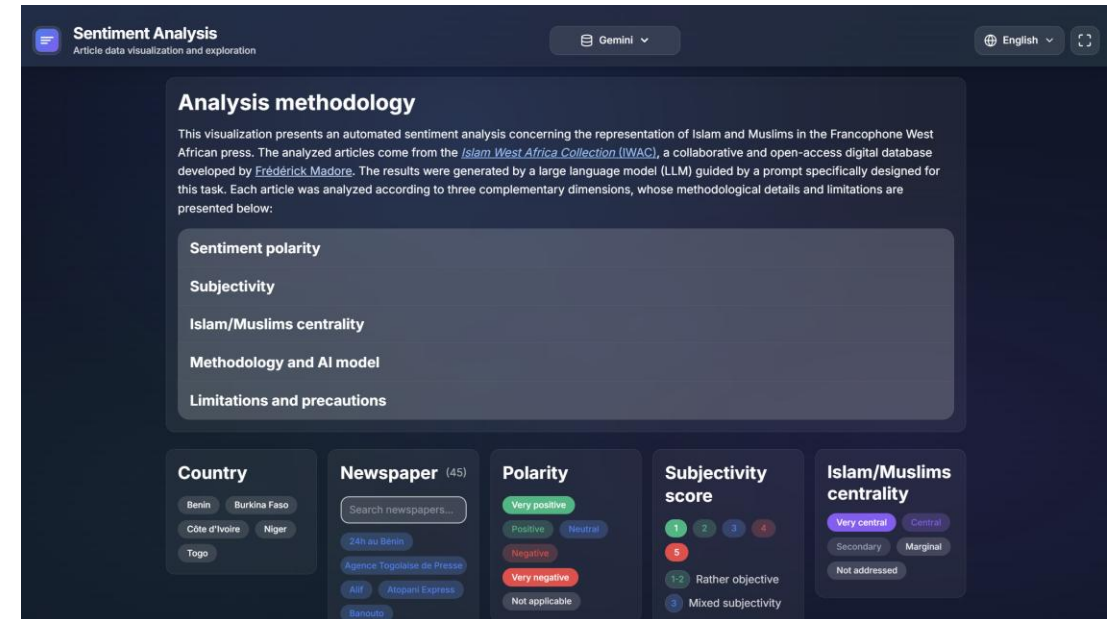
- Digital archiving creates “mountains” of sources (Milligan 2019)
- Researchers digitize vast collections left unread for years
- Traditional close reading cannot scale to millions of words
- Distant reading reveals patterns across entire corpora
- LLMs enable distant reading at unprecedented scale



Image generated by ChatGPT

Sentiment Analysis of Islam in West African Newspapers

- Question: how are Islam and Muslims portrayed?
- Corpus: 10,639 articles, analyzed with ChatGPT & Gemini
- Three dimensions: polarity, subjectivity, centrality
- Interactive dashboard for trends, heatmaps, distributions



<https://fmadore.github.io/IWAC-sentiment-analysis/>

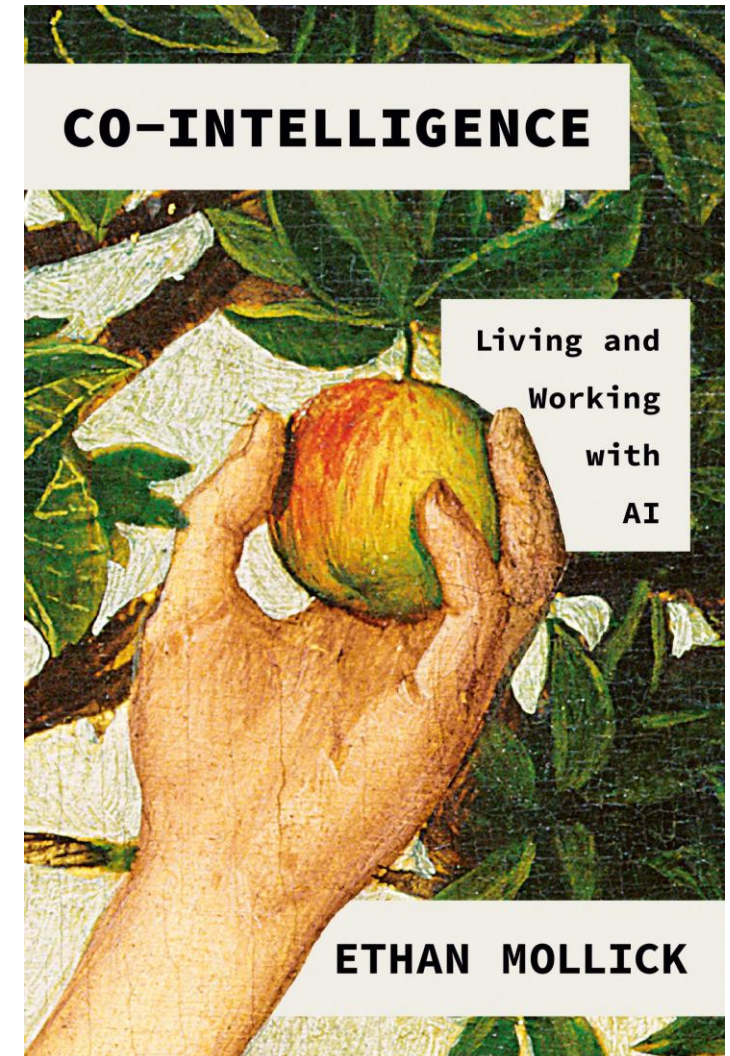
Terrorism Coverage and Positive Portrayals of Islam

- Expected: terrorism terms linked to negative portrayals
- Extremists described as “madmen of Allah,” using “Islam as a cover”
- But many articles with these terms coded as positive
- Positive framing emphasized Islam’s peace, tolerance, and dissociation from violence
- Coverage of the 2017 attack on the Café Aziz Istanbul in Ouagadougou was noted as discussing “a terrorist attack and the ensuing investigation, without mentioning Islam or Muslims, or establishing any religious connection with the event or its perpetrators.”



LLMs as Co-Intelligence in Historical Research

- **€15 and 24 hours replaced months of manual coding**
- **Real value: consistency across thousands of texts**
- LLMs evaluate each article independently, without fatigue
- **Serve as “co-intelligence” (Mollick 2024), not replacements**
- **AI triages scale; humans provide context and interpretation**



From Keyword Limits to Semantic Search in Digital Collections

- Keyword search distorts outcomes, shaped by language limits
- Few users interrogate its methodological limits
- Polysemy, synonymy, and linguistic change obscure meanings
- OCR errors and uneven metadata reduce discoverability
- Researchers constrained by search opacity and user bias

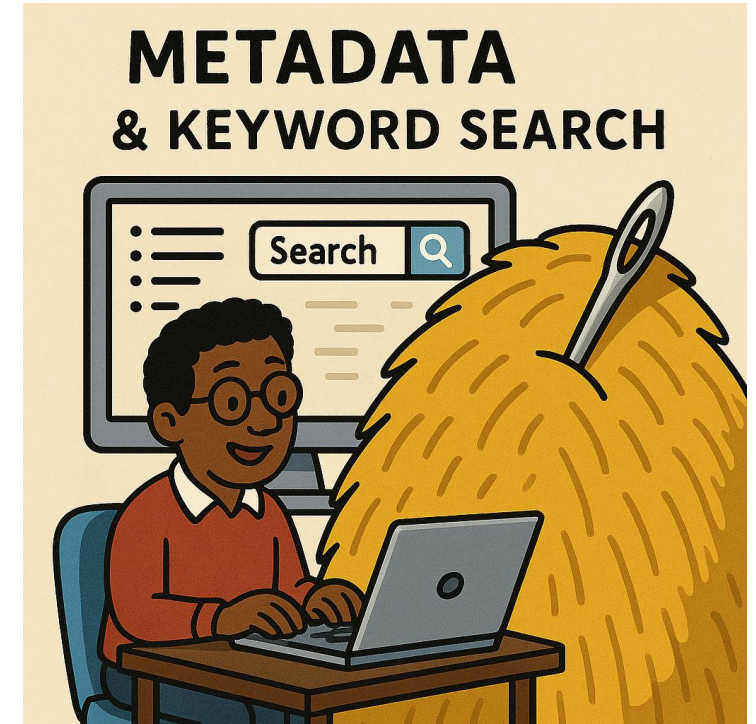
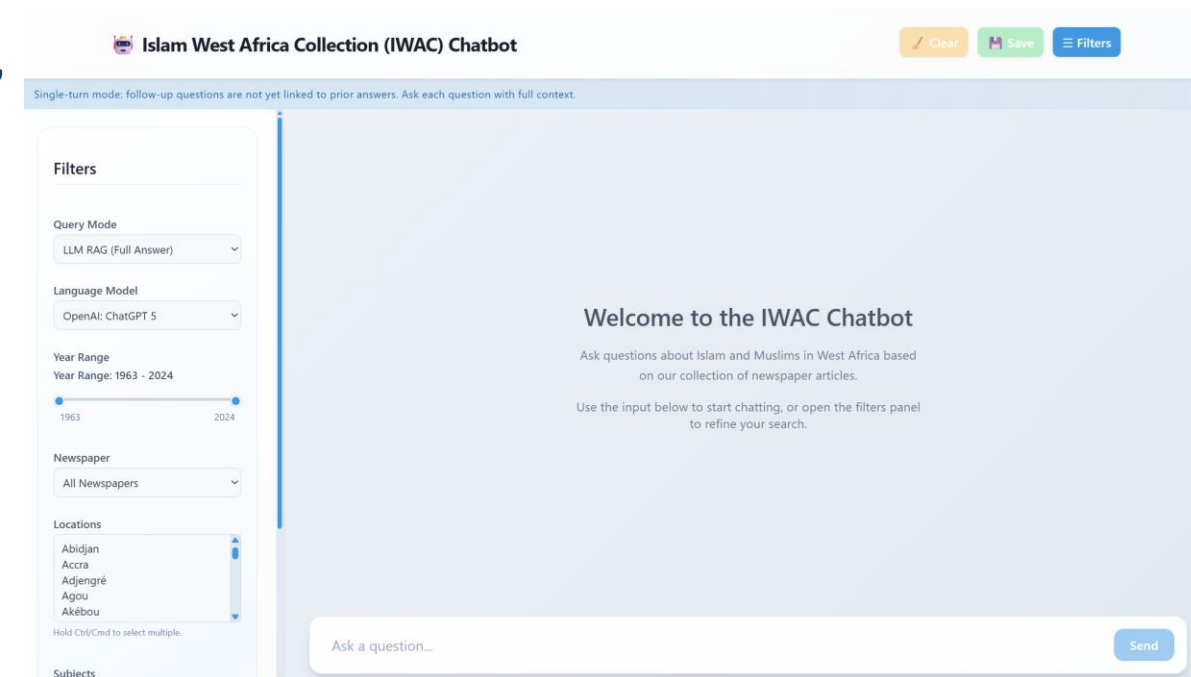


Image generated with ChatGPT:
"Finding Needles in Digital Haystacks"

RAG: From Keywords to a “GPS of Meaning”

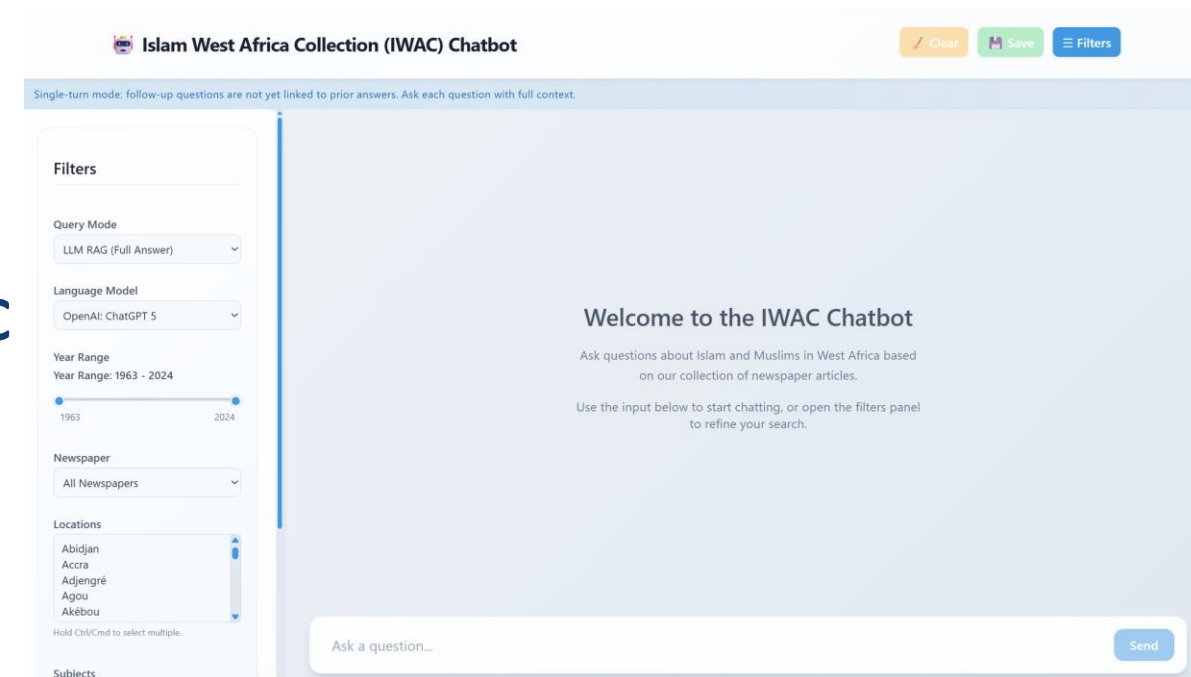
- RAG links LLMs to a curated archive
- Texts converted into semantic “signatures” or coordinates
- Similar topics cluster close on this map of meaning
- Queries retrieve conceptually related passages, not just words
- Anchors answers in IWAC sources, reducing hallucinations



<https://github.com/fmadore/IWAC-RAG>

Building and Using the IWAC Chatbot

- Prepares collection: cleaning, chunking, semantic signatures
- Vector library enables rapid retrieval of relevant passages
- LLM composes response grounded in IWAC documents
- Prompt and context engineering refine reliability and rigor
- Transparent citations ensure scholarly accountability



<https://github.com/fmadore/IWAC-RAG>

In Guise of Conclusion: Some Ethical Considerations and a Work in Progress



- **AI transforms IWAC corpus into navigable research resource**
- **Human expertise remains central to interpretation and critique**
- **Transparency, bias, and reproducibility raise ethical concerns**
- **Safeguards: open documentation, source preservation, verification**
- **Toward durable principles, not chasing latest models**