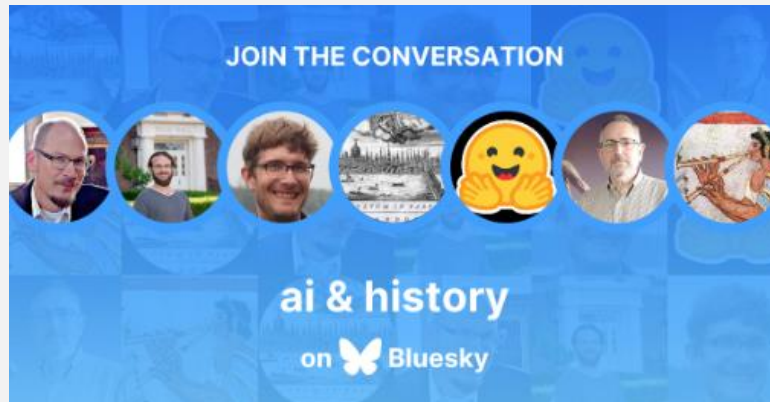


AI assistants and agents: A New Skill Set for Historians?

IHR Digital History seminar

Tuesday, November 19th 5:30pm GMT

Speaker: Colin Greenstreet ; Session chair: James Baker



A New Skill Set for Historians explores the potential for Large Language Model based intelligent assistants and agents to support historical research.

The speaker makes the case for historians and archivists acting together to build knowledgeable, effective, and serendipitous assistants and agents within the history domain, and explores two parallel routes to do so.

Firstly, the construction of large language models using retrieval augmented generation techniques to draw on specialized domain specific vectorbases. Secondly, the construction of large scale licence clear datasets of historical manuscript and printed knowledge for fine-tuning of medium-sized large language models

introduction

betweene marcellis and mallegae uppon the towerteenth
day of January last past of of the Ilane of Majyorca
was surprized and taken by three Turcks men
war of Algeir and tooke her Company prisoners and
carried parte of them to Algeir and parte to Tunis, and
amongste the reste there was one William wood a young [man]
of aboute twenty yeares of age the son of one William
Wood of Saint Katherines, Grocer who was Chirurgion
of the shipp, and him the Turcks carried a Prisoner
to the hee was sold a slave in open Marcket
thence hee was sold a slave in open Marcket
thence hee was sold a slave in open Marcket



MarineLives

@marinelives.bsky.social

1.3K followers 582 following 723 posts

Lives touched by the marine, 1574-1688. A volunteer led collaboration publishing to the Commons. Please join our ai-and-history collaboratory for all students of history and public historians: github.com/Addaci/marin...

Edit Profile

...

- Banker; management consultant; pharmaceutical executive; entrepreneur
- Co-founder of volunteer led history collaboration MarineLives, 2012
- Co-founder of mental health charity Heal Traumas International, 2022
- Tech agnostic; no coder
- First engaged with NLP in 2013
- First engaged with machine transcription in 2016
- Built 400k+ word groundtruth in 2022
- Actively exploring large language models since early 2023 (mental health; endodontics; water rights; history)
- Co-founder of a not-for-profit LLM-enabled mental health training app, 2024

research challenges

- Exponentially increasing quantities of digitized printed material and manuscripts
- Limited metadata for digitized material of all sorts (including sound and video)
- Extremely limited transcription of digitized manuscripts and poor OCR for printed documents
- Severe constraints on academic and archival resources

the opportunity

Analyze Categorize Contextualize Correct Expand
Extract Geotag Interrogate Link Map Modernize
Role play Structure Summarize Translate

editor interlocutor linguist
software engineer tutor
systematic reviewer

- Increase scale and scope of primary source access and interrogation
- Increases linkage and enrichment of large scale text bodies and metadata
- And deepen the quality of close granular reading of text
- Broaden linguistic scope of historians with limited personal language capabilities

domain specific data

Our professional attention to context, annotation, and data quality together with the institutional strength of the GLAM sector in generating semi-structured data is a huge asset in a world of LLM-enabled research.

The challenge is to achieve quality at scale.

Frontier models

- LLMs are here to stay and are largely built from contemporary data and historians are going to use them, so we might as well get as much data and metadata into frontier model pre-training as we can (under acceptable licensing conditions)

Vectorbases

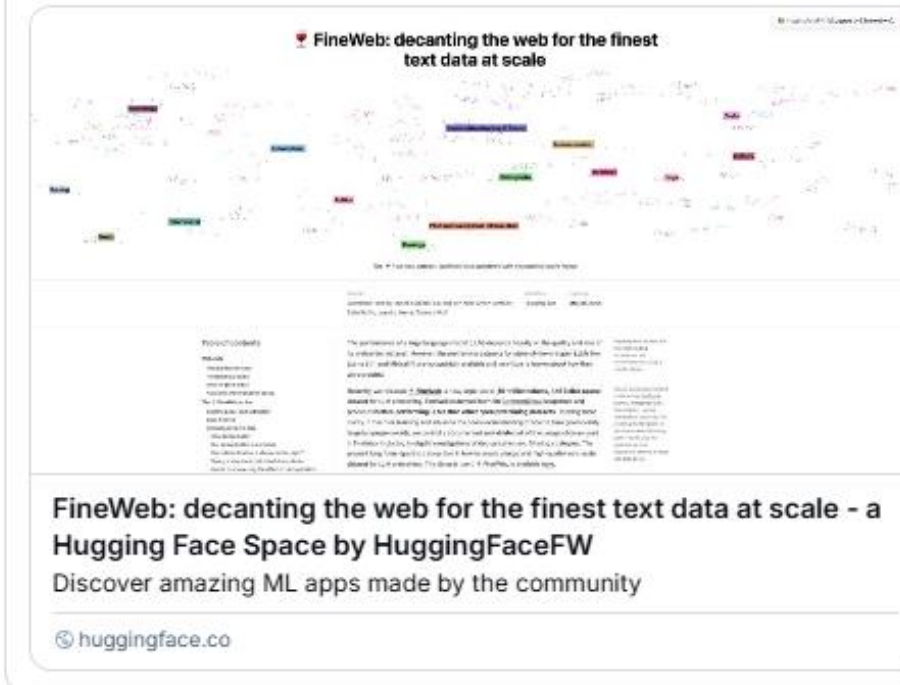
- Vectorbases of varying depth, breadth and granularity have huge potential for historical research
- Imagine a world of opensource federated vectorbases maintained by specialists available for interrogation by historians

 Daniel van Strien @danielvanstrien.bsky.social · 1d

I think one of the most important skills for using AI well for any domain, including history, will be building datasets and learning how to do good/useful evaluations for the tasks you are interested in. IMO, this is where collaboration in the field could have a lot of benefits.

 Daniel van Strien @danielvanstrien.bsky.social · 1d

I think the tide is turning on scale above all else but scale is still important so it will be more about ensuring quality at scale. See for example [huggingface.co/spaces/Huggi...](https://huggingface.co/spaces/HuggingFaceFW/fineweb)




Daniel van Strien @danielvanstrien.bsky.social · 1d

↩ Reply to you

I think GLAM also has the advantage of having a lot of data that is at least partly described/curated already.


expt 1: archival agents


Aurelius-Archives

Live
Anyone with a link

Create

Configure



Name

Aurelius-Archives

Description

Helps historians interrogate research resources and metadata provided by archives and libraries worldwide

Instructions

- * BHOL_Subject_Guides_1.0_13072024.txt for guides to specific historical topics addressed in British History Online
- * Discovery_API_1.2_TNA_12072024.txt for insights into use of the Discovery API to interrogate the TNAs online catalogue, including API Help page
- * Draft_TNA_AI_Presentation_1.0_14072024.txt for draft text of Colin Greenstreet's presentation at TNA AI week
- * GBPN_Part_One_Ver_1.0_29062024.csv
- * GBPN_Part_Two_Ver_1.0_29062024.csv
- * GBPN_Part_Three_Ver_1.0_29062024.csv

Conversation starters


What research guides do you have available?


Can you write a script to call the TNA-API?


Can you geolocate TNA records?

Knowledge

If you upload files under Knowledge, conversations with your GPT may include file contents. Files can be downloaded when Code Interpreter is enabled


New_Skill_Set_for_Histori...
Document


TNA_AI_Week_New_Skills...
PDF


Discovery_API_1.2_TNA_1...
Document


Updates pending

...

Share

Update

Preview



Aurelius-Archives


Helps historians interrogate research resources and metadata provided by archives and libraries worldwide

What research guides do you have available?

Can you write a script to call the TNA-API?

Can you geolocate TNA records?

Message Aurelius-Archives



↑

?

6

expt 1: archival agents

Demo built around the TNA API, populated with TNA resource guides, and prompt/response models to geotag and map TNA metadata



expt 1: archival agents

PROMPT/RESPONSE MODELS

PROMPT: Can you write a script to call the TNA-API?

REPOSENSE: Offers to do this for Chancery Court Records 1600-1699 using a script from 'Useful_Py_Scripts_1.12_14072024.txt' in internal knowledge

PROMPT: Can you match and geolocate geographic data in TNA records using an historical parish gazeteer?

RESPONSE: Tell me what series of geographic data you want to match? Then adjusts response according to the data. Offers to do a demo using PROB 11 data for blacksmiths from 1600 to 1699. Attempts to find parish names from column named 'places' (Column AF) in 'PROB_11_Blacksmiths_1600-1699_Ver.1.1_14072024.csv', and to match them to parish names in a search of GBPN_Part_One_Ver_1.0_29062024.csv, GBPN_Part_Two_Ver_1.0_29062024.csv, and GBPN_Part_Three_Ver_1.0_29062024.csv. The column named 'places' is a text string, which contains the place of residence. The Place of residence may simply be 'London' (which is not a parish), or it may contain a parish and county name, like 'Sotby, Lincolnshire', where 'Sotby' is the parish and 'Lincolnshire' is the County, or 'Lindfield, Sussex', where 'Lindfield' is the parish and 'Sussex' is the County.

PROMPT: Can you geolocate TNA records?

RESPONSE: Yes. Let me demonstrate this with Blacksmiths probate records from the PROB 11 series for the period 1600 to 1699. I will use a pre-existing download of the Blacksmith data from the TNA website in my internal knowledge, I will extract parish information (where available from Column I in the Blacksmith dataset) and I will match these parish data to the Great British Place Name dataset which I have in my internal knowledge, which contains geolocation data. I will create a downloadable file of all Blacksmiths, with their names, will descriptions, and the parrish I have matched them to, together with the historic county, historic parish name, and the geolocation latitute and longitude for that parish. If you like, I can also map these data for you with popups for the names of the Blacksmiths, their parish, the geolocation data, and a wikipedia link to any information I can find on the parish. Offers to do a demo using PROB 11 data for blacksmiths from 1600 to 1699. Attempts to find parish names from column named 'places' (Column AF) in 'PROB_11_Blacksmiths_1600-1699_Ver.1.1_14072024.csv', and to match them to parish names in a search of GBPN_Part_One_Ver_1.0_29062024.csv, GBPN_Part_Two_Ver_1.0_29062024.csv, and GBPN_Part_Three_Ver_1.0_29062024.csv. The column named 'places' is a text string, which contains the place of residence. The Place of residence may simply be 'London' (which is not a parish), or it may contain a parish and county name, like 'Sotby, Lincolnshire', where 'Sotby' is the parish and 'Lincolnshire' is the County, or 'Lindfield, Sussex', where 'Lindfield' is the parish and 'Sussex' is the County

expt 1: archival agents

PROMPT/RESPONSE MODELS

PROMPT: Can you write a script to call the TNA-API?

REPOSENSE: Offers to do this for Chancery Court Records 1600-1699 using a script from 'Useful_Py_Scripts_1.12_14072024.txt' in internal knowledge

PROMPT: Can you match and geolocate geographic data in TNA records using an historical parish gazeteer?

RESPONSE: Tell me what series of geographic data you want to match? Then adjusts response according to the data. Offers to do a demo using PROB 11 data for blacksmiths from 1600 to 1699. Attempts to find parish names from column named 'places' (Column AF) in 'PROB_11_Blacksmiths_1600-1699_Ver.1.1_14072024.csv', and to match them to parish names in a search of GBPN_Part_One_Ver_1.0_29062024.csv, GBPN_Part_Two_Ver_1.0_29062024.csv, and GBPN_Part_Three_Ver_1.0_29062024.csv. The column named 'places' is a text string, which contains the place of residence. The Place of residence may simply be 'London' (which is not a parish), or it may contain a parish and county name, like 'Sotby, Lincolnshire', where 'Sotby' is the parish and 'Linconshire' is the County, or 'Lindfield, Sussex', where 'Lindfield' is the parish and 'Sussex' is the County.

PROMPT: Can you geolocate TNA records?

RESPONSE: Yes. Let me demonstrate this with Blacksmiths probate records from the PROB 11 series for the period 1600 to 1699. I will use a pre-existing download of the Blacksmith data from the TNA website in my internal knowledge, I will extract parish information (where available from Column I in the Blacksmith dataset) and I will match these parish data to the Great British Place Name dataset which I have in my internal knowledge, which contains geolocation data. I will create a downloadable file of all Blacksmiths, with their names, will descriptions, and the parrish I have matched them to, together with the historic county, historic parish name, and the geolocation latitute and longitude for that parish. If you like, I can also map these data for you with popups for the names of the Blacksmiths, their parish, the geolocation data, and a wikipedia link to any information I can find on the parish. Offers to do a demo using PROB 11 data for blacksmiths from 1600 to 1699. Attempts to find parish names from column named 'places' (Column AF) in 'PROB_11_Blacksmiths_1600-1699_Ver.1.1_14072024.csv', and to match them to parish names in a search of GBPN_Part_One_Ver_1.0_29062024.csv, GBPN_Part_Two_Ver_1.0_29062024.csv, and GBPN_Part_Three_Ver_1.0_29062024.csv. The column named 'places' is a text string, which contains the place of residence. The Place of residence may simply be 'London' (which is not a parish), or it may contain a parish and county name, like 'Sotby, Lincolnshire', where 'Sotby' is the parish and 'Lincolnshire' is the County, or 'Lindfield, Sussex', where 'Lindfield' is the parish and 'Sussex' is the County

Imagine: taking a photograph on a Samsung phone; doing HTR on the fly in the cloud; extracting the named entities; and entering the linked open data into your personal cloud based archive, while still in the archive

expt 2: hca vectorbase

[link to notebooklm - hca](#)

≡ NotebookLM

Sources ⓘ

Select all sources

HCA_13_39_Full_Volu...

HCA_13_40_Full_Volu...

HCA_13_41_Full_Volu...

HCA_13_42_Full_Volu...

HCA_13_44_Full_Volu...

HCA_13_45_Full_Volu...

HCA_13_46_Full_Volu...

HCA_13_47_Full_Volu...

HCA_13_48_Full_Volu...

HCA_13_49_Full_Volu...

HCA_13_50_Full_Volu...

HCA_13_51_Full_Volu...

HCA_13_52_Full_Volu...

HCA_13_53_Full_Volu...

HCA_13_54_Full_Volu...

High Court of Admiralty Deposition Books: HCA 13/39-HCA 13/79

ⓘ Add note

✓ Select all

📄 Saved Response

Volume statistics

HCA 13/ Volume Statistics
The sources contain information about the volume number, number of folios, lines, size, status, and notes for each HCA 13/ volume. This information can be presented in the following table:

Volume	Folios	Lines	Size	Status	Notes
HCA 13/39	237	13,328	552 KB	In Progress	ff.1r-237r
HCA 13/40	237	13,328	552 KB	In Progress	ff.1r-237r

1 citation

📄 Saved Response

Trade and Law

17th Century Maritime Trade and Law: A Study Guide

Short Answer Questions:

1. What types of information are present in the HCA 13/39 document?

2. What are "Ad" and "nescit deponere" frequently used for in the source excerpts?

📄 Saved Response

Weather events

Weather Events Described in the Sources

The sources frequently mention **storms** and **tempestuous weather** at sea. These events are often cited as the cause of damage to ships and cargo. The intensity of these weather events varies and is often described in relative terms, such as "violent," "extream," "greate," "extraordinary," and "tempestuous." [1-27]

10 citations

📄 Saved Response

Ottoman Turkish ports

Trade with Ottoman Turkish Ports in the HCA Sources

The High Court of Admiralty deposition sources offer a glimpse into the dynamics of trade with Ottoman Turkish ports during this period. They reveal a complex web of commercial interactions involving English merchants, Turkish officials, and other European powers.

10 citations

What was the typical trade route for English ships in the Straits?

What was the primary reason for conflict between English ships and Turkish ships in the Straits?

View Chat

41 sources

Start typing...

→

* Notebook guide

NotebookLM can make mistakes, so double-check it.

10

expt 2: hca vectorbase

HCA_13_39_Full_Volume_Processed_Text_EDITED_Ver.1.2_07072024.txt

Source guide

he also knoweth that the
 Basshawe of Tunes
 and Customer gave
 theire severall testimo
 nials written in tur
 kysh language to the said
 Puntys concerning the
 premisses ffor that he
 was present and sawe
 them sealed with their
 seales in inke as the
 fashion is there

Ad nonum affirmat et deponit dvera esse that the City of
 Tunes is and for many yeares past hath byn a fre [?XXXX]
 and place of publick and fre trade, whither merchants of many
 nations especially french englishe Italion and duch merchants doe
 Commonly frequente, and have free accesse to buy and sell there
 ffor this examine hath almost foure yeares space now past known
 the said place and seene the publick and free trade there

Ad decimum deponit hee knoweth of his owne knowledge that
 sundry turkes and others Inhabitants of Tunes are hreate
 merchants and trade in sundry places subiecte to the greate Turck
 especially in Alexandria Constantinople and Tripoly in
 [?Ciria] and usually buy and sell cotton and cotton yarne flaz
 rice Indico and such other goodes as Alexandria dn Ciria
 yeld, ffor that many englishe shippes have byn freighted and
 laden by them with such goodes of this examines knowledge

Ad undecimum deponit he certerly knoweth that the goodes said

HCA 13/39 f.58v Annotate [MACHINE TRANSCRIBED] [CHECKED]

merchandizes specified in the said schedules are goods which growe

High Court of Admiralty Deposition Books: HCA 13/39-HCA 13/79

Settings

Share



Trade with Ottoman Turkish Ports in the HCA Sources

The High Court of Admiralty deposition sources offer a glimpse into the dynamics of trade with Ottoman Turkish ports during this period. They reveal a complex web of commercial interactions involving English merchants, Turkish officials, and other European powers.

- **Scanderoon**, as previously discussed, emerges as a vital node in this trade network. It serves as a primary source for commodities like cotton wool and galls destined for markets across the Mediterranean and beyond [1](#) [2](#) [3](#) [4](#).
- **Constantinople** (present-day Istanbul) features prominently as a major commercial center and a hub for passenger traffic. English ships engaged in transporting passengers and their goods between Constantinople and other Ottoman ports, including **Tunes**, **Alexandria**, and **Tripoli** [5](#).
- **Smyrna** (present-day Izmir), another key Ottoman port, is mentioned as a location where English ships were subject to temporary detention due to political tensions between the Ottoman Empire and England [6](#).
- **Alexandria**, a major Egyptian port under Ottoman control, emerges as a significant source of goods like cotton, cotton yarn, and drugs. These commodities are traded by Turkish merchants in other Ottoman territories like **Tunes** and **Tripoli** and transported on English ships [7](#) [8](#) [9](#).
- **Tunes**, a port city in North Africa, plays a crucial role in the trade network. The sources mention English merchants operating there and interacting with Turkish officials like **Osman Dory**, the "Captaine of the Jenesires," who exerted considerable control over trade activities [10](#). They also describe the presence of **High Changett**, a consul representing both the French and English nations, indicating the multinational nature of commerce in the region [11](#).

Trade routes connect these Ottoman ports with each other and extend to other parts of the Mediterranean, including Italian cities like **Naples**, **Venice**, and **Genoa**, as well as ports in the Levant like **Sio** (present-day Sitia, Crete) and **Cyprus** [5](#) [7](#) [12](#) [13](#). These interconnected routes facilitated the flow of goods and people across the region.

The sources highlight some of the **challenges** faced by English merchants operating in Ottoman territories. They faced political uncertainties, as exemplified by the temporary detention of English ships in Smyrna and the restrictions imposed by powerful local figures like Osman Dory in Tunes [6](#) [10](#). They also navigated a complex environment of international relations, where they interacted with merchants and officials from other European nations like France and Venice [11](#) [12](#).

The High Court of Admiralty records, through depositions and legal disputes, provide a unique perspective on the intricacies of trade with Ottoman Turkish ports during this era. While this information offers valuable insights, further investigation is needed to create a more comprehensive understanding of the economic and political contexts surrounding this trade.



Save to note

What was the typical trade route for English ships in the Straits?

What was the primary reason for conflict between English ships and Turkish ships in the

Close Chat

41 sources

Start typing...



Notebook guide

11

expt 3: summarization

+ ≡ analytical-ontological-summarization-prompt-17.8-Claude v0 Unsaved

Prompt Evaluate

<instructions>
Instructions:

Carefully read and analyze the entire document.
Extract all relevant information according to the framework below.
When you refer to specific quotes from the text, assign them a number and include the full quoted text in a dedicated section at the end of your analysis.
Organize your analysis into a single CSV file with the following columns:
Section
Subsection
Category
Description
Quote Number (if applicable)
Represent the hierarchical structure of the framework by using the Section and Subsection columns.
For entries that don't fit neatly into the Category or Description columns (e.g., cross-references, summaries), use those columns creatively to convey the information.
Maintain consistent detail level across all categories.
Flag any uncertainties or ambiguities.
Provide confidence levels for all analytical elements in a consolidated table within the CSV.
Conduct analysis meticulously and comprehensively
Give replies for ALL variables specified in the prompt, inserting "not specified"
</instructions> if there is no relevant data

<Analysis Framework>
Analysis Framework:

1. CORE_DOCUMENT_METADATA

- Reference number: {ref_number}
- Case name: {case_name}
- Proctors: {proctors}
- Date of document: {doc_date}
- Document type: {doc_type}
- Physical condition/completeness: {condition}
- Language(s) used: {languages}

2. REQUIRED_ANALYTICAL_ELEMENTS

- Key individual 1 [KI]: {key_individual} {key-individual_occupation} {key-individual-role-in-case} {key_individual_age} {key_individual_residence_parish} {key_individual_residence_town} {key_individual_residence_country}
- Key individual 2 [KI]: {key_individual} {key-individual_occupation} {key-individual-role-in-case} {key_individual_age} {key_individual_residence_parish} {key_individual_residence_town} {key_individual_residence_country}
- Key individual n [where n = number of key individuals] [KI]: {key_individual} {key-individual-role-in-case} {key-individual_occupation} {key_individual_age} {key_individual_residence_parish} {key_individual_residence_town} {key_individual_residence_country}
- Mariner behavior [MB]: {mariner_behavior}
- Language skills and literacy [LS]: {language_skills}
- Signatures/marks analysis [SM]: {signatures}

2.2 Maritime Assets

- Key ships [KS]: {key_ships}
- Vessel specifications [VS]: {vessel_specs}
- Operational status [OS]: {op_status}

Test Case Generate X

deposition-text

Enter an example value...

case_name

Enter an example value...

proctors

Enter an example value...

deposition-date

Enter an example value...

folio-ref

Enter an example value...

special-focus

Enter an example value...

expt 3: summarization

```
ANTHROPIC Dashboard Workbench Settings

+ ≡ analytical-ontological-summarization-prompt-17.8-Claude v0 Unsaved Prompt Evaluate

DETAILED_CATEGORIZATION,Personnel_information,Witness status,"Direct witness as ship's master",34
DETAILED_CATEGORIZATION,Personnel_information,Professional capacity,"Master mariner",35
DETAILED_CATEGORIZATION,Personnel_information,Employment terms,"One voyage only with this vessel",36

DETAILED_CATEGORIZATION,Commercial_transactions,Trade agreements,"Goods consigned to Coymans in Amsterdam",37
DETAILED_CATEGORIZATION,Commercial_transactions,Documentation,"Bills of lading signed in triplicate",38
DETAILED_CATEGORIZATION,Commercial_transactions,Customs compliance,"Goods cleared through customs houses at Santa Cruz and Guarachiche",39

DETAILED_CATEGORIZATION,Maritime_operations,Navigation details,"Teneriffe to Amsterdam route interrupted by captures",40
DETAILED_CATEGORIZATION,Maritime_operations,Operational decisions,"French captors removed part of cargo; Commonwealth forces seized remainder",41

CROSS_REFERENCE_MATRIX,,Cargo-Owner Relations,"Each owner's goods marked distinctly and documented separately",42
CROSS_REFERENCE_MATRIX,,Legal-Commercial Interface,"Spanish ownership claims vs Commonwealth prize jurisdiction",43
CROSS_REFERENCE_MATRIX,,Documentation Chain,"Customs clearance → Bills of lading → Prize court claims",44

CONFIDENCE_LEVELS,,High confidence,"Cargo specifications; Loading details; Witness's direct observations",
CONFIDENCE_LEVELS,,Medium confidence,"Events during French capture; Ownership claims",
CONFIDENCE_LEVELS,,Low confidence,"Details of second capture by Commonwealth forces",
CONFIDENCE_LEVELS,,Speculative elements,"Ultimate fate of cargo and vessel",

SUMMARY_SECTIONS,Executive Summary,"Spanish-owned cargo aboard Dutch-mastered vessel captured first by French privateers then by English Commonwealth forces, leading to ownership claims in English Admiralty Court",
SUMMARY_SECTIONS,Key Findings,"1. Complex international trade network involving Spanish merchants, Dutch shipping, and English/French naval forces; 2. Detailed cargo documentation system; 3. Multiple jurisdictional issues",
SUMMARY_SECTIONS,Historical Context,"Document illustrates complexities of maritime trade during Anglo-Spanish tensions and Anglo-Dutch rivalry",

METADATA,,Analysis version,"Version 17.8",
METADATA,,Date of analysis,"Current date",
METADATA,,Analyst details,"GPT-4",
METADATA,,Sources consulted,"Original HCA deposition",
METADATA,,Review status,"Pending review",

QUOTATIONS:
1. "The clayme of ffrancisco Lopez, John Baptista de ffon[?tey] [?Pontey Paxes] and of Arancia Daranga [?Mederas] Spaniards"
2. "Luytien Hendersen Sinck of hoorne in North-holland Mariner late Master of the said shipp... aged 40 yeares"
3. "ffrancisco Lopez a Spaniard and a Subject of the king of Spaine living at Oratavo in the Island of Teneriffe"
4. "don John Baptista de Ponty ffontey Paxes living at Guarachicha"
5. "Asensio darangina Maderas also a Merchant dwelling at the said Teneriffe"
6. "taken by Captaine Nicholas heaton in the Saphire frigatt"
7. "taken by Peter Collaart under pretence of a Commission from the king of ffrance"
8. "to be delivered to Balthazar and John Coymans Merchants of Amsterdam"
9. "the said shipp the Santa Maria de Guarda Lupa"
10. "One hundred and thirty halfe Chests of sugars and of six hundred and Twenty hides"
11. "marked with the first marke in the margent"
12. "six great ffatts or vessells of sugar... fifty fower hides"
13. "from the Teneriffe aforesaid... to Amsterdam"
14. "their the said producents Agents and Correspondents there"
```


expt 3: summarization

Better performance by GPT40i-preview

Worse performance by GPT40i-preview

GPT4o GPT40i-preview

ChatGPT 4o with canvas

Section	Subsection	Category	Description (OUTPUT A)	Description (OUTPUT B)	Differences
REQUIRED_ANALYTICAL_ELEMENTS	Language skills and literacy	Language skills	Bookkeeping and correspondence in English	Witness is literate; Bills of lading filled in handwriting of Jacob Lucy	OUTPUT B adds specific details related to Jacob Lucy, which OUTPUT A misses.
REQUIRED_ANALYTICAL_ELEMENTS	Signatures/marks analysis	Signatures	Signature of Cornelius van Bommel present, legible	Bills of lading signed by Master Peter de la Sale; Marked with 'WS' and numbers 20-24	OUTPUT B mentions the specific markings on the bills of lading and adds more detail on who signed it.
2.2 Maritime Assets	Key ships	Key ships	John Baptist; Saint Ives Scout; Saint Peter	The John Baptist; The Saint Ives Scout; The Saint Peter; The Charity (Hamburgher ship)	OUTPUT B adds an additional ship, "The Charity."
2.3 Commercial Elements	Financial transactions	Financial transactions	Customs paid for goods in London	Customs paid; Goods cost £228 12s 6d	OUTPUT B adds the specific value of the goods.
2.4 Geographic Framework	Key locations	Key locations	London, Portsmouth, Saint	London; Portsmouth; Saint	OUTPUT B omits Hamburgh,

Comparison of GPT4p and GPT40i-preview for analytical ontological summarization task using same prompt and same inputs

Deposition Analysis Csv

```
1 Section,Subsection,Category,Description,Quote Number
2 CORE_DOCUMENT_METADATA,Reference number,Reference number,HCA 13/70 ff.287v-288v,not specified
3 CORE_DOCUMENT_METADATA,Case name,Case name,The Clayme of Lucas Lucy for goods taken out of the John Baptist by Edward Payne Commander of the Vess
4 CORE_DOCUMENT_METADATA,Proctors,Proctors,Suckley; Smith,not specified
5 CORE_DOCUMENT_METADATA,Date of document,Document date,March 16th 1654/55,not specified
6 CORE_DOCUMENT_METADATA,Document type,Document type,Deposition,not specified
7 CORE_DOCUMENT_METADATA,Completeness,Completeness,Complete but with some illegible portions due to age,not specified
8 CORE_DOCUMENT_METADATA,Languages used,English,not specified
9
10 REQUIRED_ANALYTICAL_ELEMENTS,Key individual 1,Key individual,Cornelius van Bommel,Merchant,Witness,35,Saint Catherine Colemans,London,England
11 REQUIRED_ANALYTICAL_ELEMENTS,Key individual 2,Key individual,Lucas Lucy,Merchant,Claimant,not specified,not specified,not specified,not specified
12 REQUIRED_ANALYTICAL_ELEMENTS,Key individual 3,Key individual,Walter Sepps,Merchant,Consignee,not specified,Hamburgh,not specified,Germany
13 REQUIRED_ANALYTICAL_ELEMENTS,Key individual 4,Key individual,Hugh Salsbury,Agent,Correspondent,not specified,Portsmouth,not specified,England
14 REQUIRED_ANALYTICAL_ELEMENTS,Key individual 5,Key individual,William Jennings,Merchant,Correspondent,not specified,Plymouth,not specified,England
15 REQUIRED_ANALYTICAL_ELEMENTS,Key individual 6,Key individual,Andrew Jennings,Merchant,Correspondent,not specified,Penryn,Cornwall,England
16 REQUIRED_ANALYTICAL_ELEMENTS,Mariner behavior,Mariner behavior,Not specified,not specified
17 REQUIRED_ANALYTICAL_ELEMENTS,Language skills,Bookkeeping and correspondence in English,not specified
18 REQUIRED_ANALYTICAL_ELEMENTS,Signatures,Signature of Cornelius van Bommel present, legible,not specified
19
20 2.2 Maritime Assets,Key ships,Key ships,John Baptist; Saint Ives Scout; Saint Peter,not specified
21 2.2 Maritime Assets,Vessel specifications,Vessel specifications,Not specified,not specified
22 2.2 Maritime Assets,Operational status,Operational status,Trading vessels engaged in merchant voyages,not specified
23
24 2.3 Commercial Elements,Key commodities/materials,Key commodities/materials,Three bales of Perpetuana's; Two bales of Minneken bayes,not specifie
25 2.3 Commercial Elements,Financial transactions,Financial transactions,Customs paid for goods in London, not specified
26 2.3 Commercial Elements,Contractual undertakings,Contractual undertakings,Goods consigned to Walter Sepps for transportation to Saint Sebastians,
27
28 2.4 Geographic Framework,Key locations,Key locations,London, Portsmouth, Saint Sebastians, Penryn, Plymouth, Hamburgh,not specified
29 2.4 Geographic Framework,Route analysis,Route analysis,From London to Saint Sebastians via Portsmouth,not specified
30 2.4 Geographic Framework,Jurisdictional boundaries,Jurisdictional boundaries,English jurisdiction,not specified
31
32 2.5 Temporal Elements,Chronological sequence,Chronological sequence,March to June 1653,not specified
33 2.5 Temporal Elements,Weather events,Weather events,Not specified,not specified
34 2.5 Temporal Elements,Military events,Military events,Seizure by an English private man of war,not specified
35
36 2.6 Documentary Analysis,Key documents,Key documents,Bills of lading; Letters of advise from Hugh Salsbury, William Jennings, Andrew Jennings,not
37 2.6 Documentary Analysis,Period-specific language,Period-specific language,17th-century maritime and trade terminology,not specified
38 2.6 Documentary Analysis,Documentary chain,Documentary chain,Goods consigned to Walter Sepps, then transported to Portsmouth by Hugh Salsbury and
39 2.6 Documentary Analysis,Legal/arbitration questions,Legal/arbitration questions,Rightful ownership of goods seized by an English privateer,not specif
40 2.6 Documentary Analysis,Warrant obtained for the recovery of goods,not specified
41
42 2.7 Legal Framework,Jurisdictional matters,Jurisdictional matters,English jurisdiction over seized goods,not specified
43
44 2.8 Critical Analysis,Discrepancies,Discrepancies,Unclear ownership marking used for preservation of goods, not specified
45 2.8 Critical Analysis,Biases and limitations,Biases and limitations,Possible biases due to involvement of privateers and potential conflicts of i
46 2.8 Critical Analysis,Goods owned by Lucas Lucy were seized during transport; attempts to recover goods involved multip
47
48 3. LOCATION_TAGGING_SYSTEM,Ports,Port arrival,Saint Sebastians,not specified
49 3. LOCATION_TAGGING_SYSTEM,Ports,Port intermediate,Portsmouth, Penryn, Plymouth,not specified
50 3. LOCATION_TAGGING_SYSTEM,Maritime Spaces,Sea area,English Channel,not specified
51 3. LOCATION_TAGGING_SYSTEM,Maritime Spaces,Coastal zone,Coast of Cornwall,not specified
52 3. LOCATION_TAGGING_SYSTEM,Maritime Spaces,River system,Not specified,not specified
53
54
55 4. EVENT_CLASSIFICATION,Maritime Operations>Loading event>Loading of bales onto the Saint Peter in Portsmouth,not specified
```

Provides evidence of literacy

Provides detail of markings on bill of lading

Identifies additional ship

Provides detail of value of good

Misses one geographical location

expt 4: fine-tuning

[marinelives on hugging face](#)
[raw-htr-groundtruth-dataset](#)

- [Hugging Face MarineLives organisation](#)
[Colin Greenstreet; Thiago Krause, Tobias Hodel, David Brown, Sharon Howard]
- Exploring the potential for fine-tuned small LLMs to support the process of cleaning Raw HTR output after the machine transcription of English High Court of Admiralty depositions
- Exploring Retrieval Augmented Generation with small to medium sized LLMs with a Pinecone vectorbase of Engl;ish High Court of Admiralty depositions

The screenshot displays the Hugging Face web interface. At the top, the 'Hugging Face' logo and a search bar are visible. Below the navigation bar, the 'Datasets' section shows the 'MarineLives/raw-htr-handchecked-groundtruth-small' dataset. The dataset card includes a 'Dataset card' tab, a 'Viewer' tab, and a 'Files and versions' tab. The 'Viewer' tab is active, showing a table with two columns: 'Raw-HTR Text' and 'Hand-corrected Groundtruth'. The table contains several rows of text, including 'amongste them the goodes articulate' and 'And this hee saith is true'. To the right of the dataset card, there are statistics: 'Downloads last month: 32', 'Size of downloaded dataset files: 96.6 kB', 'Size of the auto-converted Parquet files: 60.4 kB', and 'Number of rows: 697'. Below the dataset card, the 'Spaces' section shows a Space named 'MarineLives/early-modern-legal-rag' which is currently 'Running'.

expt 4: fine-tuning

EAHistoriChat

C18th

A chatbot designed to respond in the style of Early American written texts. It has limited ability to deal with multi-turn conversations, and has some formatting issues in its answers, due to training on data which retained linebreaks.

A Fine-tuned version of [Mistral-Hermes 2](#), trained on synthetic question-answer pairs to replicate Early American prose. The training dataset consists of curated paragraphs from the [Evans-TCP](#) corpus and uses a quantized version of [Mistral-Nemo-Instruct](#) to generate questions for which these paragraphs serve as appropriate answers. Fine-tuning was conducted using the Axolotl framework on this custom dataset.

The idea came from [Mark L. Thompson](#) at the University of Groningen, with design and implementation by [Michiel van der Ree](#), also at the University of Groningen.

Resources: (1) [GitHub repository](#) (2) [Hugging Face model card](#)

MonadGPT:

C17th

A finetune of [Mistral-Hermes 2](#) on 11,000 early modern texts in English, French and Latin, mostly coming from EEBO and Gallica. It can be used in conversation mode, answering in an historical language and style, and using historical and dated references. The training dataset for MonadGPT consists of 10,797 rows, and is 10,3 M+B in size (auto-converted Parquet files).

Developed by French digital humanities researcher [Pierre-Carl Langlais](#), aka [Alexander Doria](#). [Langlais](#) is co-founder of the French private company [Pleias](#), which focuses on open source large language models.

MacBERTh

MacBERTh and GysBERT are language models (more specifically, BERT models) pre-trained on historical textual material (date range: 1450-1950). MacBERTh is trained on English and GysBERT is trained on Dutch.

Pretrained on ca. 3.9B tokens, drawn from EEBO, ECCO, COHA, CLMET3.1, EVANS, Hansard Corpus.

MacBERTh has been utilized in studies to assess its performance on Early Modern English data. For instance, the paper "How BERT Speaks Shakespearean English? Evaluating Historical Bias in Contextual Language Models" (2024) examines MacBERTh's capabilities in understanding Shakespearean English.

Resources:

(1) [Hugging Face MacBERTh model card and model](#) (2) Manjavacas, Enrique & Lauren Fonteyn. 2022. Adapting vs. Pre-training Language Models for Historical Languages. Journal of Data Mining & Digital Humanities jdmdh:9152. <https://doi.org/10.46298/jdmdh.9152> (3) [Miriam Cuscito, Alfio Ferrara, Martin Ruskov](#), How BERT Speaks Shakespearean English? Evaluating Historical Bias in Contextual Language Models, submitted 7 Feb 2024, <https://doi.org/10.48550/arXiv.2402.05034>

expt 4: fine-tuning



Daniel van Strien @danielvanstrien.bsky.social · 1d

In many cases it's unlikely to be feasible to train/fine-tune an LLM but building robust evaluation datasets can help make it much easier to measure and compare performance of different LLMs (open and closed). There are some useful pointers for doing this here: [github.com/huggingface/...](https://github.com/huggingface/evaluation-guidebook)

huggingface/evaluation-guidebook



Sharing both practical insights and theoretical knowledge about LLM evaluation that we gathered while managing the Open LLM Leaderboard and...

11

Contributors

2

Issues

788

Stars

49

Forks



GitHub - huggingface/evaluation-guidebook: Sharing both practical insights and theoretical knowledge about LLM evaluation that we gathered while managing the Open LLM ...

Sharing both practical insights and theoretical knowledge about LLM evaluation that we gathered while managing the Open LLM Leaderboard...

github.com

rag experiment

Settings

Select President

☒ FDR

☐ Reagan

Clear Conversation

Presidential Time Machine

Welcome to the CHRG-funded project "Speaking with our Sources." This is the project's Presidential Time Machine! This unique application allows you to engage in AI-powered conversations with former US presidents based on their actual speeches and public statements. Currently featuring Franklin D. Roosevelt and Ronald Reagan, the system uses advanced language models and semantic search to provide historically accurate responses in each president's distinctive voice.

Simply select a president, ask your question as if you are a journalist in the White House Briefing Room, and travel through time for a conversation with history! You can scroll down to see the speech excerpts that are retrieved with each query.

Conversation History

Date range ?

Example: January 15, 1940-April 28, 1944

Apply Date Range

Ask a Question

Enter your question:

Send

Configuration:

- Pinecone for vectorstorage
- Voyage AI for embedding and reranking models [with 1024 dimensional embeddings]
- Anthropic Haiku 3.5 for front-end conversational model

expt 5: collaboratory

The goals of the collaboratory are to:

- Build useful **research tools** for real historical research use cases which can be put into immediate effect
- Develop and document **shared knowledge** of large language models applied to historical research use cases
- Build a **community of students and public historians** interested in large language model applications to historical research

Curriculum

Colin Greenstreet edited this page 2 days ago · [5 revisions](#)

[Edit](#)
[New page](#)

Proposed topics for November and December 2024:

** Tuesday, November 26th 2024: [Prompt engineering](#)

** Tuesday, December 3rd 2024: [Summarization](#)

** Tuesday, December 10th 2024: [Named entity extraction](#)

** Tuesday, December 17th 2024: [Metadata creation](#)

Proposed topics for January 2025:

** Tuesday, January 7th 2025: Working with [Colab](#)

** Tuesday, January 14th 2025: [Vectorbases](#) (1): [NotebookLM](#)

** Tuesday, January 21st 2025: [Vectorbases](#) (2): [Tokenization and embeddings](#)

** Tuesday, January 28th 2025: [Vectorbases](#) (3) [Semantic similarity](#)

Possible future topics:

Archival APIs AI-enhanced Dublin Core compliant metadata Archival workflow Building history strategy games Creating fine-tuning datasets Distance reading Historical simulations Hugging Face; Knowledge graphs Integration with Semantic Scholar Linked Open Data Organising personal archives SQL Raw HTR text correction, modernization and summarization; Visualization Scholarly editing in the world of LLMs

The [MarineLives project](#) was founded in 2012. It is a volunteer lead collaboration dedicated to the transcription, enrichment and publication of English High Court of Admiralty depositions.

» Pages **22**

[Home page](#)
[Adoption of technology](#)
[AI assistants and agents. Nov 19, 2024 talk](#)
[Analytical ontological summarization prompt](#)
[Barbary pirate narrative summarization prompt](#)
[Batch processing of raw HTR for clean up and summarization](#)
[Bibliography](#)
[Collaboratory members](#)
[Curriculum](#)
[Early Modern English Language Models](#)
[Glossary](#)
[History prompt library](#)
[History domain training data sets](#)
[Hugging Face](#)
[Introduction to machine learning for historians](#)
[MarineLives and machine transcription](#)
[Metadata](#)
[New skill set for historians? July 19, 2024 talk](#)
[Prompt engineering](#)
[Useful tools](#)

main messages

- Dramatically lowers barriers to use of sophisticated general purpose tools
 - Dramatically speeds up research activities, and encourages experimentation at low financial and personal cost
 - Changes scale, scope and depth of research questions which can be asked and answered
 - Increases the importance of archives and libraries relative to individual historians as sources of innovation
 - Creates an opening for new centres of intellectual and technological energy
 - Increases the importance of large scale high quality well curated data sets with clear licensing terms
- Need to explore and develop AI competency
 - prompt engineering
 - data set creation
 - integration of tools & methods
 - Need to tinker, tinker, tinker
 - with tech
 - with process
 - Need to collaborate across disciplines
 - history; corpus & historical linguistics; languages; psychology; sociology
- Unclear if fine-tuned LLMs are worth developing
 - Are vector bases with frontier LLMs and RAG the way to go?

useful links

Google NotebooksLM

[email colin.greenstreet@gmail.com from a gmail account and I will give you access]

[High Court of Admiralty Deposition Books:
HCA 13/39-HCA 13/79](#)

[LitReview: LLMs in Higher Education](#)

[marinelives-ai-and-history-collaboratory](#)

[Programming Historian](#)

ai-and-history-collaboratory

[ai-and-history-collaboratory-wiki](#)

[Prompt engineering](#) [first session of our
collaboratory, Tuesday, November 26th]

[Machine learning and historical research
bibliography](#)

discussion

backup

AI assistants and agents: A New Skill Set for Historians?

Abstract: A New Skill Set for Historians explores the potential for Large Language Model based intelligent assistants and agents to support historical research. The speaker makes the case for historians and archivists acting together to build knowledgeable, effective, and serendipitous assistants and agents within the history domain, and explores two parallel routes to do so. Firstly, the construction of large language models using retrieval augmented generation techniques to draw on specialized domain specific vectorbases. Secondly, the construction of large scale licence clear datasets of historical manuscript and printed knowledge for fine-tuning of medium-sized large language models.

The speaker looks at two areas of potential impact for large language models on historical research:

- on historians' interactions with archives and on their personal archival research practices
- on the types of research questions historians can ask and answer, enabled by much larger, more complex, and more interlinked datasets and metadata, and supported by sophisticated, flexible, and easy to use analytical tools.

The speaker looks at the history of technology uptake within historical research practice and asks what needs to be done to encourage the widespread adoption and embedding of large language model enabled techniques into research practices. As a contribution to the exploration and adoption of such techniques, the speaker is launching a MarineLives-Collaboratory for doctoral students interested in the application of large language models in their own research design and research practices. This will provide the opportunity to work on [specific historical use cases](#).

The speaker illustrates his broad proposals with his own hands on work at MarineLives:

- * the publication of 6 million words of semantically searchable and summarizable English High Court of Admiralty depositions using Google's vectorbase NotebookLM
- * the creation of a bespoke Pinecone vectorbase using sentence and paragraph embeddings for interrogation by researchers
- * the fine-tuning of a mid-sized multi-lingual large language model to clean up raw machine transcriptions
- * the use of frontier large language models (Claude sonnet 3.5, Gemini1.5, GPT-4o-preview) to perform high grade analytical ontological summarization as part of a pipeline from machine transcription through to open linked data creation.

The talk concludes with a vision of multi-agent/multi-player historical simulations to be integrated into graduate teaching and looks at the structure of such a simulation of international investment in the seventeenth century.

DEBATE FORUM

CONFRONTING THE DIGITAL

OR HOW ACADEMIC HISTORY WRITING LOST THE PLOT

Tim Hitchcock

University of Hertfordshire

ABSTRACT This discussion piece argues that the design and structure of online historical resources and the process of search and discover embodied within them create a series of substantial problems for historians. Algorithm-driven discovery and misleading forms of search, poor OCR, and all the selection biases of a new edition of the Western print archive have changed how we research the past, and the underlying character of the object of study (inherited text). This piece argues that academic historians have largely failed to respond effectively to these challenges and suggests that while they have preserved the form of scholarly good practice, they have ignored important underlying principles.

Keywords: digital humanities, digital history, standards, scholarship, referencing, OCR, search

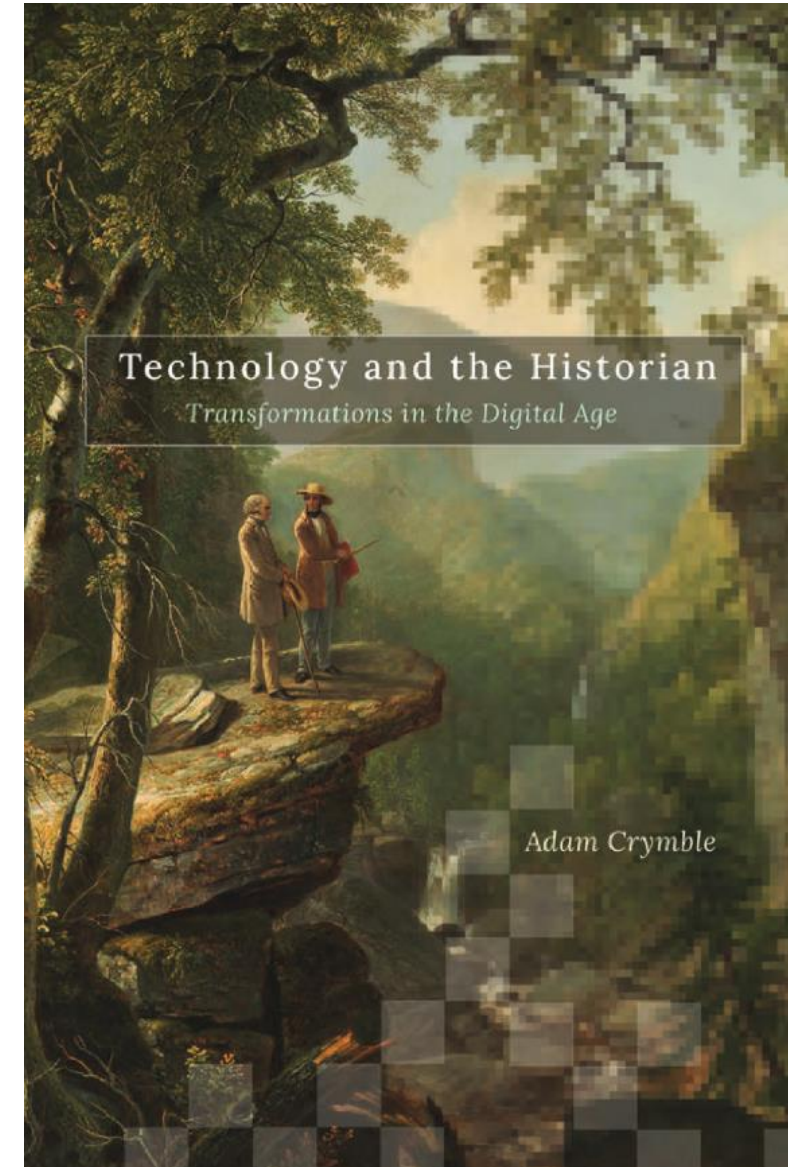
We are halfway through what has frequently been described as a revolution. In the last fifteen years trillions of words of printed text have been digitized and delivered to an eager audience in a keyword searchable form. Google books alone has digitized some seven million of what it estimates are 1.3 billion volumes,¹ and in the next ten years the pre-1900 archive of printed materials in most major languages will be available for keyword searching.² We are witnessing the creation of the Western print archive, second edition. Even now it is possible to research and write credible, evidence-based history on many topics using exclusively online sources. This is not to imply that the process is complete, or even nearly so. We have only just begun to digitize manuscript materials, ephemera, images and objects. There also remain serious issues about what should be digitized next, and what impact that selection has on the direction of scholarship.³ But what has been achieved is nevertheless remarkable. Britain has been at the forefront of the international campaign to make this happen, and its academic community – and the wider community of scholars working on British subjects – has been its greatest beneficiary. Historians working on early modern and nineteenth-century British history in particular have been gifted the most thoroughly digitized period and place in the world. Between Google Books, Early English Books Online (EEBO) and Eighteenth Century Collections Online (ECCO), *The Times* Digital

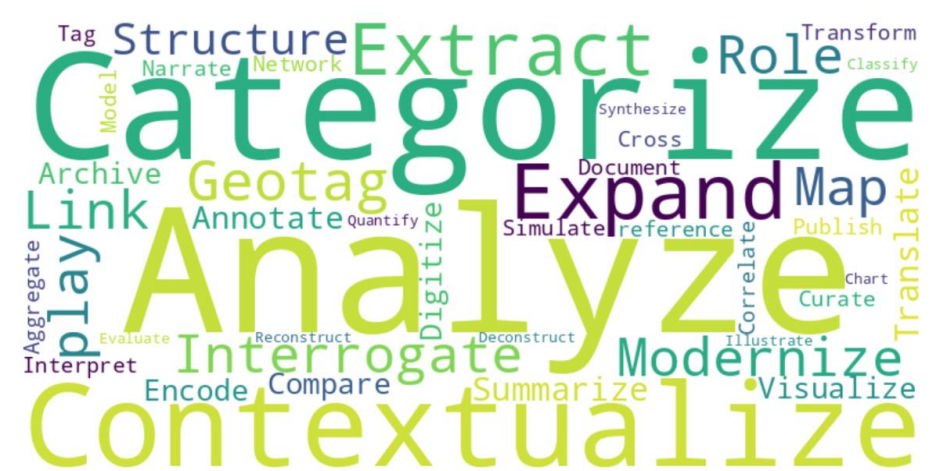
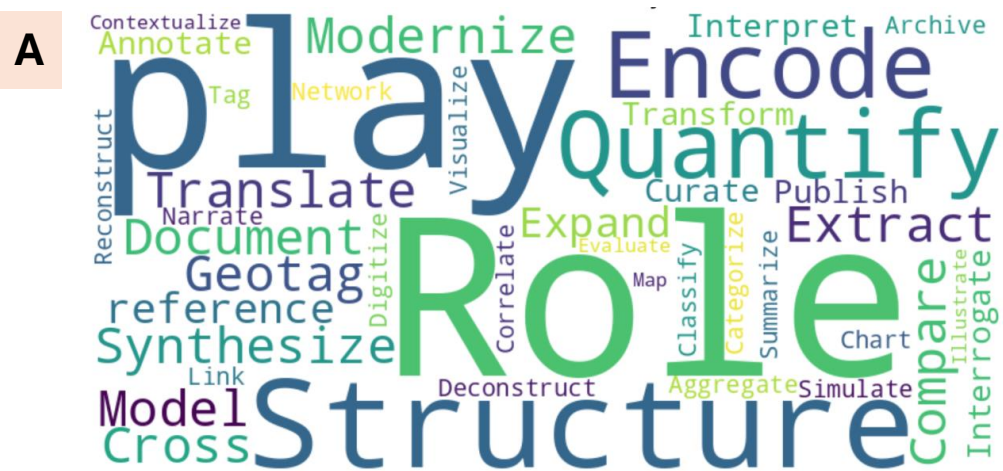
Address for correspondence: Professor Tim Hitchcock, University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB, UK. E-mail: T.Hitchcock@herts.ac.uk

Cultural and Social History, Volume 10, Issue 1, pp. 9–23 © The Social History Society 2013
DOI 10.2752/147800413X13515292098070

HITCHCOCK Confronting the Digital

9





domain specific data

History domain training data sets

Colin Greenstreet edited this page yesterday · [7 revisions](#)

TABLE OF CONTENTS

(alphabetical)

- British History Online (BHOL)
- Common Corpus
- Common Crawl
- Early English Books Online (EEBO-ProQuest)
- Europeana APIs
- Evans-TCP
- Hathitrust
- Historic England APIs
- Huygens Institute
- Internet Archives APIs
- Library of Congress
- National Archives (TNA)
- Text Creation Partnership (TCP I and TCP II)
- Vrije Universiteit Amsterdam















But historical domain data is rather randomly incorporated in current pre-training datasets

experimentation

My Notebooks

+ Create new

✓ [Grid Icon] [List Icon] Title ▾

 High Court of Admiralty Deposition Books: HCA... Oct 4, 2024 · 41 sources	 Intelligent Agents for Historians Oct 3, 2024 · 5 sources	 Introduction to NotebookLM Dec 6, 2023 · 8 sources	 LitReview: LLMs in Higher Education Oct 27, 2024 · 13 sources	 LitReview: MT5 & UMT5 Oct 28, 2024 · 2 sources
 LLM Technology Oct 30, 2024 · 1 source	 Machine learning for historians Oct 27, 2024 · 10 sources	 marinelives-ai-and- history-collaboratory Nov 13, 2024 · 2 sources	 MarineLives: 12 Years of Tweets Nov 13, 2024 · 0 sources	 Mental Health Therapist Support App Oct 28, 2024 · 2 sources
 Programming Historian Oct 29, 2024 · 4 sources	 Student Notebooks Nov 9, 2024 · 1 source	 The Metadata Machine Nov 12, 2024 · 0 sources	 The TRT2: Teaching Recovery Techniques... Nov 2, 2024 · 9 sources	

Prompt engineering

Colin Greenstreet edited this page 9 minutes ago · [4 revisions](#)

[Edit](#)[New page](#)

Typology of prompts

- Information Retrieval Prompts: To extract specific information such as dates, names, places, core concepts
- Text Generation Prompts: To create abstracts, analyses, narratives, summaries, reports
- Code Generation Prompts: To generate code for tasks like data analysis, data linkage, or data visualization.
- Reasoning Prompts: To interrogate and reason about connectivity, causality, sequencing

Diction, syntax, style, and rhetoric

- Prompts are essentially instructions: Just like any form of communication, the way you construct a prompt influences how it's received and interpreted. Clear, concise, and well-structured prompts lead to better results from the AI.
- Diction matters: The specific words you choose in a prompt can significantly impact the AI's response. For example, asking the AI to "describe" an event will yield a different result than asking it to "analyze" or "evaluate" it. Historians are trained to be sensitive to the nuances of language, and this skill is crucial in prompt engineering.
- Syntax shapes the response: The grammatical structure of your prompt guides the AI's understanding. Using complete sentences, proper punctuation, and clear phrasing helps the AI grasp the intended meaning and generate a more coherent and relevant response.
- Style influences the output: Just as there are different styles of historical writing, there are different styles of prompting. A formal and precise prompt might be appropriate for factual information retrieval, while a more creative and open-ended prompt might be better for generating imaginative narratives.
- Rhetoric adds layers of meaning: Rhetorical devices like metaphors, analogies, and rhetorical questions can be used in prompts to guide the AI's reasoning and elicit more nuanced responses.

Anthropic prompt engineering guide

[Prompt generator](#)

[Be clear and direct](#)

► Pages **13**

[Home page](#)

[AI assistants and agents. Nov 19, 2024 talk](#)

[Batch processing of raw HTR for clean up and summarization](#)

[Bibliography](#)

[Collaboratory members](#)

[Curriculum](#)

[Glossary](#)

[Introduction to Machine Learning for Historians](#)

[MarineLives and machine transcription](#)

[New Skill Set for Historians? July 19, 2024 talk](#)

[Prompt engineering](#)

Clone this wiki locally

<https://github.com/Addaci/marinelive>



retrieval augmented generation

Spaces | MarineLives/early-modern-legal-rag | like 0 | Running | Logs

App | Files | Community | Settings |

Chatbot

What is retrieval augmented generation and why is it an interesting approach for a domain which has specialized knowledge?

Retrieval Augmented Generation (RAG) is a natural language generation technique that combines text retrieval and language generation to produce more accurate and relevant responses. In RAG, the chatbot first retrieves relevant information from a knowledge base or external sources, and then generates a response by synthesizing the retrieved information using natural language generation algorithms.

This approach is particularly interesting for domains that have specialized knowledge because it enables the chatbot to provide more accurate and detailed responses based on the specific context and information available. Unlike traditional natural language generation techniques that rely solely on training data to generate responses, RAG allows the chatbot to access a much larger corpus of information, which can significantly improve the quality and relevance of the generated responses.

For example, in the healthcare domain, a RAG-based chatbot can provide more detailed and accurate responses about specific medical conditions, treatments, and symptoms by retrieving information from medical journals, clinical studies, and other authoritative sources. This can be particularly helpful for patients who have specific questions or concerns, as the chatbot can provide more personalized and detailed responses based on their unique circumstances.

Overall, RAG is an interesting approach for a domain with specialized knowledge because it enables the chatbot to provide more accurate and detailed responses by leveraging the power of text retrieval and natural language generation.


Additional Inputs

dataset building task evaluation

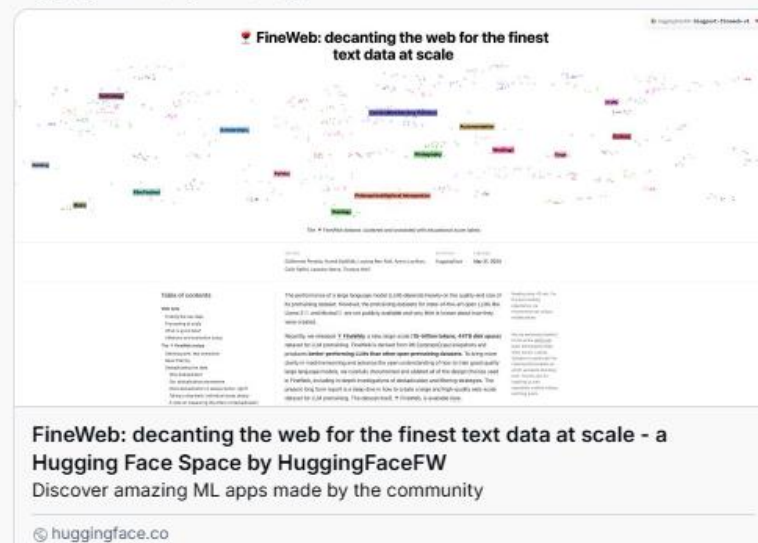
 **Daniel van Strien** @danielvanstrien.bsky.social · 16h

I think one of the most important skills for using AI well for any domain, including history, will be building datasets and learning how to do good/useful evaluations for the tasks you are interested in. IMO, this is where collaboration in the field could have a lot of benefits.

quality at scale

 **Daniel van Strien** @danielvanstrien.bsky.social · 16h

I think the tide is turning on scale above all else but scale is still important so it will be more about ensuring quality at scale. See for example [huggingface.co/spaces/Huggi...](https://huggingface.co/spaces/HuggingFaceFW/fineweb)



robust evaluation benchmarks

 **Daniel van Strien** @danielvanstrien.bsky.social · 16h

In many cases it's unlikely to be feasible to train/fine-tune an LLM but building robust evaluation datasets can help make it much easier to measure and compare performance of different LLMs (open and closed). There are some useful pointers for doing this here: [github.com/huggingface/...](https://github.com/huggingface/evaluation-guidebook)

huggingface/evaluation-guidebook



Sharing both practical insights and theoretical knowledge about LLM evaluation that we gathered while managing the Open LLM Leaderboard and...

 11

Contributors

 2

Issues

 788

Stars

 49

Forks



GitHub - huggingface/evaluation-guidebook: Sharing both practical insights and theoretical knowledge about LLM evaluation that we gathered while managing the Open LLM ...

Sharing both practical insights and theoretical knowledge about LLM evaluation that we gathered while managing the Open LLM Leaderboard...

github.com

LLM + web

Please give me some background on these academics The IHR Digital History seminar conveners - Alexandra Ortolja-Baird (Sheffield), Alice Kinghorn, Jack Newman (Antwerp), James Baker (Southampton), Sarah Middle (Science Museum/Open University), and Tessa Hauswedell (UCL).



The Institute of Historical Research (IHR) Digital History Seminar is convened by a group of scholars with diverse expertise in digital history and humanities:

Dr. Alexandra Ortolja-Baird

Dr. Ortolja-Baird is a Lecturer in History and Digital Humanities at the University of Sheffield's Digital Humanities Institute. Her research focuses on early modern European intellectual history, digital humanities, and cultural heritage studies, particularly concerning Italy, Britain, and the Atlantic world during the seventeenth and eighteenth centuries. She employs digital tools such as network analysis and text mining to enhance historical research and is committed to critical digital humanities approaches. UNIVERSITY OF SHEFFIELD

Dr. Alice Kinghorn

Dr. Kinghorn completed her PhD in History at the University of Bristol in 2024, examining the Church of England's involvement in transatlantic slavery through Anglican missionary societies in the early nineteenth-century Caribbean. She has collaborated on research projects related to institutional involvement in slavery and has published articles on these topics. HISTORIANS AT BRISTOL

Dr. Jack Newman

Dr. Newman is a historian specializing in digital history. He is affiliated with the University of Antwerp, where he contributes to research and teaching in the field of digital humanities.

Dr. James Baker

Dr. Baker is the Director of Digital Humanities at the University of Southampton. His work encompasses digital history, digital humanities, and the application of computational methods to historical research. He has a background in both history and computer science, facilitating interdisciplinary approaches to historical studies. UNIVERSITY OF SOUTHAMPTON

Dr. Sarah Middle

Dr. Middle is associated with the Science Museum and The Open University. Her research interests include digital humanities, metadata, and the accessibility of digital collections. She focuses on enhancing the discoverability and usability of digital resources in cultural heritage institutions.

Dr. Tessa Hauswedell

Dr. Hauswedell is affiliated with University College London (UCL). Her research interests lie in digital humanities, European cultural history, and the development of digital resources for historical research. She has been involved in projects that explore the intersection of digital technologies and historical studies.

These conveners bring a wealth of experience and diverse perspectives to the IHR Digital History Seminar, fostering discussions on the integration of digital methodologies in historical research.

Sources



Message ChatGPT



Ready for a masters?

Dr Alexandra Ortolja-Baird - The University of Sheffield

Dr Alexandra Ortolja-Baird School of History, Philosophy and Digital Humanities. Lecturer in...

Historians at Bristol

PhDone! Dr Alice Kinghorn | Historians at Bristol

December 18, 2023 — Dr Alice Kinghorn. Posted on 18 December 2023 by administrator. Dr Alice...

University of Southampton

Research staff | Humanities | University of Southampton

Contact us +44(0)23 8059 5000 +44(0)23 8059 3131
Address University of Southampton

Search Results

Historians at Bristol

PhDone! Dr Alice Kinghorn | Historians at Bristol

December 18, 2023 — Dr Alice Kinghorn. Posted on 18 December 2023 by administrator. Dr Alice...

Arts Matter

How a PGR Internship Prepared Me for Publishing a Co-Authored Article

August 17, 2023 — Alice Kinghorn is a PhD History candidate with research interests in Anglican...

Arts Matter

bristol – Arts Matter - University of Bristol

August 17, 2023 — Alice Kinghorn is a PhD History candidate with research interests in Anglican...

Oxford Academic

Encoding the haunting of an object catalogue: on the potential of ...

October 19, 2021 — The project sought to contribute to ongoing conversations in historical,...

UCL Blogs

New article 'Encoding the haunting of an object catalogue' - UCLDH Blog

October 19, 2021 — Alexandra Ortolja-Baird and Julianne Nyhan have co-authored a new article on...

Institute of Intellectual History

Cameralism and the Enlightenment — Institute of Intellectual History

May 27, 2020 — In this episode, Alexandra Ortolja-Baird speaks with Dr Ere Nokkala about his...

Ready for a masters?

Dr Alexandra Ortolja-Baird - The University of Sheffield

Dr Alexandra Ortolja-Baird School of History, Philosophy and Digital Humanities. Lecturer in...

Ready for a masters?

Digital Humanities Institute Staff - The University of Sheffield

Dr Alexandra Ortolja-Baird Lecturer in History and

process

Q: I need a daily report on key sales activities—covering tasks completed, events held, leads generated, and the status of opportunities.

Text-to-SQL Workflow Environment

Database

Diverse System



Real Apps



Complex Schema

Table 1:	activity_date	account_id	...	type
...
Table 495:	opportunity_id	stage_name	...	amount

(8695 columns omitted)

Documents



Database Metadata
External Knowledge
SQL Dialect Docs
Query Interface

Codebase

macros/ models/
data/ schema.yml
project.yml leads.sql
salesforce.db ... +20 files



Language Models

SQL

(e.g. Google/Snowflake SQLs ...)

```
WITH opportunity AS (  
  SELECT {{ dbt.date_trunc('day','close_date') }}  
         AS close_date,  
  CASE  
    WHEN is_won THEN 'Won'  
    WHEN NOT is_won AND is_closed THEN 'Lost'  
    WHEN NOT is_closed AND LOWER(forecast_category)  
  IN ('pipeline','best case') THEN 'Pipeline'  
    END AS status  
  FROM {{ var('opportunity') }}  
)  
... [+ 100 lines omitted]  
LEFT JOIN event on ds.date_day =  
salesforce_event.activity_date  
LEFT JOIN opportunities_created  
ON ds.date_day = opportunities_created.created_date  
LEFT JOIN opportunities_closed  
ON ds.date_day = opportunities_closed.close_date
```

SQLs /
Python

Exec
feedback

understanding



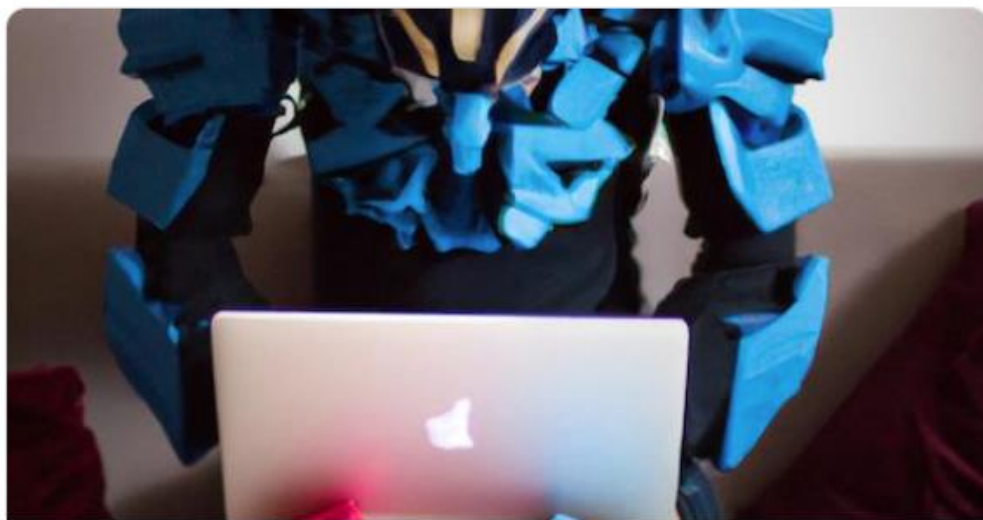
Mat Miller

@matdmiller.bsky.social

In my blog post, I tackle [@karpathy.bsky.social](#) 'Let's build GPT' to construct a transformer 🤖 model from scratch, leveraging the foundational knowledge I gained from the fantastic [@fastdotai](#) courses by [@jph.bsky.social](#)

Curious about how transformers work?

[blog.matdmiller.com/posts/2023-0...](#)




Mat's Blog - Transformers From Scratch

[blog.matdmiller.com](#)


November 14, 2024 at 4:58 AM 🗨️ Everybody can reply

Mat Miller: "In my blog post, I" × notebook.ipynb - Colab × Gemini ×

← → ↺ 🏠 🔍 🗒️ 🗑️ ⭐ 📄 🖼️ 🎵 📺 2m

 Explain the variables in this snippet ^

```
#| output: true
vocab = sorted(list(set(text)))
vocab_size = len(vocab)
print('Vocab Size:', vocab_size)
print('Vocab:', vocab)
```

 This code snippet is all about creating the vocabulary for your transformer model. Let's break down the variables:

- `text` : This variable should already hold the entire text content of 'input.txt', which you loaded in the previous code snippet. It's a string containing all of Shakespeare's works.
- `set(text)` : This creates a set from the `text` string. Sets in Python are collections of unique elements, so this effectively removes all duplicate characters from the text. This gives you every unique character used in Shakespeare's works.
- `list(set(text))` : This converts the set back into a list. This is done because sets are unordered, and we want a specific order for our vocabulary.
- `sorted(list(set(text)))` : This sorts the list of unique characters in alphabetical order. This creates an ordered vocabulary where each character has a specific position or index.
- `vocab` : This variable stores the final sorted list of unique characters, representing the vocabulary of your transformer model. This means your model will understand and work with these specific characters.
- `vocab_size = len(vocab)` : This line calculates the number of unique characters in the `vocab` list and stores it in the `vocab_size` variable. This tells you the size of your vocabulary.


In simple terms, this code creates an ordered list of all the unique characters present in Shakespeare's works. This list acts as the vocabulary for your transformer model. The `vocab_size` variable stores the number of unique characters in this vocabulary.

Understanding the vocabulary is crucial because your transformer will use it to convert text into numerical representations and vice versa. Each character in the vocabulary will be assigned a unique index, which will be used by the model during training and inference.

[nature](#) > [scientific reports](#) > [articles](#) > article

Article | [Open access](#) | Published: 14 November 2024

AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably

[Brian Porter](#)  & [Edouard Machery](#)

[Scientific Reports](#) **14**, Article number: 26133 (2024) | [Cite this article](#)

[Metrics](#)

Abstract

As AI-generated text continues to evolve, distinguishing it from human-authored content has become increasingly difficult. This study examined whether non-expert readers could reliably differentiate between AI-generated poems and those written by well-known human poets. We conducted two experiments with non-expert poetry readers and found that participants performed below chance levels in identifying AI-generated poems (46.6% accuracy, $\chi^2(1, N = 16,340) = 75.13, p < 0.0001$). Notably, participants were more likely to judge AI-generated poems as human-authored than actual human-authored poems ($\chi^2(2, N = 16,340) = 247.04, p < 0.0001$). We found that AI-generated poems were rated more favorably in qualities such as rhythm and beauty, and that this contributed to their mistaken identification as human-authored. Our findings suggest that participants employed shared yet flawed heuristics to differentiate AI from human poetry: the simplicity of AI-generated poems may be easier for non-experts to understand, leading them to prefer AI-generated poetry and misinterpret the complexity of human poems as incoherence generated by AI.