

Exploring Islamic manuscript catalogues with large language models: building a Q&A system and insightful visualizations

Alicia González Martínez

“The Evolution of Islamic Societies (c.600-1600 CE):
Algorithmic Analysis into Social History”



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

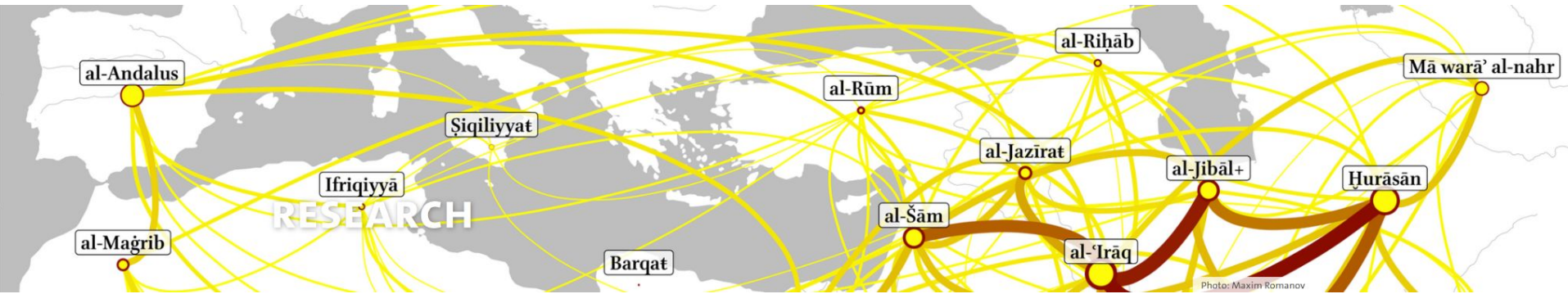


Abstract

The recent groundbreaking advancements in Natural Language Processing (NLP) and Artificial Intelligence (AI) have transformed the field of digital humanities, enabling us to address complex research questions with unprecedented precision. This talk highlights the potential of large language models to push forward the field of Islamic manuscript studies. While traditional digital techniques have their limitations, large language models offer an efficient and scalable solution for extracting structured data from non-digitized or partially digitized manuscript catalogues. With the assistance of AI tools, we can significantly reduce the time and effort required to process and utilize these valuable resources.

In this presentation, I will showcase how advanced digital humanities techniques can be applied to: (1) digitize and structure a manuscript catalogue, (2) support interactive exploration and question-answering capabilities, (3) create data visualizations that provide new insights into the contents of the catalogue. By demonstrating these possibilities, this talk aims to inspire researchers in Islamic manuscript studies to explore the vast potential of large language models for advancing research in the field.

The Evolution of Islamic Societies (c. 600–1600 CE)



600

1600 CE

Automatically Annotated Corpus



	SHR	NSB	NSB	NSB	NAS	NAS	NAS	NAS	ISM	Onomastic Entities
	أحمد بن محمد بن محمد الشهاب بن الصدر بن الصلاح الأنصاري القاهري الشافعي ويعرف بابن صدر الدين .									Nasab (onomastic section)
Teacher	ولد سنة خمس وتسعين وسبعمئة تقريبا ونشأ حفظ القرآن والمنهاج رفيقا للوالد عند الفقيه الشمس السعدي									Birth Date
	وعرض علي جماعة واشتغل قليلا وسمع شيخنا وغيره ومما سمعه ختم البخاري بالظاهرية وتنزل بالبيبرسية وتكسب بالشهادة في حانوت باب القوس داخل باب القنطرة وفي سوق الرقيق ولم يكن فيها بالماهر معرفة وخطا ولكنه كان لا بأس به سكونا ومحافظة على الجماعة ثم انجماعا واقتصادا في معيشته مع دريهمات بيده ربما يعامل فيها وقد حج غير مرة وجاور . مات في ليلة الاثنين منتصف رمضان سنة أربع وثمانين وصلى عليه من الغد ودفن بحوش البيبرسية وأوصى بثلثه لمعينين وغيرهم رحمه الله وإيانا .									Death Date

~100 Million words

Large Language Models (LLMs)

It's a statistical system that tries to imitate human language.

It models or "learns" semantics from a vast corpus of human texts

"you shall know a word by the company it keeps", John Firth (1957)

A **bear** wandered into the campsite.

I can't **bear** this pain anymore.

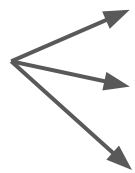
Large Language Models (LLMs)

It's a statistical system that tries to imitate human language.

It models or "learns" semantics from a vast corpus of human texts

"you shall know a word by the company it keeps", John Firth (1957)

I can't bear ___?___



Next word	Probability
this	0.88
wandered	0.03
...	...

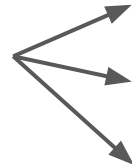
Large Language Models (LLMs)

It's a statistical system that tries to imitate human language.

It models or "learns" semantics from a vast corpus of human texts

"you shall know a word by the company it keeps", John Firth (1957)

I can't bear ?



Next word	Probability
this	0.88
wandered	0.03
...	...

$$\begin{bmatrix} 0.3 \\ 0.5 \\ -0.1 \\ 5.4 \\ 0.9 \\ -2.0 \end{bmatrix}$$

word embeddings

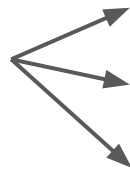
Large Language Models (LLMs)

It's a statistical system that tries to imitate human language.

It models or "learns" semantics from a vast corpus of human texts

"you shall know a word by the company it keeps", John Firth (1957)

I can't bear ___?___



Next word	Probability
this	0.88
wandered	0.03
...	...



Stochastic parrot

Large Language Models (LLMs)



Stochastic parrot



Large Language Models (LLMs)



Stochastic parrot



Artificial intelligence is not intelligent

Large Language Models (LLMs)



Stochastic parrot



Artificial intelligence is not intelligent

LLMs imitate human language but lack true understanding

Large Language Models (LLMs)



Stochastic parrot



Artificial intelligence is not intelligent

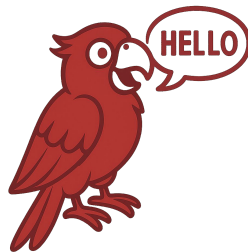
LLMs imitate human language but lack true understanding

LLMs reflect all the biases and misinformation present in the training data

Large Language Models (LLMs)



Stochastic parrot



Artificial intelligence is not intelligent

LLMs imitate human language but lack true understanding

LLMs reflect all the biases and misinformation present in the training data

LLMs commit mistakes

Large Language Models (LLMs)

LLMs excel in many tasks!

Summarization

Domain specific Chatbots

Creative writing

High quality translations

Structuring unstructured data

PUBLICATIONS

DE

L'ÉCOLE DES LANGUES ORIENTALES VIVANTES

II^E SÉRIE — VOLUME 'X

LES MANUSCRITS ARABES DE L'ESCURIAL

Derenbourg, Hartwig; Renaud, H.P.J.; Lévi-Provençal, E. Les manuscrits arabes de l'Escurial. Paris: Ernest Leroux, Paul Geuthner, 1884-1928. (3 v)

<https://rbme.patrimonionacional.es/s/rbme/page/-rabes>

Derenbourg Catalogue

Machine-readable representation of the text

Optical Character Recognition (OCR)

70.

Titre : السفر الاول من المرادى « Le premier volume d'Al-Mourâdi. » C'est le premier volume du commentaire sur l'*Alfiyya* d'Ibn Mâlik, par Selams ed-Din Hasan ibn Kâsim Al-Mourâdi (cf. mss. 4 et 12). Commencement : الحمد لله والشكر لله وصلاته على محمد خير نبي ارسله اما بعد فهذا مختصر توضيح لمقاصد الفية ابن ملك الخ (cf. éd. de Dieterici, p. ١٣).

Papier. Écriture Magrèbine. 130 feuillets. 27 lignes par page. Sans date.

image

70.

Titre: السفر الاول من المرادى. Le premier volume l'Al Mouradi. C'est le premier volume du commentaire sur l'*Alfiyya* d'Ibn Malik, par Selams ed-Din Hasan ibn Kâsim Al-Mouradi (cf. muss. 4 et 12). Commencement: الحمد لله والشكر لله وصلاته على محمد خير نبي ارسله اما بعد فهذا مختصر توضيح المقاصد الفية ابن ملك الخ (cf. éd ce Dieterici, p. ٢١٣)

Papier. Écriture Magrèbine. 130 feuillets. 27 lignes par page. Sans date.

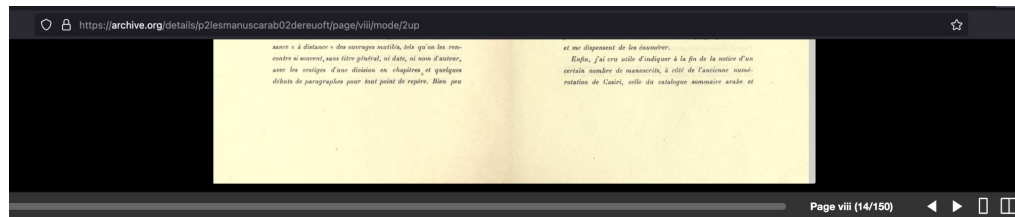
text

Derenbourg Catalogue

Machine-readable representation of the text

<https://archive.org>

Optical Character Recognition (OCR)



Les manuscrits arabes de l'Escurial

by Derenbourg, Hartwig, 1844-1908; Escorial. Biblioteca

Publication date 1884
Topics Manuscripts, Arabic, Manuscripts
Publisher Paris E. Leroux
Collection robarts; toronto; university_of_toronto
Contributor Robarts - University of Toronto
Language French
Volume 2, pt.2
Item Size 218.2M

26

Addeddate 2008-10-28 16:55:57
Associated-names Escorial. Biblioteca
Call number ABA-2348
Camera 1Ds
External-identifier um:oclc:record:978586871
Foldoutcount 0
Identifier p2lesmanuscarab02dereuoft
Identifier-ark ark:/13960/t3708bb1x
Lccid 331202
Ocr_converted abbyy-to-hocr 1.1.37

2,702 Views

5 Favorites

DOWNLOAD OPTIONS

ABBY GZ 1 file
B/W PDF 1 file
CHOCR 1 file
CLOTH COVER DETECTION LOG 1 file
DAISY For users with print-disabilities Temporarily Unavailable
EPLUB Temporarily Unavailable
FULL TEXT 1 file
HOCH 1 file
ITEM TILE 1 file

Titre : ,j^ S^ J-o^l ^ ^^^^^ j^j^ ~j^ w^-
Commentaire du Mugiz d'Ibn an-Nafis (cf. supra, n'' 792
et 828), par [Gamâl ad-dîn (ou Sams ad-dîn) Abu Abd
AUâh Muhammad] al-Aksarâ'i [at-Tabrizî] f vers 800 /
1397; cf. Hâggi IJalifa, V, 304; Brockelmann, I, 407, III ^^
\$e, etS/)/., I, 824.

L'ouvrage, dont il existe plusieurs exemplaires à la biblio-
thèque Bodléienne et à Leyde, est intitulé : U ^^Jl J^a.,
ilj! 5 jSiùJl ^ i^^ « Explication du Mûgiz et des diffi-
cultés et énigmes qu'il contient ». Ses divisions sont celles du
Mûgiz, en 4 /\i. Partout, on trouve : ...Jy^ ...^^1 JU.
Incipit : wi^^ *if ^JJaJ! jU A*j j ...^^^^IjJl ^j t^ a^ssh!

Copie datée du dimanche 9 Ramadan 840/17 Mars 1437.
Autre exemplaire (incomplet), infra, n» 867 g-

Papier. Écriture orientale. 222 feuillets, les deux premiers plus
modernes. 26 lignes par page. Dimensions : 0.18 X 0.27. (Cas. 831.)

Derenbourg Catalogue

Machine-readable representation of the text

Optical Character Recognition (OCR)

70.

Titre : السفر الاول من المرادى « Le premier volume d'Al-Mourâdi. » C'est le premier volume du commentaire sur l'*Alfigya* d'Ibn Mâlik, par Seham ed-Din Hasan ibn Kâsim Al-Mourâdi (cf. mss. 4 et 12). Commencement : الحمد لله والشكر لله وصلاته على محمد خير نبي ارسله اما بعد فهذا مختصر توضيح لمقاصد الفية ابن ملك الخ. La fin manque, et le volume s'arrête au milieu du chapitre intitulé : اعمال اسم الفاعل (cf. éd. de Dieterici, p. 112).

Papier. Écriture Magrèbiue. 130 feuillets. 27 lignes par page. Sans date.

Tesseract
fr+ara

very poor quality, useless

10,

Titre: السش الاول من المرادى La premier volume d'AT Mourhdi,s C'est le premier volume du commentaire «ur || "مطل" Mdik, par Seham ed-Din Llasan Ibn Kâsin AlMouradi (ef, mess. 4 et 12} Conmenvement : a ad 214

5

+

À

ا

...

£

ام ابن ملك AA المقاصد La fin manque, et le volume Dietcrici, p, ae),

Xe

l'upier, Bcriture Magrèbiue. 140 feuillets, 27 يورا par page, Sas dits

Derenbourg Catalogue

Machine-readable representation of the text

Optical Character Recognition (OCR)

70.

Titre : السفر الاول من المرادى « Le premier volume d'Al-Mourâdi. » C'est le premier volume du commentaire sur l'*Alfiyya* d'Ibn Mâlik, par Selams ed-Din Hasan ibn Kâsim Al-Mourâdi (cf. mss. 4 et 12). Commencement : الحمد لله والشكر لله وصلاته على محمد خير نبي ارسله اما بعد فهذا مختصر توضيح لمقاصد الفية ابن ملك الخ La fin manque, et le volume s'arrête au milieu du chapitre intitulé : اعمال اسم الفاعل (cf. éd. de Dieterici, p. 112).

Papier. Écriture Magrêbine. 130 feuillets. 27 lignes par page. Sans date.

Google
Vision API

: Titre السفر الاول من المرادى -Le premier volume l'Al-Mouradi. C'est le premier volume du commentaire sur l'*Alfiyya* d'Ibn Malik, par Selams ed-Din Hasan ibn Kâsim Al-Mouradi (cf. mss. 4 et 12). Commencement: والشكر لله وصلاته على محمد خير نبي ارسله اما بعد فهذا مختصر 4 توضيح المقاصد الفية ابن ملك الخ La fin manque, et le volume s'arrête : an milieu du chapitre intitule اعمال اسم الفاعل ofil de) 657 Dieterici, p. cir).

Papier. Écriture Magrêbine. 130 feuillets. 27 lignes par page. Sans date,

not perfect, but fairly good

Derenbourg Catalogue

Machine-readable representation of the text

Optical Character Recognition (OCR)

70.

Titre : السفر الاول من المرادى « Le premier volume d'Al-Mouradi. » C'est le premier volume du commentaire sur l'Alfiyya d'Ibn Malik, par Selams ed-Din Hasan ibn Kâsim Al-Mouradi (cf. muss. 4 et 12). Commencement : الحمد لله والشكر لله وصلاته على محمد خير نبي ارسله اما بعد فهذا مختصر توضيح المقاصد الفية ابن ملك الخ La fin manque, et le volume s'arrête au milieu du chapitre intitulé : اعمال اسم الفاعل (cf. éd. de Dieterici, p. ٢١٣).

Papier. Écriture Magrêbine. 130 feuillets. 27 lignes par page. Sans date.

Google
Vision API

: Titre السفر الاول من المرادى -Le premier volume l'Al Mouradi. C'est le premier volume du commentaire sur l'Alfiyya d'Ibn Malik, par Selams ed-Din Hasan ibn Kâsim Al-Mouradi (cf. muss. 4 et 12). Commencement: 4 الحمد لله والشكر لله وصلاته على محمد خير نبي ارسله اما بعد فهذا مختصر توضيح المقاصد الفية ابن ملك الخ La fin manque, et le volume s'arrête : an milieu du chapitre intitule اعمال اسم الفاعل ofil de) 657 Dieterici, p. cir).

Papier. Écriture Magrêbine. 130 feuillets. 27 lignes par page. Sans date,

rule-based
norm

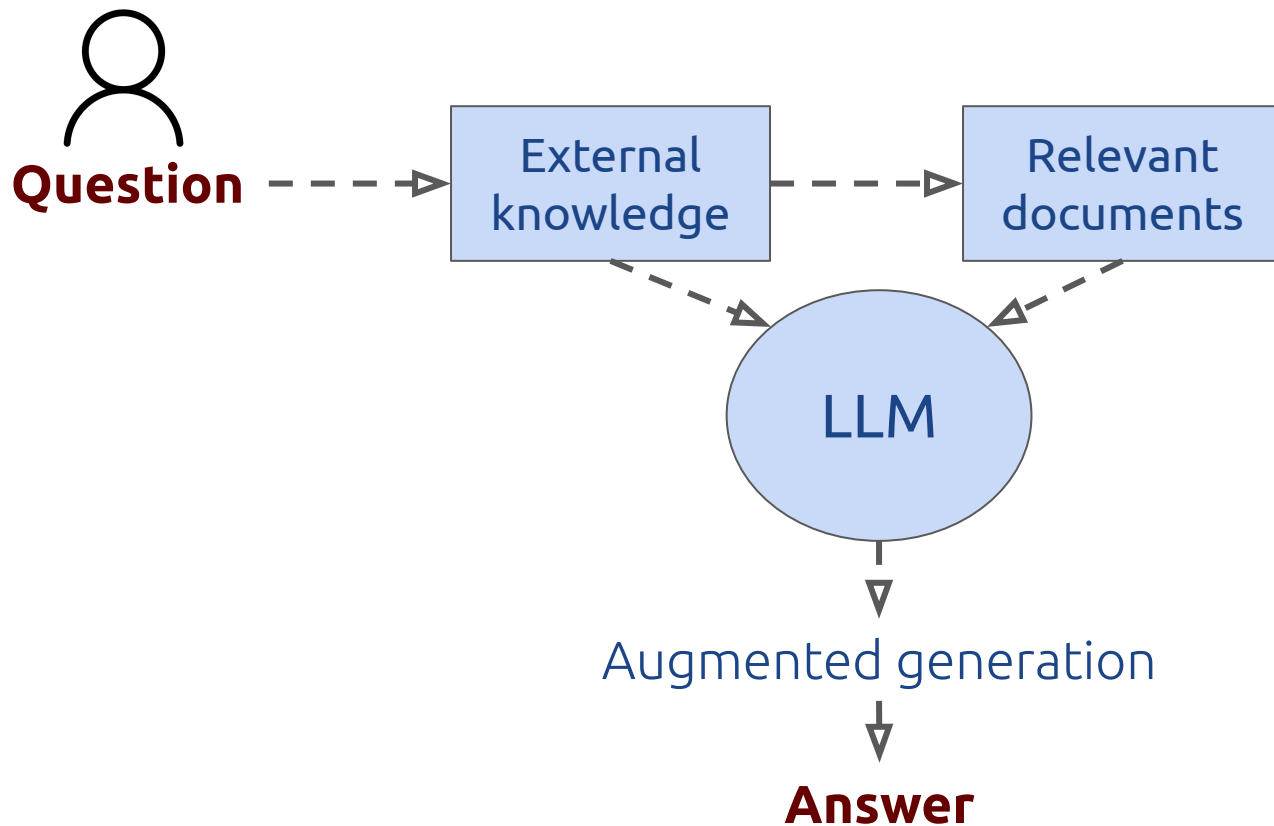
Titre: السفر الاول من المرادى. Le premier volume l'Al Mouradi. C'est le premier volume du commentaire sur l'Alfiyya d'Ibn Malik, par Selams ed-Din Hasan ibn Kâsim Al-Mouradi (cf. muss. 4 et 12). Commencement: الحمد لله والشكر لله وصلاته على محمد خير نبي ارسله اما بعد فهذا مختصر توضيح المقاصد الفية ابن ملك الخ La fin manque, et le volume s'arrête an milieu du chapitre intitule: اعمال اسم الفاعل *cf. éd. ce Dieterici, p. ٢١٣)

Papier. Écriture Magrêbine. 130 feuillets. 27 lignes par page. Sans date,

still not perfect, but best result so far
I had to apply post-correction, but it is a relatively fast task

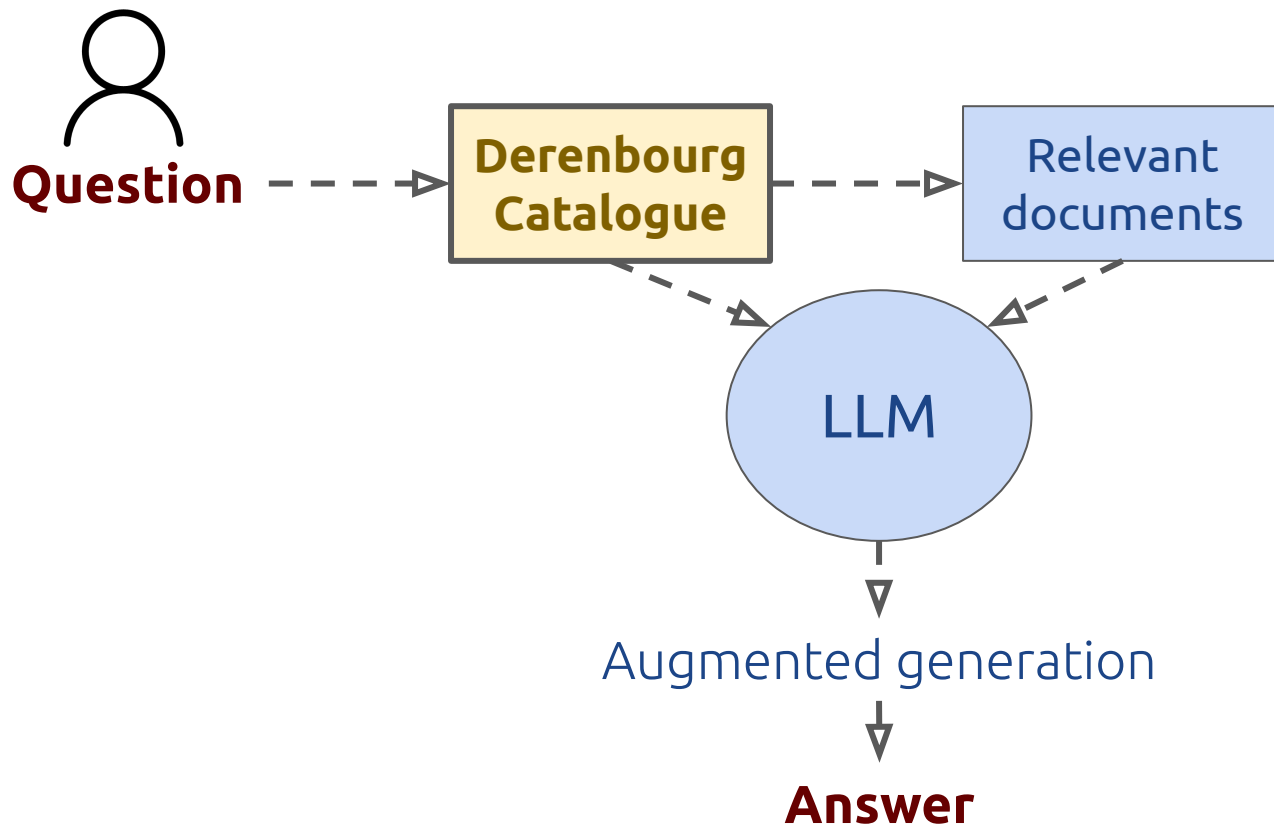
Using Retrieval-Augmented Generation (RAG) for Q&A

technique that enhances the capabilities of LLMs by integrating information retrieval systems



Using Retrieval-Augmented Generation (RAG) for Q&A

technique that enhances the capabilities of LLMs by integrating information retrieval systems



Using Retrieval-Augmented Generation (RAG) for Q&A

Open WebUI

<https://openwebui.com>

gpt-4.1 ▾ +



Derenbourg

Collection

Search the catalogue for authors explicitly recorded as having died between 800 and 810 AH.

OI gpt-4.1

Based on the provided context, the only author explicitly recorded as having died between 800 and 810 AH is:

- Abou Zaid Abd er-Rahmân ibn 'Alî ibn Sâlih Al-Makoûdî, who is stated to have died in 807 AH (1404 CE).
- The precise reference: "Celui-ci, d'après une longue biographie insérée aux fol. 5 r° à 7 v°, naquit en 736 de l'Hégire (1335 ap. J.-Ch.) et mourut en 807 (1404 ap. J.-Ch.)"

References from 1 0033.txt 2 0008.txt and 31 more ▾



Commentaire de 'Abd el-Wâhid ibn Aḥmad ibn Moḥammad Al-Ḥasanî sur un poème qui n'a rien de grammatical, d'Aboû Zaid 'Abd er-Raḥmân ibn 'Alî ibn Ṣâlih Al-Makoûdî, le commentateur de l'*Alfiyya* d'Ibn Mâlik (voir les manuscrits 6, 2°; 7, 1°). Celui-ci, d'après une longue biographie insérée aux fol. 5 r° à 7 v°, naquit en 736 de l'Hégire (1335 ap. J.-Ch.) et mourut en 807 (1404 ap. J.-Ch.).

Using Retrieval-Augmented Generation (RAG) for Q&A

Open WebUI

<https://openwebui.com>

gpt-4.1 ∨ +



Which catalogue entries contain multiple works?

o1 gpt-4.1

Several catalogue entries in the provided context contain multiple works within a single manuscript. These are typically identified by numbered sections (e.g., 1°, 2°, 3°) or by explicit mention of more than one text or commentary within the same manuscript description. Here are examples:

1. Manuscript described as containing:

- «Traité des cent régents grammaticaux (العوامل المائة), par Abd al-Kahir Al-Djordjant.»
- «Commentaire abrégé d'Al-Makondî sur l'Alfiyya d'Ibn Mâlik.»
- «Traité des flexions, intitule: مراخ الارواح «Le repos des esprits», par Ahmad ibn Ali ibn Masoad.»

These are described together in one entry.

2. Manuscript with:

- «Livre intitulé: Les trésors réservés (الازهرية).
- «Livre des régents grammaticaux» by Sirâdj ed-Din Mahmoud ibn Yousouf de Herat (الهروى).
- «Commentaire sur les Signes évidents» (شرح الآيات البينات).

All listed consecutively in a single manuscript.

3. Another example:

- «Traité des propositions grammaticales», by Aboû 'I-Kasim Abd er-Rahman ibn Ishalk AzZadjdjâdjî.
- «Livre intitulé: La langue pure» (النصيح), attributed to Tha'lab.
- «Les mots dont le premier radical peut recevoir les trois voyelles», by Koutroub.

Multiple works, all within one described manuscript.

4. A manuscript listed as:

- Commentaire de Mas'ouîd ibn 'Omar At-Taftâzânî sur le traité de Izz ed-Din Az-Zandjani
- «Mataw' Mubarak yachetamil 'ala matin ar-Rahbiyy ↓ a sharhiha» (versified treatise on inheritance, commentary after)

Structuring unstructured data

Frew shot prompt

Autre exemplaire d'un premier volume du commentaire sur l'Alfiyya, par Al-Mourâdî. L'auteur est nommé ici, en title: بدر الدين ابو على حسن بن قاسم بن عبد الله بن على المرادى المالكي. Manuscrit daté de 780 de l'Hégire (1378 ap. J.-Ch.). Le dernier chapitre commenté et sur les أسماء اللازمة (cf. l'édition de Dieterici, p. 106-109).

1 Papier. Écriture Magrébine, 114 feuillets. 27 lignes par page.

```
{
  "catalogue_entry": [
    {
      "book": ["Commentaire sur l'Alfiyya"],
      "date_of_book": {"hijri": 780, "CE": 1378},
      "author": [
        "Al-Mourâdî",
        "بدر الدين ابو على حسن بن قاسم بن عبد الله بن على المرادى المالكي"
      ],
      "date_of_author": null,
      "book_commented_upon": ["Alfiyya"],
      "date_of_book_commented_upon": null,
      "author_of_book_commented_upon": [],
      "date_of_author_of_book_commented_upon": null,
      "date_of_publication": [],
      "incipit": null,
      "explicit": null,
      "material": "Papier",
      "script": "Magrébine",
      "number_of_folios": 114,
      "lines_per_page": [27]
    }
  ]
}
```

Structuring unstructured data

kabikaj.github.io/derenbourg

Search in Derenbourg Catalogue

Book	Book Date (Hijri)	Book Date (CE)	
Author	Author Date (Hijri)	Author Date (CE)	
Book Commented Upon	Book Commented Upon Date (Hijri)	Book Commented Upon Date (CE)	
Author of Book Commented Upon	Author of Book Commented Upon Date (Hijri)	Author of Book Commented Upon Date (CE)	
Publication Place	Publication Date (Hijri)	Publication Date (CE)	
Material	Script	Number of Folios	Lines per Page
Incipit			
Explicit			
Search Text			
Filter Results		Reset Filters	

Found 205 manuscripts

Le livre de Siboûyéh, گاب سیبویه

Book Dates: 629 (Hijri) / 1231 (CE)

Author: Aboû Bischr Amr ibn 'Othman ibn Kanbar, سيبيويه

Author Dates: 180 (Hijri) / 796 (CE)

Publication: Unknown

Book Commented Upon: N/A

Book Commented Upon Dates: ? (Hijri) / ? (CE)

Material: Papier

Script: Magrébine

Folios: 271

Lines per Page: 27

Author of Book Commented Upon: Unknown

Author of Book Commented Upon Dates: ? (Hijri) / ? (CE)

Structuring unstructured data

kabikaj.github.io/derenbourg

Search in Derenbourg Catalogue

Book	780	Book Date (CE)	
Author	Author Date (Hijri)	Author Date (CE)	
Book Commented Upon	Book Commented Upon Date (Hijri)	Book Commented Upon Date (CE)	
Author of Book Commented Upon	Author of Book Commented Upon Date (Hijri)	Author of Book Commented Upon Date (CE)	
Publication Place	Publication Date (Hijri)	Publication Date (CE)	
Material	Magrébine	Number of Folios	Lines per Page
Incipit			
Explicit			
Search Text			
Filter Results	Reset Filters		

Found 1 manuscript

Commentaire sur l'Alfiyya

Book Dates: 780 (Hijri) / 1378 (CE)

Author: Al-Mourâdi, بدر الدين ابن علي حسن بن قاسم بن عبد الله بن علي المرادي المالكي

Author Dates: ? (Hijri) / ? (CE)

Publication: ? (Hijri) / ? (CE)

Book Commented Upon: Alfiyya

Book Commented Upon Dates: ? (Hijri) / ? (CE)

Material: Papier

Script: Magrébine

Folios: 114

Lines per Page: 27

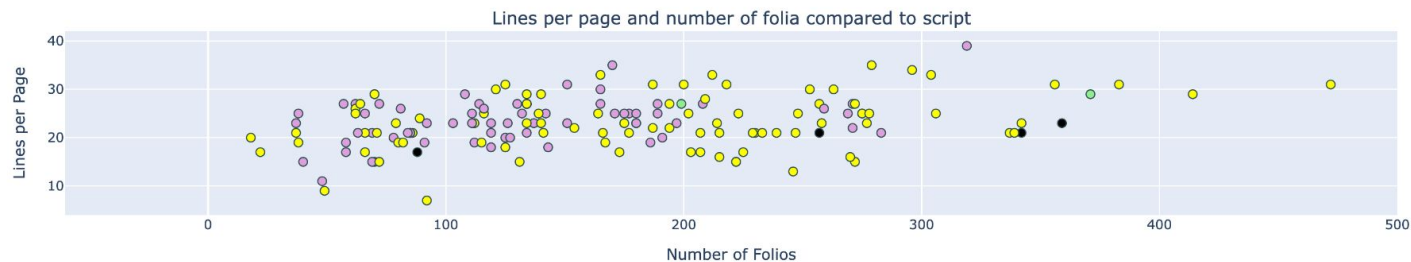
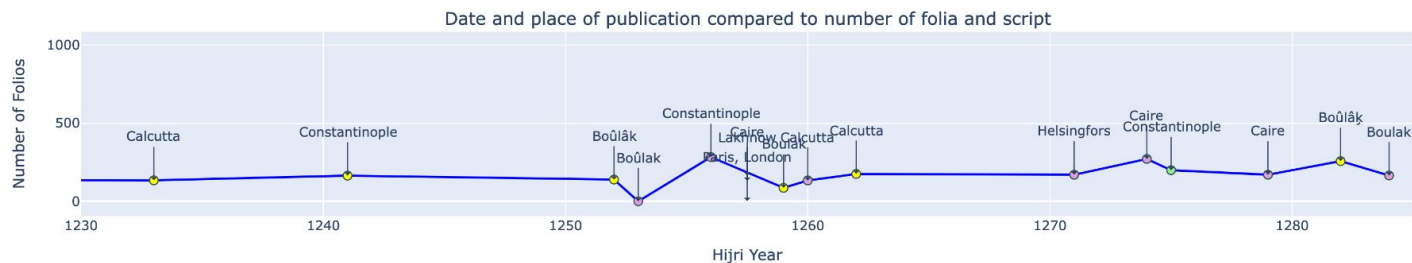
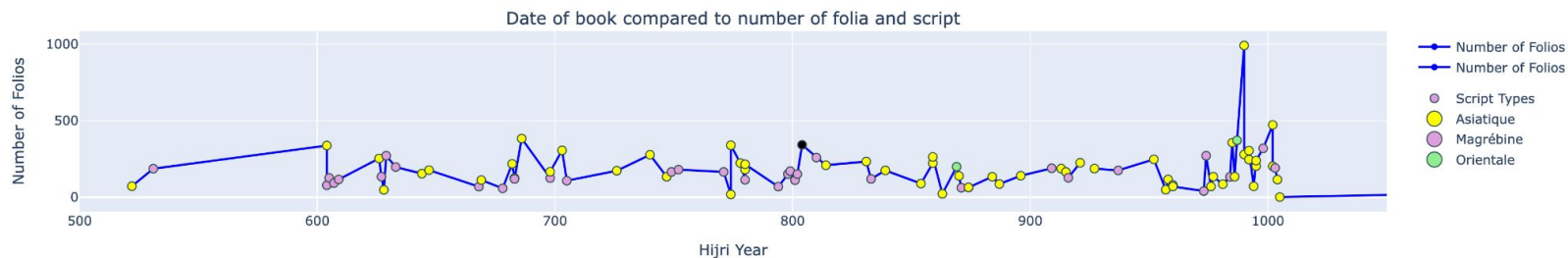
Author of Book Commented Upon: Unknown

Author of Book Commented Upon Dates: ? (Hijri) / ? (CE)

Visualizing the data

kabikaj.github.io/derenbourg

Sample of Manuscript Plottings



How can LLMs enhance humanities research

How can LLMs enhance humanities research

Semantic search: fast retrieval of knowledge

How can LLMs enhance humanities research

Semantic search: fast retrieval of knowledge

Speed up data curation

How can LLMs enhance humanities research

Semantic search: fast retrieval of knowledge

Speed up data curation

Understand global trends and patterns

Thank You

kabikaj.github.io/derenbourg

github.com/kabikaj/derenbourg