

# Digital Tools for Humanists, Working with AI Summer School 2025 University of Pisa

## AI Meets the Archive: Refining Generative Tools for Historical Research

13 to 27 June 2025 (26 June 2025)



Dr Seamus Ross,  
Professor, Faculty of Information, University of Toronto

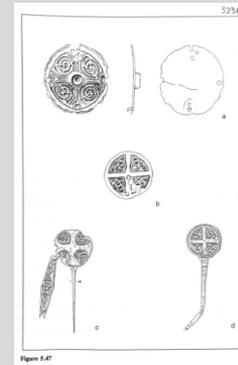
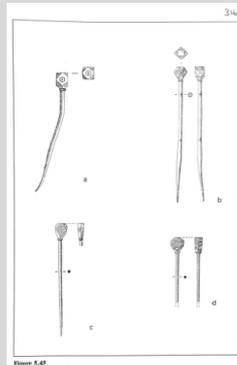
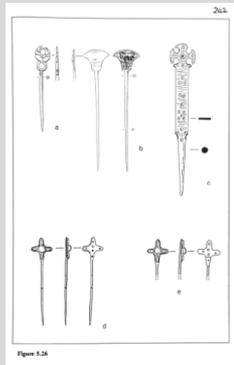
(NOTE: All Audio, Video, Images, etc used is derived from Europeana.)

# Overview of the class

- Welcome and Introduction
- Who am I
- Overview of the day
  - ❖ Lecture
  - ❖ Experimentation
  - ❖ Discussion
  - ❖ Timetable & What we will cover
  - ❖ 09:00 – 10:30 Lecture followed by Activity
  - ❖ 10:30 – 11:00 Break
  - ❖ 11:00 – 12:30 Complete Activity & Continue Lecture
  - ❖ 12:30 – 14:00 Lunch
  - ❖ 14:00 – 15:30 Discussion
  - ❖ 15:30 – 16:00 Break
  - ❖ 16:00 – 17:00 Activity
  - ❖ 17:00 – 17:30 Discussion

# Overview: Who am I and what is the focus of this session

- Research focuses on Digital Humanities and Curation
- Doctoral Studies in Archaeology
  - Studied Anglo-Saxon Dress fasteners (pins) in the 1980s
  - Read hundreds of excavation reports
  - Visited over 100 museums and archaeological research units
  - Used Databases and Expert Systems
  - Thesis:  
<https://ora.ox.ac.uk/objects/uuid:3976b772-fccd-41fe-b8c7-f4ae08ac0295>



Collar Form: incised lines creating the appearance of a ring, Shaft Form: swollen, Shaft Section at Level 1: round, Shaft Section at Level 2: round (?possibly faceted below this point); Degree of Completeness: complete, Context: IAS 4801 0973; Type of Material: copper alloy, Main Manufacturing Process: cast, Secondary Work (if noted): incised lines for collar; Length: 54 mm, Head Diameter/Width: 5x5 mm, Shaft Diameter at Position 1: 1.95 mm, Shaft Diameter at Position 2: 2.10 mm, Shaft Diameter at Position 3: 1.95 mm; Head Decoration: none, Shaft Decoration: none; Reference: Wade & Hatten pers comm.

Group: LXIX, Sub-group: ii, Variant: a, Sub-Variant: 2.B; Site: Southampton, Hants; Museum or Unit: Southampton Archaeological Unit, Accession or Small Find Number: SOU 24 014; Head Form: hot-air-balloon headed, Collar Form: ring, Shaft Form: swollen, Shaft Section at Level 1: round, Shaft Section at Level 2: round; Context: F7023; Degree of Completeness: very tip lost, Type of Material: copper alloy, Main Manufacturing Process: cast; Length: 50 mm, Head Diameter/Width: 5 mm; Head Decoration: wrythen, Decoration Application Method: incised or cast, Shaft Decoration: none; Reference: Brisbane pers comm.

Group: LXIX, Sub-group: ii, Variant: a, Sub-Variant: 2.B; Site: Southampton, Hants; Museum or Unit: Southampton Archaeological Unit, Accession or Small Find Number: SOU 24 003; Head Form: hot-air-balloon headed, Collar Form: ring, Shaft Form: swollen, Shaft Section at Level 1: round, Shaft Section at Level 2: round; Context: F7022; Degree of Completeness: very tip lost, Type of Material: copper alloy, Main Manufacturing Process: cast, Secondary Work (if noted): none; Length: 53 mm, Head Diameter/Width: 7 mm; Head Decoration: none, Shaft Decoration: none; Reference: Brisbane pers comm.

Group: LXIX, Sub-group: ii, Variant: a, Sub-Variant: 2.C; Site: Brandon, Suffolk; Museum or Unit: Suffolk Archaeological Unit, Accession or Small Find Number: 3390 (?); Head Form: hot-air-balloon, Collar Form: ring collar, Shaft Form: hipped, Shaft Section at Level 1: round, Shaft Section at Level 2: round; Degree of Completeness: incomplete, Type of Material: copper alloy, Main Manufacturing Process: cast, Secondary Work (if noted): filed; Length: 65 mm, Head Diameter/Width: 4 mm, Shaft Diameter at Position 1: 2 mm; Head Decoration: none, Shaft Decoration: none; Reference: Carr pers comm.

Group: LXIX, Sub-group: ii, Variant: a, Sub-Variant: 2.C; Site: Brandon, Suffolk; Museum or Unit: Suffolk Archaeological Unit, Accession or Small Find Number: 2294 (?); Head Form: hot-air-balloon-headed, Collar Form: ring collar, Shaft Form: hipped, Shaft Section at Level 1: round, Shaft Section at Level 2: faceted; Degree of Completeness: point lost, Type of Material: copper alloy, Main Manufacturing Process: cast, Secondary Work (if noted): filed; Length: 44 mm, Shaft Diameter at Position 1: 2 mm; Head Decoration: none, Shaft Decoration: none; Reference: Carr pers comm.

Group: LXIX, Sub-group: ii, Variant: a, Sub-Variant: 2.C; Site: Brandon, Suffolk; Museum or Unit: Suffolk Archaeological Unit, Accession or Small Find Number: 2308 (?); Head Form: hot-air-balloon-headed, Collar Form: double ring collar, Shaft Form: hipped, Shaft Section at Level 1: round, Shaft Section at Level 2: square; Degree of Completeness: point lost, Type of Material: copper alloy, Main Manufacturing Process: cast, Secondary Work (if noted): filed; Length: 63 mm, Head Diameter/Width: 6 mm, Shaft Diameter at Position 1: 2 mm; Head Decoration: none, Shaft Decoration: none; Reference: Carr pers comm.

Group: LXIX, Sub-group: ii, Variant: a, Sub-Variant: 2.C; Site: Southampton, Hants; Museum or Unit: Southampton Archaeological Unit, Accession or Small Find Number: SOU 169 1; Head Form: hot-air-balloon, Collar Form: ring collar, Shaft Form: swollen, Shaft Section at Level 1: round, Shaft Section at Level 2: round; Context: 8419(T1); Degree of Completeness: point lost, Type of Material: copper alloy, Main Manufacturing Process: cast; Length: 52 mm, Head Diameter/Width: 6 mm; Head Decoration: wrythen, Decoration Application Method: cast, Shaft Decoration: none; Reference: Brisbane pers comm.

Group: LXIX, Sub-group: ii, Variant: a, Sub-Variant: 2.C; Site: Southampton, Hants; Museum or Unit: Southampton Archaeological Unit, Accession or Small Find Number: SOU 169 213; Head Form: hot-air-balloon headed, Collar Form: ring, Shaft Form: hipped, Shaft Section at Level 1: round, Shaft Section at Level 2: round; Degree of Completeness: complete, Type of Material: copper alloy, Main Manufacturing Process: cast, Secondary Work (if noted): none; Length: 68 mm, Head Diameter/Width: 8 mm, Head

```

        subtype1 = unavailable.
    endif.
Castle Dyke wedge_headed type LVI rule:
    if ident_rate = probable | maybe
    and Metal Type = copper alloy
    and Metal Manufacturing Process = forged
    and Head Type = flat plate
    and Flat Plate Head Form = wedge
    and Collar Type = no_collars
    and Length lt 70
    and Shaft Form = plain tapering
    then Type = Castle Dyke wedge_headed type LVI.
        subtype1 = unavailable.
    endif.

Sibetswold bird_headed type LVII rule:
    if ident_rate = probable | maybe
    and Metal Type = copper alloy | silver
    and Metal Manufacturing Process = cast
    and Head Type = flat plate
    and Zoomorphic Head = present
    and Zoomorphic Head Form = bird_head
    and Collar Type # faceted knob
    and Head Decoration Application Method = cast
    and Cast Decoration Type = animal ornament
    then Type = Sibetswold bird_headed type LVII.
        subtype1 = unavailable.
    endif.
Cross_headed type LVIII rule:
    if ident_rate = probable | maybe
    and Metal Type = copper alloy | silver
    and Metal Treatment = gilt | none
    and Metal Manufacturing Process = cast | forged
    and Head Type = flat plate
    and Flat Plate Head Form = cruciform

```

```

then run h65d.
endif.
if Subtype = Sibton globular_headed subtype LXXXIX v
then run h65e.
endif.
if Type = Ely spatulate lozenge-headed type LXXX
then run h66.
endif.
if Type = Watch_winder type LXXXI
then run h67.
endif.
if Type = Bramble_headed type LXXXII
then run h68.
endif.
if Type = Swing_ringed type LXXXIII
and status(Subtype) = unknown
then run h69.
endif.
if Subtype = Spiral_ringed subtype
then run h69a.
endif.
if Subtype = Ringed_crutch_headed subtype
then run h69b.
endif.
if Subtype = Polyhedral_headed subtype
then run h69c.
endif.
if Subtype = Loop_headed_plain_ringed subtype
then run h69d.
endif.

```

750

How complete is this pin?

1. complete
2. more than seventy five percent complete
3. more than fifty percent complete
4. less than fifty percent complete

=? 1

What type of metal is the pin made from? (s)

1. Copper Alloy
2. Iron
3. Silver
4. Gold

=? 1

Examining the pin which process involved in manufacturing the pin (the primary process only)?

1. cast
2. composite
3. forged or hammered
4. folded Or rolled
5. strip drawn
6. undetermined
7. undeterminable

(Multiple answers allowed)

=? 1

What type of head does this pin have?

1. Knob head (s)
2. Flat plate head (s)
3. Ring passes through head (s)
4. Ring attached to head swings back and forth
5. Select this option if the head is another form, say moulded

=? 1

Which term best describes the primary geometric shape of the head?

1. faceted
2. disc
3. cruciform
4. hot air balloon like

Which of the following items were used to make this pin

1. beaded wire
2. garnets
3. glass
4. glass bead
5. chain
6. gold foil
7. amethyst

(Multiple answers allowed)

=? 1

Does the pin have an attachment mechanism, such as a hole or loop?

1. present
2. absent

=? 2

Is the head of the pin decorated

1. ring
2. dot
3. Ring-and-Dot
4. wrythen
5. spirals
6. filigree
7. animal ornament
8. plant Or vine ornament
9. interlace

# Compare Claude, ChatGPT 4, DeepSeek R1

## ❖ Clarity

- ❖ Does the model understand and respond well to the prompt?

## ❖ Depth

- ❖ Does one prompt yield more nuanced or thoughtful answers?

## ❖ Misinterpretation

- ❖ Does the prompt lead to hallucinations or errors?

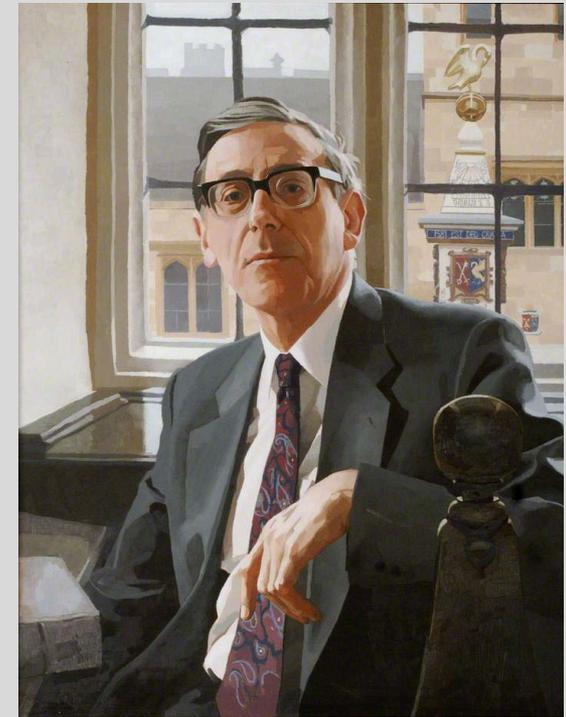
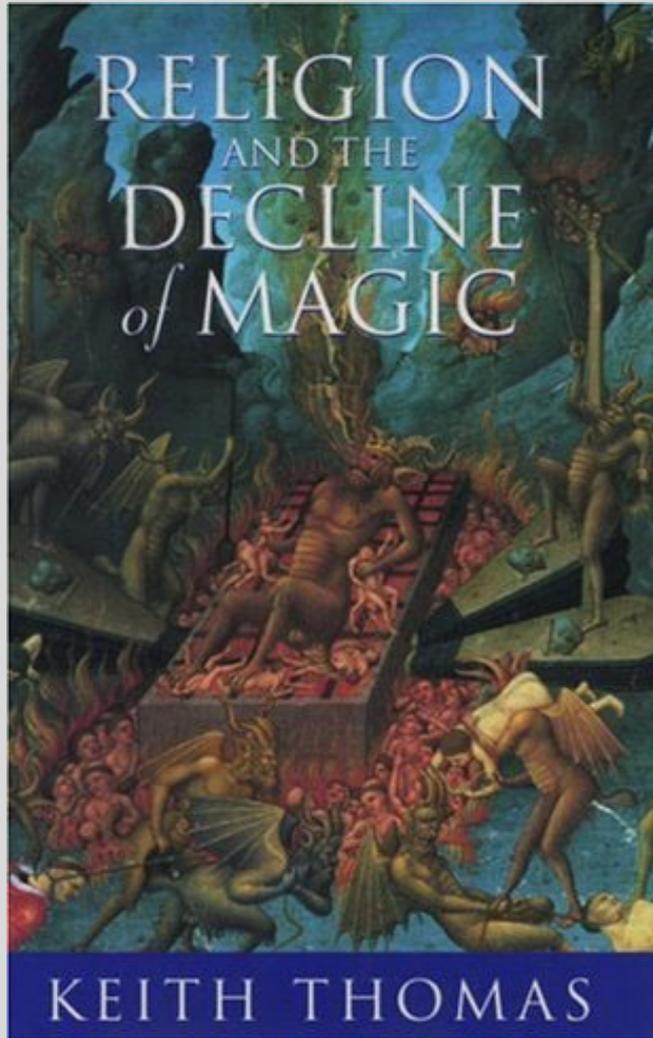
## ❖ Tone of Response

- ❖ Is the output more factual, speculative, literary, etc.?

# Comparing Claude, ChatGPT 4, DeepSeek R1

- ❖ Is the LLM the problem or is the prompt the problem?
- ❖ Is the way the source material is presented a problem?
- ❖ Was it a pre-processing issue?
  - ❖ Was the text poor quality because the scan was inadequate?
  - ❖ Was the problem with the OCRing of the text?

# Doing Humanities Scholarship



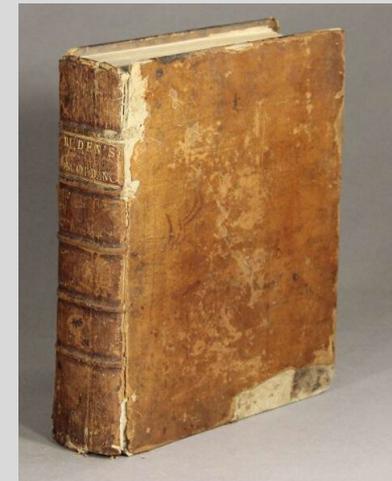
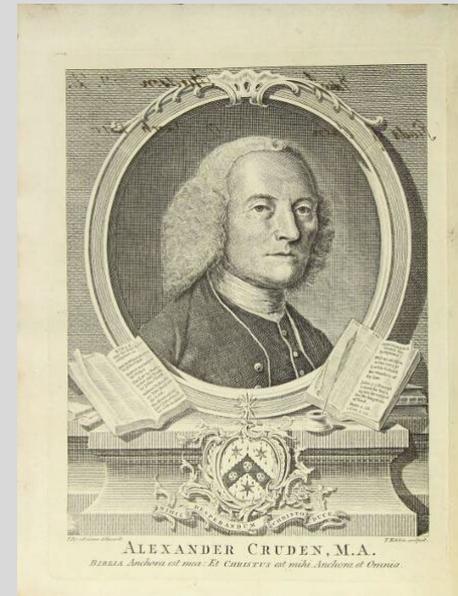
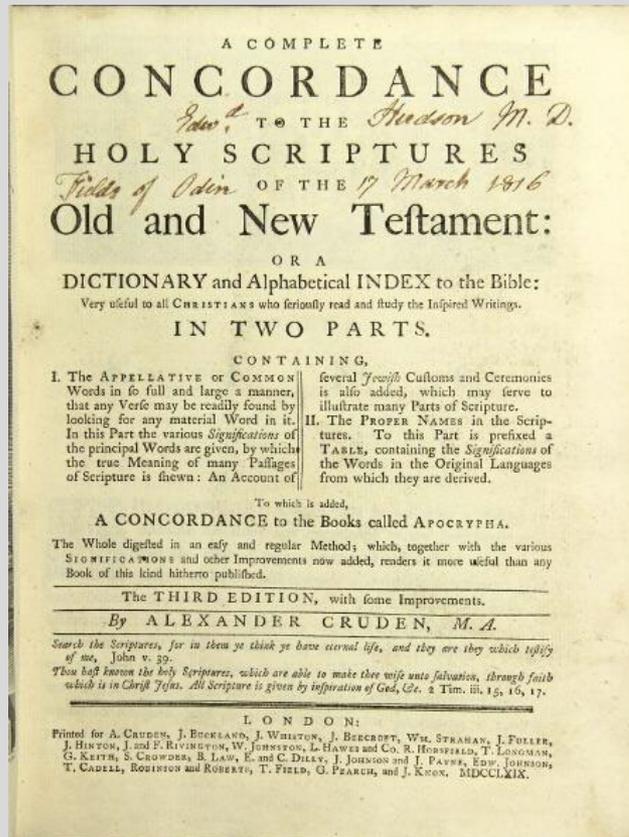
Keith Thomas, [\*Religion and the Decline of Magic\*](#), (Scribner, 1971)

# Impact of the digital

- ❖ Digital technologies change the nature of humanities scholarship.
- ❖ Creates new opportunities.
- ❖ But it also redefines what we think scholarship in the humanities is
- ❖ And it redefines approaches to dissemination
- ❖ Looking mainly at humanities as seen through the vantage of analogue resources

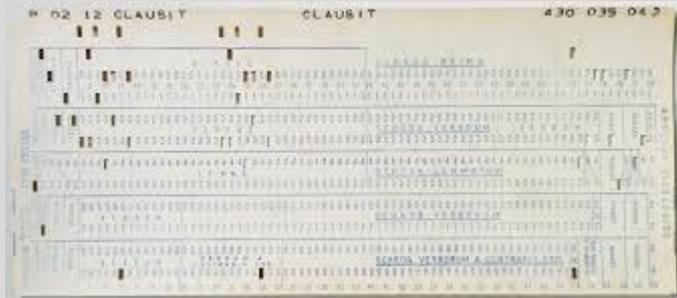
# The Concordance

- ❖ Alexander Cruden (1699 – 1770)
- ❖ *A Complete Concordance to the Holy Scriptures* (1737, 1761, and 1769)



# DH and Concordance Scholarship

- ❖ Father Roberto Busa (1913-2011)
- ❖ [Index Thomisticus](#) (30 years in the making) -- 56 volumes by 1980, CDROM (1989), web (2005)
- ❖ <http://www.corpusthomisticum.org/it/index.age>



Images from: <https://alchetron.com/Roberto-Busa>

CORPUS THOMISTICUM  
**INDEX THOMISTICUS**  
by Roberto Busa SJ and associates  
web edition by Eduardo Bernot and Enrique Alarcón  
English version

Search:

[concordances](#) [terms](#) [works](#) [options](#) [new search](#)

**FOUND 4439 CASES IN 2506 PLACES**



**CASE 1. PLACE 1. Super Sent., lib. 1 q. 1 pr. [...]**<sup>-132</sup> Tertio ex ordinatione modi procedendi, ibi: *ut autem quod quaeritur facilius occurrat, titulos quibus singulorum librorum capitula distinguuntur, praemisimus. Ad evidentiam hujus sacrae doctrinae, quae in hoc libro traditur, quaeruntur quinque: 1 de necessitate ipsius; 2 supposito quod sit necessaria, an sit una, vel plures; 3 si sit una, an practica, vel speculativa: et si speculativa, utrum sapientia, vel **scientia**, vel intellectus; 4 de subjecto ipsius; 5 de modo.*

**CASE 2. PLACE 2. Super Sent., lib. 1 q. 1 a. 1 arg. 1. [...]**<sup>-2</sup> Sicut enim dicit Dionysius in epistola ad Polycarpum, *philosophia est cognitio existentium; et constat, inducendo in singulis, quod de quolibet genere existentium in philosophia determinatur; quia de creatore et creaturis, tam de his quae sunt ab opere naturae, quam de his quae sunt ab opere nostro. Sed nulla doctrina potest esse nisi de existentibus, quia non entis non est **scientia**. Ergo praeter physicas disciplinas nulla doctrina debet esse.*

Home Search Browse About Resources News MyCRSBI

# The Corpus of ROMANESQUE SCULPTURE in Britain & Ireland

Home

## Welcome to CRSBI

The Corpus of Romanesque Sculpture is a complete online record of all the surviving Romanesque sculpture in Britain and Ireland, at more than 5000 sites. It provides us with a unique window on the aesthetics, beliefs, daily life, preoccupations, humour and technical skills of the artists and people of this creative and formative era from the late 11th century to the late 12th century.

Every entry is freely available and includes information on the historical and architectural context of the building, a first-class photographic record, and a scholarly description of the sculpture. Our work continues and many sites are already available on this website.

[Read More](#)

### Recent Images

CRSBI Annual Lecture 2019, Monday 29 April: Dr Ron Baxter

Dr Ron Baxter, our research director, gave a lecture entitled The Surviving Architecture and Sculpture of Reading Abbey, 8 pm at The Courtauld Institute of Art, Vernon Square.

CRSBI Annual Lecture archive

Each year since 2011 the CRSBI has had an annual lecture given by a distinguished scholar on a subject related to Romanesque architecture or sculpture. The annual lectures are hosted by the Courtauld Institute of Art and take place in the Kenneth Clark Lecture Theatre, except where noted below.

New Chairman for CRSBI Management Board

The Trustees and the Management Board of the Corpus of Romanesque sculpture in Britain and Ireland are pleased to announce the appointment of a new Chair of the Management Board, Dr Xavier Dectot. Xavier Dectot brings both scholarship and managerial expertise to the project and, since 2016, has been the Keeper of Art and Design at National Museums Scotland in Edinburgh.

Home Search Browse About Resources News MyCRSBI

# The Corpus of ROMANESQUE SCULPTURE in Britain & Ireland

Home » Search Results » St Michael, Knill, Herefordshire

## St Michael, Knill, Herefordshire

[Previous site](#) | [Return to search](#) | [Next site](#)

[Download as PDF](#)

### Feature Sets (2)

- Furnishings

### Description

Knill is a village in NW Herefordshire, close to the Welsh border and some 3 miles SW of Presteigne. It is on a wooded promontory on the N bank of the Hindwell Brook, a tributary of the Lugg, and the church and Knill Court are at the S of the village, alongside the brook. The name Knill is derived from OE *cnoll*, meaning knoll (Ekwall, 282). The church consists of a late-12th chancel with a plain N lancet, a nave that is mostly the result of a restoration of 1873-76, with a 14th S doorway under a timber porch, and an early 13th W tower.

### History

Knill was held as a manor of 2 hides by Osbern FitzRichard in 1086. He was the son of Richard Scrupe, a Norman who settled in Herefordshire before the Conquest and who built the castle which bears his name. The Domesday Survey records that Knill and a number of other manors 'was and is waste' and, by way of explanation, adds, 'it lies in the Welsh March'. It is clear that there was no church there at that time.

### Features

#### Furnishings

#### Fonts

#### Font



Exterior from NE

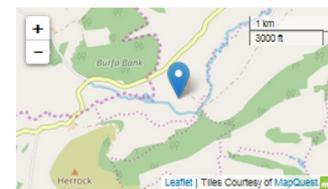


Exterior from SE



Interior to E

#### Location



[Full screen](#)

#### Site Location

Knill

National Grid Reference

#### Diocese

now: Hereford

medieval: Hereford

Destination

# The Corpus of ROMANESQUE SCULPTURE in Britain & Ireland

Home » Results

### Search Terms

fonts

### Filter By

- Country
- Boundaries (now)
- Boundaries (traditional)
- Dedication (medieval)
- Dedication (now)
- Diocese (medieval)
- Diocese (now)
- Type of

### Search Results

For the term **fonts**. Showing 1282 results.

Our picture archive will document every example of Romanesque sculpture in Britain and Ireland. At present our project is partially complete but our editors are adding material to the database every week, and our team is hard at work to complete the project.

Showing 1282 results.



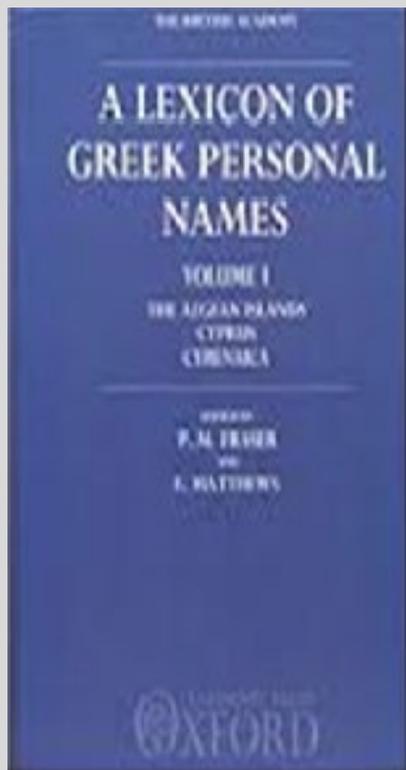
5000 km  
3000 mi

# LEXICON OF GREEK PERSONAL NAMES

LGN  
LEXICON

## Purpose and Scope

To collect and publish with documentation all known ancient Greek personal names (including non-Greek names recorded in Greek, and Greek names in Latin), drawn from all available sources (literature, inscriptions, graffiti, papyri, coins, vases and other artefacts), within the period from the earliest Greek written records down to, approximately, the sixth century A.D.



## Images illustrating sources for names



Tombstones



Vases



Papyri



Seals & Stamps



Names 'in situ'



Inscriptions



Coins



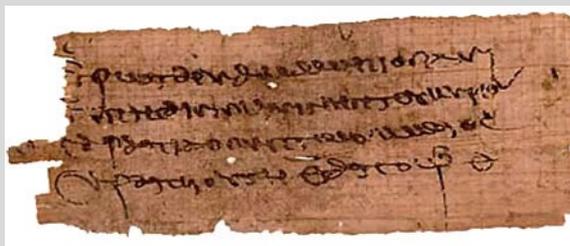
Gems

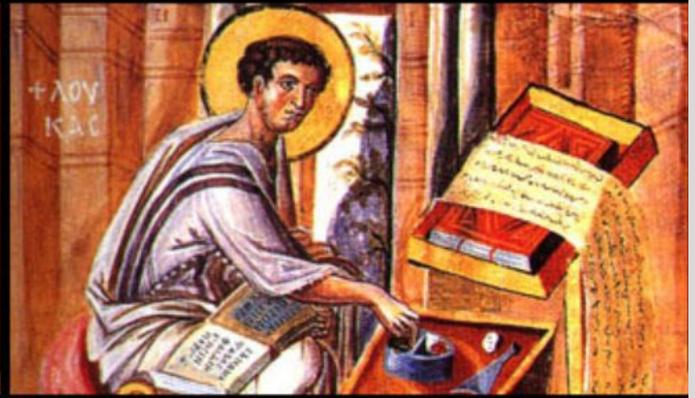


Manuscripts



Miscellaneous





### Welcome to the TLG®

The Thesaurus Linguae Graecae (TLG) is a Special Research Project at the University of California, Irvine. Founded in 1972 the TLG has collected and digitized most literary texts written in Greek from Homer to the fall of Byzantium in AD 1453. Its goal is to create a comprehensive digital library of Greek literature from antiquity to the present era. [\(more\)](#)

### News

Aug. 19, 2015: The corpus has been updated with 121 new texts from 63 authors [\(more\)](#)

Feb. 4, 2015: The TLG has released a new web site and search engine. [\(more\)](#)

October 6, 2014: The TLG is pleased to release the online version of the Lexikon zur byzantinischen Gräzität (LBG) [\(more\)](#)

Feb. 23, 2011: The TLG is pleased to announce the release of the *Online Liddell-Scott Jones*. [\(more\)](#)

October 29-30, 2009: *Reclaiming the Past, Envisioning the Future*: A conference to celebrate the 37th TLG anniversary. [Click here for the anniversary video \(more\)](#)

### Search the Online TLG®

Subscribers only

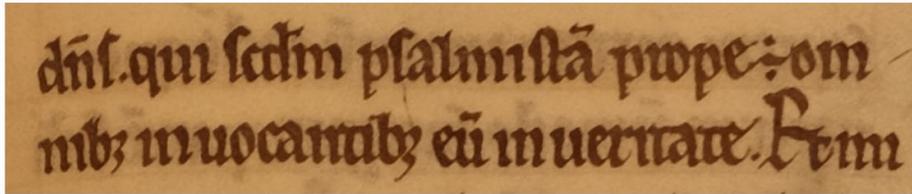
- Individuals
- Institutions

### Online Liddell-Scott-Jones (LSJ)



## Abbreviations

The `<ex>` (editorial expansion) tag indicates letters that you have added to expand an abbreviation:

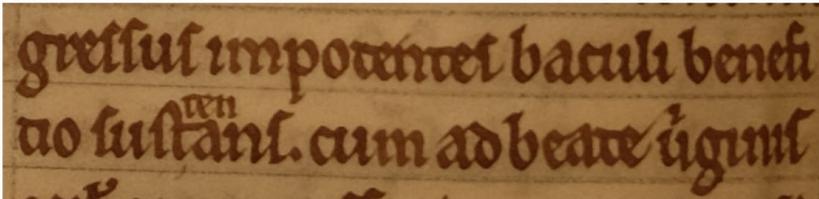


Oxford, Bodleian Library, Digby 177, fol. 19r

```
<lb/>d<ex>omi</ex>n<ex>u</ex>s. qui  
s<ex>e</ex>c<ex>un</ex>d<ex>u</ex>m  
psalmista<ex>m</ex> prope <ex>est</ex>  
om<lb break="no"/>nib<ex>us</ex>
```

## Additions

The `<add>` (addition) element allows you to note text added in a correction:



Oxford, Bodleian Library, Digby 177, fol. 6v

```
benefi<lb break="no"/>tio sus<add place="above">ten</add>tans.  
cum ad beate uirginis
```

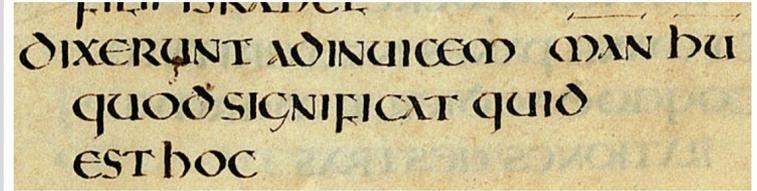
Source:

<https://andrewdunning.ca/transcribing-medieval-manuscripts-tei>

## Stylistic features

There are many TEI elements for marking [lists](#), different kinds of [quotations](#), and [emphasis](#); these are handy for showing your interpretation of a text. It can be useful to keep track of the stylistic features in a manuscript that prompt you to use these.

For example, the scribes of the [Codex Amiatinus](#) like to draw a line above any foreign words. You can mark this in your text along with noting that it's a word in a different language:



Florence, Biblioteca Medicea Laurenziana, Amiatino 1, fol. 64v

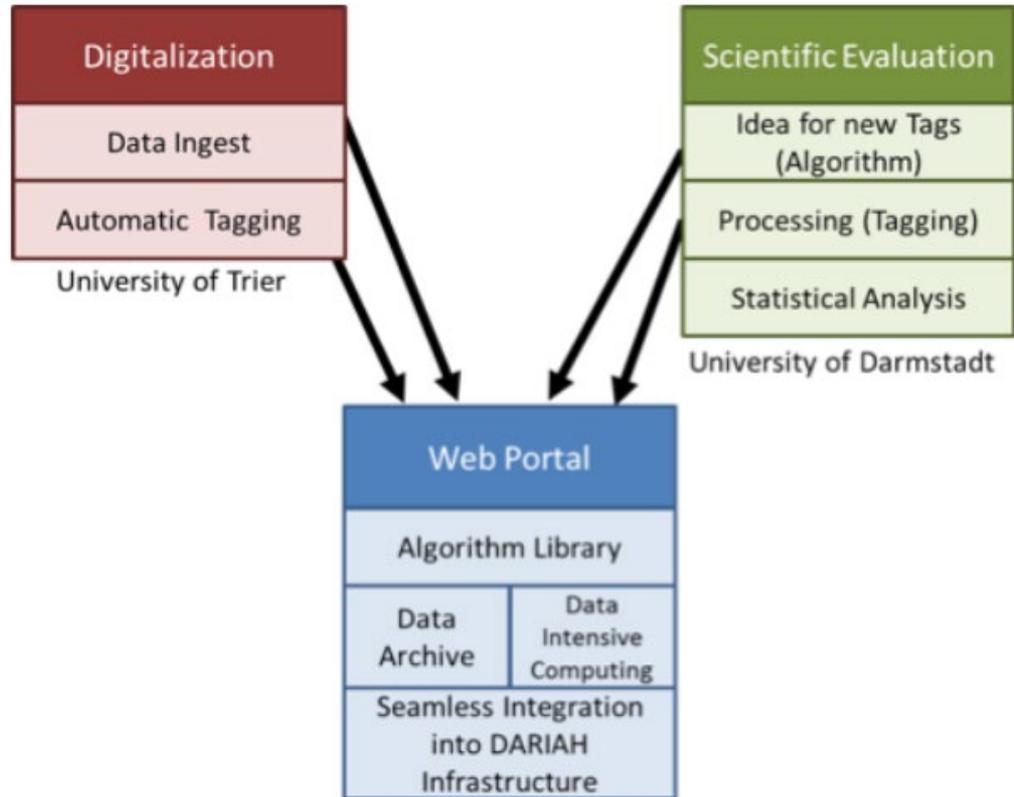
- Finding the right character
- Anatomy of TEI tags
- Paragraphs and headings
- New pages, lines, and columns
- Stylistic features
- Additions
- Deletions
- Substitutions
- Abbreviations
- Damage
- Unclear or illegible letters
- Editorial interventions
- Quotations
- Names

# Working with Manuscripts

## eCodicology - algorithms for automatic tagging of medieval mss



Green: page size, red: image size, yellow: text size, blue: initials.



Karlsruhe Institute of Technology

Collaboration within eCodicology.

<http://ipelsdf1.lsdm.kit.edu/cms/index.php/nav-pro-projects/nav-pro-act-lsdma/nav-pro-act-lsdma-dlcl-key/96-cat-projects/cat-projects-active/151-art-lsdma-dlcl-key-ecodicology>

# Working with Analogue Manuscript Collections

- Caveats to reuse:
  - Researchers normally gravitate to new questions and new materials.
  - When they revisit material, it often requires a complete reworking of the underlying data resource or its re-annotation.
  - Standards of imaging and user expectations are constantly changing
  - Datafication of manuscript collections mean different things to different researchers depending upon the collection.
  - That said comparative study depends upon consistency in representation, annotation, metadata, and analysis



- HOME
- ABOUT IDP
- COLLECTIONS
- EDUCATION
- CONSERVATION
- TECHNICAL
- ARCHIVES
- SITE MAP
- HELP
- CONTACT IDP

## The International Dunhuang Project: The Silk Road Online

IDP is a ground-breaking international collaboration to make information and images of all manuscripts, paintings, textiles and artefacts from Dunhuang and archaeological sites of the Eastern Silk Road freely available on the Internet and to encourage their use through educational and research programmes.



Bookmark and Share

IDP partner institutions which both provide data for and act as hosts to the multilingual website and database are:

- [The British Library, London IDP Directorate \(English Version\)](#)
- [The National Library of China, Beijing \(Chinese version\)](#)
- [The Institute for Oriental Manuscripts, St Petersburg \(Russian version\)](#)
- [Ryukoku University, Kyoto \(Japanese version\)](#)
- [The Berlin-Brandenburg Academy of Science and Humanities \(German version\)](#)
- [The Dunhuang Academy, Dunhuang \(Chinese version\)](#)
- [Bibliothèque nationale de France, Paris \(French version\)](#)
- [Research Institute of Korean Studies, Seoul \(Korean version\)](#)

**IDP NEWS**  
[Read the Latest Issue Here](#)  
[Join the IDP mailing list](#)

Major IDP partners providing data include:

- [The Library of the Hungarian Academy of Sciences, Budapest](#)
- [The British Museum, London](#)

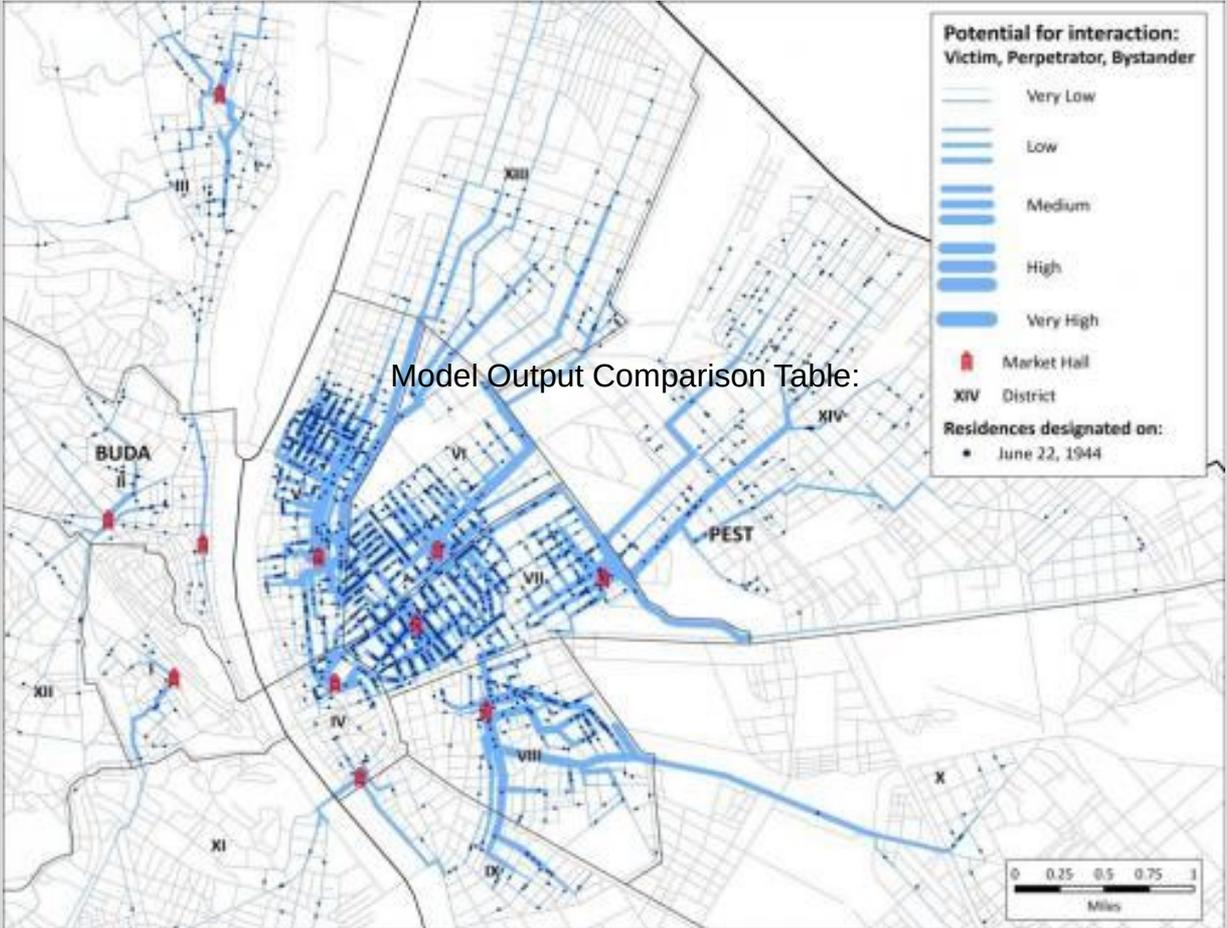
**SUPPORT IDP**  
Sponsor a Sutra  
Become a Silk Road Supporter

**IDP DATABASE**  
SEARCH THE IDP DATABASE  
  
GO

ADVANCED SEARCH  
CATALOGUE SEARCH  
BIBLIOGRAPHY SEARCH  
SEARCH TIPS

**SUPPORT IDP**  
SPONSOR A SUTRA  
MAKE A DONATION

# Enhancing Understandings of the Holocaust



From Prof Tim Cole (Bristol) Presentation at EHRI Conference (Amsterdam, 2019)

# Digital Humanities

- ❖ Not static
- ❖ Production of Scholarly Resources
- ❖ Datafication of sources
- ❖ Digitisation (scanning, multi-modal representations, 3D scans)
- ❖ Tagging (manual, automatic)
- ❖ Represents amenable to automatic reasoning



## A Public Record at Risk: The Dire State of News Archiving in the Digital Age

By Sharon Ringel  
and Angela Woodall

MARCH 28, 2019



SHARE  
ON TWITTER



SHARE  
ON FACEBOOK



EMAIL  
THIS STORY

[Executive Summary](#) | [Introduction](#) | [Methodology](#) | [Perceptions of News Preservation](#) | [The Intricacy of Archiving Digital News](#) | [Approaches to Preservation](#) | [Conclusion](#) | [Appendix: Additional Resources](#) | [Acknowledgments](#) | [Citations](#)

### EXECUTIVE SUMMARY

This research report explores archiving practices and policies across newspapers, magazines, wire services, and digital-only news producers, with the aim of identifying the current state of archiving and potential strategies for preserving content in an age of digital distribution. Between March 2018 and January 2019, we conducted interviews with 48 individuals from 30 news organizations and preservation initiatives.

What we found was that the majority of news outlets had not given any thought to even basic strategies for preserving their digital

### ABOUT THE TOW CENTER FOR DIGITAL JOURNALISM

The Tow Center for Digital Journalism at Columbia's Graduate School of Journalism is a research center exploring the ways in which technology is changing journalism, its practice and its consumption – as we seek new ways to judge the reliability, standards, and credibility of information online.

### TOW REPORTS

FRIDAY, SEPTEMBER 6TH, 2019

#### Guide to Native Advertising

Ava Sirrah

## CONCLUSION

Preservation is a multi-pronged process that technology can assist. But ultimately, maintaining news for the future depends on deliberate practices that involve planning around tasks such as migrating content to new formats, assigning consistent metadata, and indexing. Like most media organizations, the individuals interviewed for this report care about maintaining access to the news. But they are at a loss for what to do and may doubt their ability to prioritize preservation.

Sharon Ringel and Angela Woodall, 2019 (March 28), *A Public Record at Risk: The Dire State of News Archiving in the Digital Age*, [https://www.cjr.org/tow\\_center\\_reports/the-dire-state-of-news-archiving-in-the-digital-age.php](https://www.cjr.org/tow_center_reports/the-dire-state-of-news-archiving-in-the-digital-age.php)

# How is the Digital Reshaping Archives

News Opinion Sport Culture Lifestyle More ▾

UK ► UK politics Education Media Society Law Scotland Wales Northern Ireland

**Health policy**

## Covid contracts: minister replaced phone before it could be searched

Government expected to disclose James Bethell's correspondence relating to award of £85m of contracts for Covid tests

- Coronavirus - latest updates
- See all our coronavirus coverage

**Rowena Mason** Deputy political editor  
Wed 4 Aug 2021 13:11 BST



▲ A government lawyer's witness statement said Bethell replaced his phone in early 2021 and it may no longer be possible to retrieve information about dealings with Abingdon. Photograph: Roger Harris/UK Parliament

Labour has called for an inquiry into the use of WhatsApp within the government, after it emerged a health minister replaced his mobile phone before it could be searched for information relevant to £85m of deals that are subject to a legal challenge.

James Bethell, who oversaw the award of Covid contracts, is one of those under scrutiny over the way deals for personal protective equipment (PPE) and tests were allocated at the height of the pandemic.

As part of legal proceedings issued by the Good Law Project, the government is expected to disclose Lord Bethell's correspondence including by email, WhatsApp and SMS relating to the award of £85m of contracts for antibody tests to Abingdon Health.

The secretary of state has a responsibility to preserve and search documents for information relevant to the case from the point at which judicial review proceedings were issued in late 2020, under the government's "duty of candour".

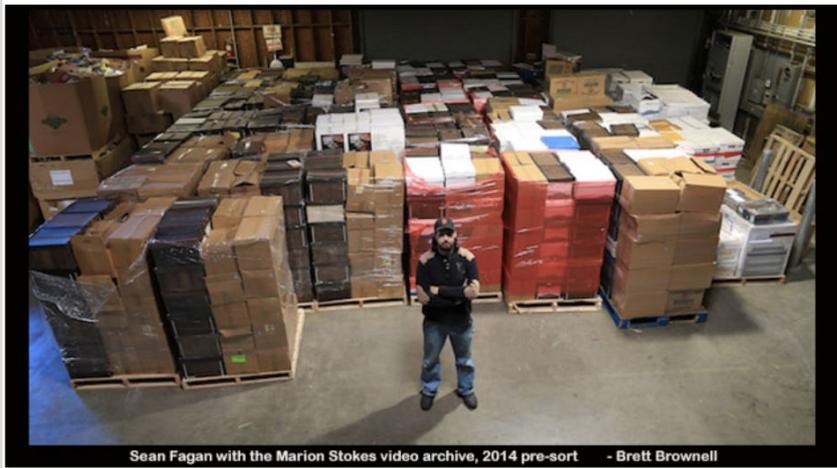
Advertisement



The statement said Bethell had used his official email account as well as his private email account to send and receive emails relevant to the contracts, and that he had also used his mobile phone for SMS and WhatsApp messages. But it said Bethell had confirmed that about six months ago his phone was broken and replaced and that his new phone did not contain the phone data.

Government lawyers revealed Bethell had not been issued with a "preservation notice" requiring him to save documents because ministers' official correspondence was routinely saved as a matter of course. However, this did not cover government business conducted by private means.

Bethell is already under investigation by the Information Commissioner's Office (ICO) over the use of private emails for government business, prompted by revelations that his former boss **Matt Hancock** was using a private account at the height of the pandemic.

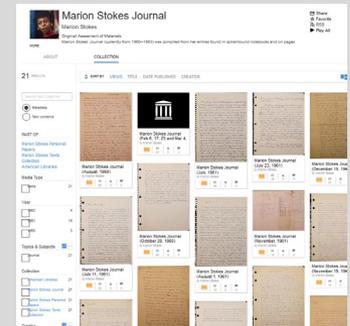
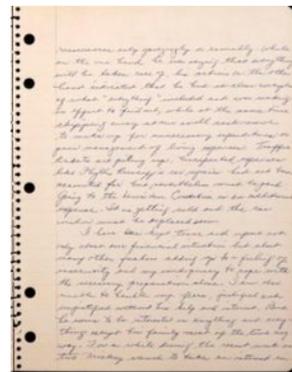


Sean Fagan with the Marion Stokes video archive, 2014 pre-sort - Brett Brownell

- 35 years of off-air recording of TV News
- More than 71,716 VHS tapes
- 840,000 hours of news casts
- Began on 4 November 1979

## Personal Journals

Ms. Stokes was a committed diarist, note taker and list maker. Under the leadership of archivist [Jackie Jay](#), The Internet Archive has been digitizing the contents of 55 bankers boxes of her papers that include her personal journals, magazines, newspapers, civic organization pamphlets, leaflets and handbills. Some of her earliest (1960 & 1961) hand-written journal entries are now publicly available and can be viewed [here](#). More will be added as they are scanned and QC'd.



# Archives as places of discovery



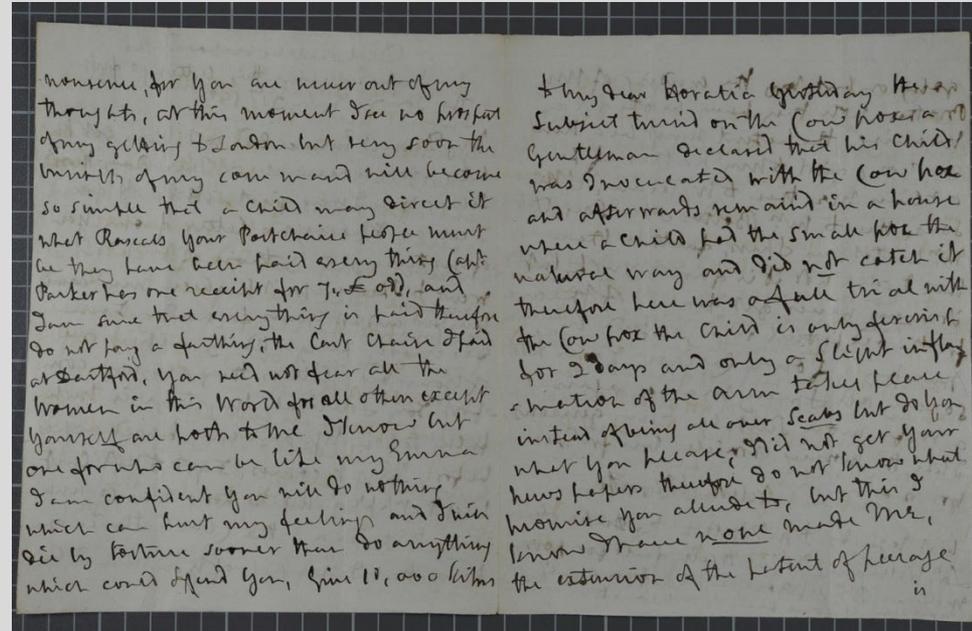
Vice-Admiral Horatio Nelson 1758-1805, 1st Viscount Nelson (BHC2887)  
Painter: Lemuel Francis Abbott  
oil on canvas, 1798  
National Maritime Museum, Greenwich, London, Greenwich Hospital Collection  
<https://collections.rmg.co.uk/collections/objects/14360.html>



Horatia Nelson, 1801-81 (BHC2886)  
oil on canvas, c.1815  
Artist: unidentified  
National Maritime Museum, Greenwich, London, Nelson-Ward Collection  
<https://collections.rmg.co.uk/collections/objects/14359.html>



Emma Hamilton  
by George Romney  
oil on canvas, circa 1785  
NPG 294  
© National Portrait Gallery, London  
<https://www.npg.org.uk/collections/search/use-this-image/?mkey=mw02854>



Horatio Nelson to Emma Hamilton, 31 July 1801.  
(© National Maritime Museum. Greenwich),  
<https://www.rmg.co.uk/stories/blog/curatorial/lord-nelson-smallpox-question-vaccination>

## Search the LGPN database

**LGNP Name Search** offers online analysis of all the primary names published in LGPN I (Aegean Islands, Cyprus, Cyrenaica), II A (the revised version of Attica), III.A (Peloponnese, Western Greece, Sicily, Magna Graecia) and III.B (Central Greece), IV (Macedonia, Thrace, Northern Shores of the Black Sea), and VA (Coastal Asia Minor: Pontos to Ionia), a total of 35,982 names.

Dialect forms, of which there are very many in LGPN III.B, stand as separate forms. We intend in due course to set up a system allowing the search to take account of dialect.

The searching does not include 'secondary' names (i.e., conjectured restorations, names containing diacritics, orthographic variants etc., which, in the published volumes, are recorded in the final brackets); unpublished names, or addenda and corrigenda since publication.

### The search results show:

all names in which the search string occurs, the volume(s) of LGPN in which they occur, and their frequencies.

### Common uses of the searching facility include identifying:

names with a particular root or termination e.g., all names ending in -κροτη

name-forms based on a particular deity e.g. names based on Isis (Ἰσιδιωρος, Ἰσιδοτος etc.)

names matching a string of characters in a fragmentary inscription

### Database Search

The database search can be accessed [here](#).

## Lexicon of Greek Personal Names

Name (unaccented):  [Keyboard](#)

Place:

Start date:  End date:  [show advanced search options](#)

[Show number of results](#) [Show results](#)

[View search](#) | [Name catalogue](#) | [Place catalogue](#) | [Status and profession catalogue](#)

You searched for the following:

**status:** doctor

420 results

[Direct link for this search](#)

Download these results as: [XML](#) [JSON](#) [KML](#) [CSV](#) [RDF](#)

Table [Map](#)

isplay  records

ID	Vol.	PubID	Name	Sex	Place	Floruit
<a href="#">V1-10056</a>	1	20	<a href="#">Διφιλος</a>	[m.]	<a href="#">Siphnos</a>	<a href="#">iv/iii BC</a> <a href="#">RE (17) (doctor)</a>
<a href="#">V1-15586</a>	1	8	<a href="#">Αραπίδης</a>	[m.]	<a href="#">Kos</a>	<a href="#">s.ii BC</a> <a href="#">Milet I (3) 184-5 (doctor)</a>
<a href="#">V1-15655</a>	1	1	<a href="#">Μνασίμαχος</a>	[m.]	<a href="#">Kos</a>	<a href="#">m.iv BC</a> <a href="#">Hp., Epid. vii 112 ([Mvη]σίμαχος; doctor)</a>
<a href="#">V1-15919</a>	1	109	<a href="#">Φίλιππος</a>	[m.]	<a href="#">Kos</a>	<a href="#">240BC</a> <a href="#">PP 16640; = HE 1049 (f. Καφισοφών; doctor)</a>
<a href="#">V1-17103</a>	1	4	<a href="#">Φυλότιμος</a>	[m.]	<a href="#">Kos</a>	<a href="#">m.iii BC</a> <a href="#">RE (1); HG 14, 1 &amp; p. 38 (s. Τιμόλυκος; doctor)</a>
<a href="#">V1-17117</a>	1	76	<a href="#">Απολλωνίδης</a>	[m.]	<a href="#">Kos</a>	<a href="#">v BC</a> <a href="#">RE (32) (doctor)</a>
<a href="#">V1-17119</a>	1	10	<a href="#">Θεσσαλός</a>	[m.]	<a href="#">Kos</a>	<a href="#">s.iv BC</a> <a href="#">RE (5) (s. Ἰννοκράτης; f. Ἰννοκράτης; doctor)</a>
<a href="#">V1-17123</a>	1	114	<a href="#">Ἡρακλείδης</a>	[m.]	<a href="#">Kos</a>	<a href="#">vi/v BC</a> <a href="#">RE (53) (s. Ἰννοκράτης; f. Ἰννοκράτης; doctor)</a>
<a href="#">V1-17411</a>	1	10	<a href="#">Φίλιππος</a>	[m.]	<a href="#">Kos</a>	<a href="#">c.242-200BC</a> <a href="#">SEG XXVII 510 (s. Νικαρχος; doctor)</a>
<a href="#">V1-17452</a>	1	5	<a href="#">Γνωσιδικος</a>	[m.]	<a href="#">Kos</a>	<a href="#">arch.</a> <a href="#">Suda I 565 (f. Ἰννοκράτης; doctor)</a>
<a href="#">V1-17453</a>	1	10	<a href="#">Ἰννοκράτης</a>	[m.]	<a href="#">Kos</a>	<a href="#">vi/v BC</a> <a href="#">Suda I 565 (s. Γνωσιδικος; f. Ἡρακλείδης; doctor)</a>
<a href="#">V1-17454</a>	1	12	<a href="#">Ἰννοκράτης</a>	[m.]	<a href="#">Kos</a>	<a href="#">s.iv BC</a> <a href="#">Suda I 566 (s. Θεσσαλός; doctor)</a>
<a href="#">V1-17455</a>	1	11	<a href="#">Ἰννοκράτης</a>	[m.]	<a href="#">Kos</a>	<a href="#">iv BC</a> <a href="#">Suda I 567 (s. Ἀράκων; doctor)</a>
<a href="#">V1-17457</a>	1	9	<a href="#">Ἰννοκράτης</a>	[m.]	<a href="#">Kos</a>	<a href="#">?</a> <a href="#">Suda I 568 (s. Ουμβραίος; doctor)</a>
<a href="#">V1-17460</a>	1	13	<a href="#">Ἰννοκράτης</a>	[m.]	<a href="#">Kos</a>	<a href="#">hell.</a> <a href="#">Suda I 569 (s. Πραξιόνης; doctor)</a>

## BEAZLEY ARCHIVE POTTERY DATABASE (BAPD)

[Show a sample vase](#)

The BAPD is the world largest database of ancient Greek painted pottery (Greek vases). It contains records of almost 130,000 ancient pots and about 250,000 images. Nearly all of the pots included were made during the 6th to 4th centuries BC, and about three quarters Classical Art Research Centre in Oxford. The great majority of its images were collected by the great pottery expert, Sir John Beazley (1885-1970). However, the Archive grew further after his death, and the BAPD contains even more images acquired later from a variety of Beazley's own drawings. Where the source of an image is not obvious please contact us for information.

The BAPD is generously supported by



SCROLL DOWN TO USE THE DATABASE

There are different methods of searching the BAPD. The 'Basic Search' option will allow you to carry out very simple filtering of the database. Even advanced researchers sometimes find this a quick and easy way of getting the results they need, but if you want access to functions are available in the 'Traditional Search', which is the old-format search form familiar to many of our long-term users.

For help with advanced searching watch this short video [coming soon].

See also

[Short Guide to the BAPD Traditional Search interface](#)

[Guide to searching CARC Open Data Programmatically](#)

A note on vase inscriptions: addition of vase inscriptions to the BAPD began in 1999 with the conversion of Henry Immerwahr's data for *A Corpus of Attic Vase Inscriptions* (CAVI). The database contains more than 8,000 inscriptions, and includes earlier updates from Rudolf on the [AVI Project website](#) at the University of Basel. Links in BAPD records also direct the user to the Basel database.

If you create an account you can preserve your Photograph Album.

[Click here to login or register \(optional\)](#)



BASIC SEARCH



<https://www.carc.ox.ac.uk/XD/B/ASP/searchOpen.asp>

# CLASSICAL ART RESEARCH CENTRE

## Beazley Archive Pottery Database - Advanced Search Form

Showing search results for: All Pottery

General search term

  
[help examples](#)

Combining search terms

When selecting multiple search terms from the lists below, combine them with:

Use \* as a wildcard when searching lists

Vase Number

Fabric

Technique

Sub Technique

Shape Name

Provenance

Date Range

Inscription Type

Inscription

Artist Name

Scholar Name

Decoration Termword

Decorated Area

Collection Name

Number

Publication Name

Reference

Miscellaneous

Results
<input type="checkbox"/> image <input type="checkbox"/> all images <input checked="" type="checkbox"/> text <input type="checkbox"/> detailed text <input type="checkbox"/> map <input type="checkbox"/> timeline <input type="checkbox"/> lightbox <a href="#">download</a>
1, ATHENIAN, Munich, Antikensammlungen, Munich, Antikensammlungen, Rome, Candelori Collection, 1539
2, ATHENIAN, Athens, Agora Museum, P13332
3, ATHENIAN, Rome, Mus. Naz. Etrusco di Villa Giulia, 63614
4, ATHENIAN, Budapest, Hungarian Museum of Fine Arts, 50.732
5, ATHENIAN, Rome, market, Basseggio, London, British Museum, B240
6, ATHENIAN, Munich, Antikensammlungen, Munich, Antikensammlungen, Rome, Candelori Collection, J989
7, ATHENIAN, Athens, P. Kanellopoulos
8, ATHENIAN, Malmaison, Empress Josephine, Paris, Musée du Louvre, F386
10, ATHENIAN, Basel, market, Münzen und Medallen A.G.
11, ATHENIAN, Geneva, market, Koutoulakis, Mannheim, Reiss-Museum, Basel, market, Münzen und Medallen A.G., CG343
12, ATHENIAN, Athens, Agora Museum, P372
14, ATHENIAN, New York (NY), private, London, market, Sotheby's, Basel, market, Münzen und Medallen A.G.
15, ATHENIAN, Basel, market, Münzen und Medallen A.G.
16, ATHENIAN, London, market, Christie's
17, ATHENIAN, London, market, Christie's
18, ATHENIAN, London, market, Christie's
19, ATHENIAN, London, market, Christie's
20, ATHENIAN, Durham (NC), Duke University, Museum of Art, London, market, Christie's, 1974.6
21, ATHENIAN, London, market, Christie's
22, ATHENIAN, Aegina, Archaeological Museum, 155
23, ATHENIAN, London, market, Bonhams, Kiel, private
24, ATHENIAN, London, market, Christie's, Northampton, Castle Ashby, 8
25, ATHENIAN, Austin (TX), University of Texas, Jack S. Blanton Museum of Art, Austin (TX), University of Texas, Jack S. Blanton
26, ATHENIAN, New York (NY), Callimanolopoulos, London, market, Christie's, Northampton, Castle Ashby, 14
27, ATHENIAN, New York (NY), market, Sotheby's, Texas, McCoy, London, market, Christie's, Northampton, Castle Ashby, 15
28, ATHENIAN, Greece, private, London, market, Christie's, Northampton, Castle Ashby, 22
29, ATHENIAN, New York (NY), market, Christie's, Texas, McCoy, London, market, Christie's, Northampton, Castle Ashby, 7
30, ATHENIAN, New York (NY), market, Sotheby's, Texas, McCoy, London, market, Christie's, Northampton, Castle Ashby, 16
31, ATHENIAN, Malibu (CA), The J. Paul Getty Museum, 75.AE.28



# Facets of Metadata Revolutions

- ❖ Pragmatics
  - ❖ What was the context of the object
  - ❖ What is the paradata?
  - ❖ What processes are associated with the object?
- ❖ Semantics
  - ❖ Specified in terms of metadata element sets
  - ❖ The meaning of metadata elements (often inherent in the element name)
- ❖ Syntax
  - ❖ Specified in terms of encoding & representation schemes
  - ❖ The structural relationships in which metadata elements may stand with one another (defined in the metadata scheme)

# Automating Appraisal and Description

- ❖ Automating the process of selecting material
- ❖ Annotation and Provenance
  - ❖ What about structural differences—say between radiological and linguistic data sets and their annotations
- ❖ Summarisation technologies (point of view)
  - ❖ What about composite documents
  - ❖ What about databases, images,
- ❖ Knowledge representation developments & tools
- ❖ What if collection development testing

# Serendipity

- ❖ Humanities Research depends experiential immersion in source materials.
- ❖ The more you examine the more original your questions are likely to be.
- ❖ Digital humanities offers the possibility for explore much more diverse arrays of resources.
- ❖ LLMs provide a wonderful mechanism to explore documentary resources.

# New ways of constructing arguments

- ❖ Digital Humanities and the “remixing” of media.
- ❖ Inspiration from the work of Martin Irvine (e.g., dialogism—essential for new expressions) and Virginia Kuhn (e.g., argumentation) in the theory of remix studies
- ❖ All Scholarship builds on scholarship of the past.
- ❖ Depends upon
  - ❖ understanding earlier scholarship
  - ❖ Asking Novel questions
  - ❖ Access to resources (primary resources and their digitization)
  - ❖ Access to metadata and paradata

# Human Sciences

- ❖ Digital Humanities = Digital Human Sciences
- ❖ Integrity
- ❖ Repeatability
- ❖ Accessibility of source material
- ❖ Curation of Data, Applications/programs
- ❖ Curation of argumentation and dissemination

# Digital Transformation of Humanities Scholarship

- ❖ Creates an environment for Interdisciplinarity and transdisciplinarity
- ❖ Emergence of integrated and dynamic papers
- ❖ Data as scientific output – building resources
- ❖ Relationship between data, tools (e.g., programs) and digital scholarship.

# Activity One

## An 18<sup>th</sup> Century Cookbook...

# Transition

# Potential Roles of LLMs for Historical Research (pt 1)

- ❖ Textual Analysis & Interpretation (analyze corpora including letters, mss, newspapers, diaries)
- ❖ Semantic analysis such as understanding nuanced meanings, contexts, or rhetorical styles in documents
- ❖ Extraction of recurring themes, sentiments, or ideological patterns
- ❖ Retrieval from digital Archives, and eventually digitized archives, by enabling semantic search, context relevance beyond keywords, investigating connections across documents, subjects, places, periods.

# Potential Roles of LLMs for Historical Research (pt 2)

- ❖ Chronological reasoning and timeline construction
- ❖ Parsing documents to flag gaps, contradictions, or anachronisms in historical narratives.
- ❖ Highlight inconsistencies or fabrications in narratives or secondary sources.
- ❖ Postulating alternative interpretations or new research directions by detecting patterns, exploring overlooked primary sources, probabilistic modelling of social, political or cultural trends.
- ❖ Support for translation and deciphering documents.
- ❖ Pattern recognition – medieval textual studies



# Education, Dissemination, Engagement, and Curation

- ❖ Dissemination through LLMs powering chatbots for cultural heritage institutions from Museums to Archives;
- ❖ Creation of Interactive narratives to support engagement with historical resources;
- ❖ Engaging Civic Society in historical processes, such supporting crowdsourcing of transcription verification, OCR validation and correction, and annotation; and,
- ❖ Pedagogy such as generating context summaries, timelines or visualizations, assisting with primary source study, and creating practice quizzes.
- ❖ Automating data curation processes in repositories and archives.

# Model Limitations Concerns

Limitation	Description
<i>Hallucination Risk</i>	Plausible but false content is often generated
<i>Temporal Bias</i>	Older historical periods are poorly represented
<i>Cultural Bias</i>	Western-dominant framing distorts non-Western narratives
<i>Marginalized Voices Absent</i>	Underrepresented in training data and model responses
<i>Corroboration Required</i>	Mitigates all other limitations; mandatory for reliability

## ❖ Recommendation:

- ❖ DeepSeek-V3 has lower hallucination rates than GPT-4, but as rigorously verified Claude

# Example Impacts of Western Training Data

Limitations	Description	Consequence
Linguistic Bias and Incomplete Coverage	LLMs trained primarily on English struggle with non-English texts, especially pre-modern, minority, or colonial languages; they favor standard modern English, neglecting dialects and historical forms.	Misinterpretation or exclusion of non-English or archaic texts in historical research.
Cultural Erasure and Eurocentrism	Western-centric narratives dominate training data, leading to biased outputs in summarization, inference, and topic detection, especially harmful in postcolonial and indigenous contexts.	Reinforcement of colonial narratives and misrepresentation of marginalized histories.
Distortion of Semantic Meaning	English-centric models often mistranslate culturally specific idioms or phrases, leading to errors in downstream NLP tasks such as named entity recognition and event sequencing.	Propagation of translation errors and misclassification in historical document analysis.
Temporal Bias in Historical Corpora	Training on recent English data fails to capture older semantic contexts; models mislabel historical figures, apply modern ideology, and omit critical terminology.	Inaccurate historical modeling and skewed understanding of events and ideologies.
Limitations in Metadata Inference	Multilingual or sparse metadata in archives is poorly handled by English-centric LLMs, which struggle with interpreting place names, local roles, and document structures.	Inability to effectively index or analyze diverse historical archives.

# “LLMs for Low-resource Languages—A call to Action”

- ❖ Recognizes that low-resource languages preserve unique cultural and intellectual traditions, but at risk due to globalization, marginalization, lack of interest, and limited technological support;
- ❖ Stress that LLMs offer new ways for studying low-resource languages, enabling tasks like translation, text generation, and historical analysis;
- ❖ Lots of challenges such as sparse and low-quality corpora and risks of cultural misrepresentation or bias;
- ❖ Emerging approaches including transfer learning, data augmentation, and multimodal integration create possibilities for LLMs constrained by minimal training data;
- ❖ LLMs can broaden historical and linguistic insights by reconstructing lost texts, modeling dialect evolution, and accessing underused cultural records;
- ❖ They draw attention to ethical risks such as cultural appropriation, loss of linguistic authenticity, and inadequate community consent, especially for Indigenous or sacred knowledge—think FAIR and CARE Principles.

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Yiheng Liu, Haiyang Sun, Yi Pan et al.  
"Opportunities and challenges of large language models for low-resource languages in humanities research." *arXiv preprint arXiv:2412.04497* (2024)



# Example Ethical Concerns

Ethical Concern	Description	Consequence
<i>Falsified History</i>	Hallucinated or invented facts	Misleads researchers
<i>Voice Erasure</i>	Excludes marginalized groups	Reinforces dominant narratives
<i>Lack of Attribution</i>	No citations for claims	Undermines academic integrity
<i>Privacy Infringement</i>	Sensitive archival use	Ethical/legal violations
<i>Cultural Misrepresentation</i>	Disrespects heritage context	Ethical appropriation
<i>Displacement of Historians</i>	Overreliance on AI	Loss of critical scholarship
<i>Bias Reproduction</i>	Encodes historical discrimination	Normalizes injustice
<i>Opaque Reasoning</i>	No explanation of outputs	Undermines transparency

# Focus today is on Instruct Models

- ❖ Large language models (LLMs) that have been fine-tuned to follow explicit user/human instructions, usually using instruction tuning and Reinforcement Learning with Human Feedback (RLHF).
- ❖ Designed to follow prompts such as, “summarize,” “extract names,” or “translate this sentence.”
- ❖ Optimal for natural language interfaces (chatbots, user queries, interactive NLP tasks)
- ❖ Trained on curated input-output pairs and refined by humans to prefer helpful, correct responses. (i.e., the model has learned from many examples of questions and their ideal answers—across many domains—so it can generalize and generate useful responses to new prompts.)

# Focus today is on Instruct Models

- ❖ Instruct model have undergone additional supervised fine-tuning using prompt-response pairs and refined by humans to prefer helpful, correct responses. (i.e., the model has learned from many examples of questions and their ideal answers—across many domains—so it can generalize and generate useful responses to new prompts.)
- ❖ This fine-tuning helps the model:
  - ❖ Better understand task framing;
  - ❖ Provide more structured or goal-oriented responses; and,
  - ❖ Be more robust to vague or ambiguous prompts.

LLMs can be tools for supporting historical inquiry they are not computational historians.

With the historian-in-the-loop they can support, exploration, discovery, analysis, source mining, theory building and validation, and dissemination.



# Activity Two, Part 1

## Letters of Mary Shelley

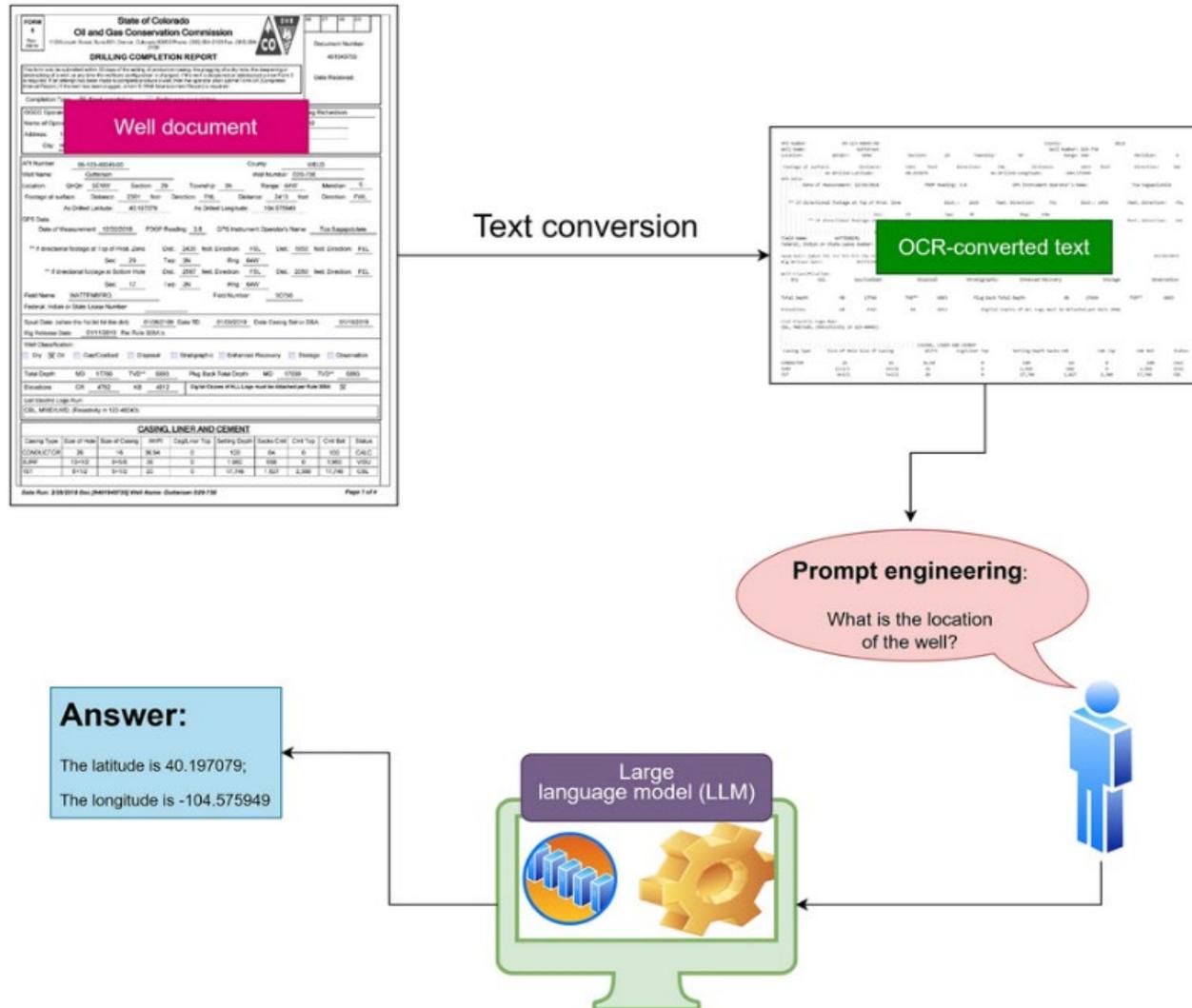
### **Summarisation & Interpretation**

# Looking beyond Humanities...

- ❖ Historical oil and gas well records from 1950s to early 2000s
- ❖ Colorado Energy and Carbon Management Commission's online system (COGIS) (150 records)
- ❖ Pennsylvania Geological Survey's EDWIN online tool (10)
- ❖ Many unstructured, handwritten, or scanned, making manual digitization labor-intensive and impractical at scale
- ❖ Developed and validated an automated information extraction workflow combining OCR and LLMs, primarily using open-source Llama 2 models.
- ❖ 100% accuracy on structured PDF documents (CO) and 70–90% accuracy on OCR-derived image-based records (PA)

Zhiwei Ma, Javier E. Santos, Greg Lackey, Hari Viswanathan, and Daniel O'Malley. "Information extraction from historical well records using a large language model." *Scientific Reports* 14, no. 1 (2024): 31702.

From: [Information extraction from historical well records using a large language model](#)



The proposed workflow for well information extraction via LLM.

Ma, et al, 2024, figure 1.

# Lessons...

- ❖ More complex prompts, larger models (70B vs. 7B or 13B), and improved OCR text quality led to better performance.
- ❖ Investing effort in prompt engineering paid off.
- ❖ The study demonstrates that LLM-based workflows are a scalable solution for geoscientific data extraction.
- ❖ Advances in LLMs make previously inaccessible historical technical records accessible. Implications for archives.

Zhiwei Ma, et al, 2024

# Next Steps

- ❖ Improve OCR performance, especially for image-based and handwritten documents;
- ❖ Customize pre-trained models for domain to improve accuracy, especially for noisy or inconsistent inputs;
- ❖ Use more powerful GPUs to run larger models;
- ❖ Explore models that can handle PDFs or images directly, eliminating OCR;
- ❖ Embed tools to assist with unit conversions and error correction;
- ❖ Expand dataset size.

Zhiwei Ma, et al, 2024

# Histolens...

- ❖ A structured LLM-based framework for exploring historical texts.
- ❖ Primarily employed OpenAI's GPT-4 API for all LLM-based analyses-- general reasoning and nuanced language understanding capabilities.
- ❖ Delivers support for layered historical interpretation.
- ❖ Case Study of Chinese political-philosophical text Yantie Lun (Discourses on Salt and Iron). Han dynasty, records debates about government monopolies with rich rhetorical and political content.
- ❖ Advances the notion of structured collaboration between machine-generated suggestions and expert interpretation.

Yifan Zeng, 2024, HistoLens: An LLM-Powered Framework for Multi-Layered Analysis of Historical Texts -- A Case Application of Yantie Lun,

<https://arxiv.org/abs/2411.09978>



# Histolens...

- ❖ Framework operates on three levels:
  - ❖ Surface-level Analysis includes summarization, named entity recognition, and rhetorical question detection.
  - ❖ Interpretive-level Analysis includes extracting speaker stances, argumentative structure, and sentiment.
  - ❖ Meta-level Analysis supporting probing for shifts in argumentation and socio-political framing across speakers and sections.
- ❖ Key Insights:
  - ❖ LLMs can support historiographical analysis
  - ❖ There are limits in terms of nuance, cultural inference, and ambiguity.
  - ❖ Human validation remains crucial: the LLM output is used as a prompt for deeper human reflection and interpretation.

Zeng, 2024.

# Textual Restoration...

- ❖ Develop and evaluate a DL model, Ithaca, to assist historians in restoring and interpreting fragmentary Greek inscriptions
  - ❖ Textual restoration of damaged ancient inscriptions,
  - ❖ Identification of the geographical location of creation,
  - ❖ Dating the Inscription
- ❖ Purpose and Broader Goal:
  - ❖ demonstrate how AI can support historical expertise
  - ❖ Improving the accuracy and speed of epigraphic workflows,
  - ❖ Making predictions interpretable and collaborative,
  - ❖ Contributing to ongoing historical debates (e.g., Athenian decrees)

Yannis Assael Thea Sommerschild, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. "Restoring and Attributing Ancient Texts Using Deep Neural Networks." *Nature (London)* 603, no. 7900 (2022): 280–83. <https://doi.org/10.1038/s41586-022-04448-z>.

# Ithaca...

- ❖ Ithaca trained on 78,608 texts from the Packard Humanities Institute (PHI) Greek inscriptions corpus.
- ❖ DL Model based on a transformer-based architecture with task-specific heads to handle the different epigraphic tasks.
- ❖ Inputs are processed as a combination of character and word embeddings, augmented with positional information.
- ❖ Performance:
  - ❖ Restoration accuracy: 61.8%, improved to 71.7% by historians.
  - ❖ Region attribution accuracy: 70.8%.
  - ❖ Date attribution accuracy: Mean error of 29.3 years (compared to 144.4 years for baseline).

A transformer-based architecture a DL model introduced by Vaswani et al in 2017 (“Attention Is All You Need”). The foundation for many state-of-the-art language models, including BERT, GPT, and Ithaca.

### ❖ Key Features of Transformer Architecture:

- ❖ Self-Attention Mechanism ensures model considers all words.
- ❖ Using positional coding transformers encode position explicitly.
- ❖ Using a Layered Structure, transformer layers each contain:
  - ❖ Multi-head self-attention (to focus on different aspects of the input),
  - ❖ Feedforward networks (to transform the representations),
  - ❖ Residual connections and normalization (to stabilize training and help gradients flow).
- ❖ Transformers allow parallel computation across tokens

# Knowing Tokens Count Matters for

- ❖ Prompt size limits:
  - ❖ GPT-4-turbo has a 128k token limit.
  - ❖ GPT-3.5-turbo has a 16k token limit (some variants only 4k).
  - ❖ Knowing token counts helps you avoid truncation or context loss.
- ❖ Pricing (for API usage):
  - ❖ Costs are based on tokens in both prompt and response.
  - ❖ Example: GPT-4-turbo charges per 1,000 tokens.
- ❖ Fine-tuning or embedding tasks:
  - ❖ Token count affects how text is chunked.
  - ❖ Longer documents must be split at logical token boundaries.
- ❖ Performance in long-context interactions:
  - ❖ If your conversation hits the limit (e.g., analyzing 100 pages), careful token management ensures nothing important is dropped.

# Tokens - View One

- ❖ A token is the basic unit of text that a system processes.
- ❖ In Natural Language Processing (NLP):
  - ❖ Word-level tokens: individual words separated by spaces or punctuation
  - ❖ Example: "Hello world!" → ["Hello", "world", "!"]
  - ❖ Subword tokens: parts of words, useful for handling unknown words or morphology
  - ❖ Example: "unhappiness" → ["un", "happy", "ness"]
  - ❖ Character-level tokens: individual characters including letters, numbers, punctuation, and spaces
  - ❖ Example: ".", "#", "A"

# Tokens - View Two

- ❖ A token is the basic unit of text that a system processes.
- ❖ In LLMs (e.g., GPT, Claude, Deepseek):
  - ❖ Tokens are subword units that balance efficiency and meaning
  - ❖ One token  $\approx$  0.75 words in English on average
  - ❖ Example: "tokenization" might be split into ["token", "ization"]
  - ❖ Models have token limits (e.g., context windows of 4K, 8K, 128K tokens)
  - ❖ Tokens have a “price” associated with them
  - ❖ LLMs enforce Token Limits, these limits include the “input” window and the “output” (i.e., your prompt, the LLMs reply and often previous prompts and replies in the conversation)

# Tokens

- ❖ They define how the model reads and understands text and are LLM uses to learn patterns, meaning, grammar, and context.
- ❖ Every model has a token limit for what it can "remember" in one interaction (its context window). If you exceed the limit, the model will lose or truncate earlier content.
- ❖ They affect cost and performance (processing time)
  - ❖ For example:
- ❖ They impact model behavior as they do not always equal words. For example they effect translation, paraphrasing, or summarization accuracy or in the case of rare words they can "confuse" the LLM.

# Do you need to worry about Tokens

Yes and No:

- ❖ Yes, for
  - ❖ Historical document analysis
  - ❖ Amalgamating archival records
  - ❖ Running multi-turn structured research queries
  - ❖ Planning to use the API
- ❖ No, for
  - ❖ Working with single sentences or short paragraphs
  - ❖ Staying within the visible chat window
  - ❖ Not approaching size or cost limits

# Optimizing Tokens

- ❖ "Please summarize this document in three sentences."  
(8 Tokens)
- ❖ "Can you provide a summary of this document using three complete sentences?" (14 Tokens)
- ❖ The second sentence makes the same request using 6 more tokens.
- ❖ In analyzing 100s of historical documents a more efficient prompt reduces costs and improves performance.
- ❖ Example: if you were examining 1000 documents the second query itself would use 6000 extra tokens

# Tokens matter when refining prompts

- ❖ Prompt engineering depends on understanding tokenization:
  - ❖ You can trim prompts to stay under the limit
  - ❖ You can test alternative phrasings to reduce token count or improve clarity
  - ❖ You can chunk documents efficiently for analysis

# Tokens matter when refining prompts

- ❖ In prompt engineering, fewer tokens can mean better control in terms of
  - ❖ Reduce the risk of unexpected behavior due to verbosity
  - ❖ Tend to guide the model more predictably (less room for ambiguity)
  - ❖ They tend to bring benefits in terms of speed, cost, control, and capacity.
- ❖ But Precision is critical and sometimes short prompts are not sufficient.

- ❖ Why Transformers Matter for Text Restoration and Attribution:
- ❖ In the case of Ithaca, the transformer architecture allows the model to:
  - ❖ understand incomplete or fragmented text by modeling dependencies that are even at distances from the object.
  - ❖ predict missing characters using both local and global context.
  - ❖ attribute authorship, region, and date based on subtle linguistic and stylistic cues found across entire inscriptions.

Activity Two, Part 2  
Letters of Mary Shelley  
**Thematic & Sentimental Evolution**

## How do the LLMs Compare

I asked DeepSeek, ChatGPT4, and Claude given a set of parameters to produce comparisons. What follows in the next three slides is their opinions.

# Constitutional AI

- ❖ Claude is an example of a LLM with an embedded Constitution.
- ❖ Constitutional AI uses a set of written principles to guide model behavior.
- ❖ After the model generates an initial response to a prompt it evaluates and revises its own responses using the constitution so the response aligns better with its principles.

# Claude AI's Comparison of itself DeepSeek R1 and ChatGPT4 (at 1 June 2025)

Feature	Claude Sonnet 4	DeepSeek R1	ChatGPT-4 (GPT-4o)
<b>Instruction Following Ability</b>	<b>Excellent</b> - Known for precise adherence to complex, multi-step instructions and structured formatting requirements. Strong performance on tasks requiring specific output formats like NER tagging.	<b>Very Good</b> - Shows strong instruction following with chain-of-thought reasoning. Particularly effective for step-by-step logical tasks and mathematical problems.	<b>Good</b> - Generally follows instructions well but can sometimes deviate from specific formatting requirements or complex multi-step processes.
<b>Factual Accuracy</b>	<b>High</b> - Generally reliable with factual information, though like all LLMs can experience hallucinations. Strong grounding in training data through Constitutional AI training.	<b>High</b> - Recent improvements have reduced hallucinations significantly. Strong performance on factual benchmarks, though being newer means less extensive real-world validation.	<b>Good</b> - Reasonable factual accuracy but known to have hallucination issues, particularly with recent events or specialized domains.
<b>Architectural Design &amp; Size</b>	<b>Proprietary</b> - Exact architecture undisclosed by Anthropic. Estimated to be a large transformer model (likely 100B+ parameters). Not open-source. Optimized for safety and alignment.	<b>Open-Source</b> - Multiple variants: R1-Zero (1.5B), R1-Distill (1.5B, 7B, 8B, 14B, 32B, 70B), R1 (671B parameters). Mixture-of-Experts architecture. Full model weights available.	<b>Proprietary</b> - GPT-4o architecture details not fully disclosed. Estimated multimodal transformer with significant parameter count. Not open-source. Optimized for speed and multimodal capabilities.
<b>Zero-shot and Few-shot Performance</b>	<b>Excellent</b> - Strong zero-shot performance across diverse tasks. Particularly effective with few-shot examples for complex reasoning and analysis tasks.	<b>Very Good</b> - Demonstrates strong zero-shot reasoning capabilities, especially in mathematics and coding. Chain-of-thought approach enhances few-shot performance significantly.	<b>Good</b> - Solid zero-shot performance, though may require more examples for optimal results on specialized tasks compared to Claude.
<b>Output Structure and Format Control</b>	<b>Excellent</b> - Superior at maintaining consistent output formats, JSON structures, and following precise formatting instructions. Low "noisy output" tendency.	<b>Good</b> - Recent updates include improved JSON output support and structured formatting. Chain-of-thought can sometimes add verbosity but generally maintains requested structure.	<b>Fair to Good</b> - Can produce structured outputs but less consistent than Claude. May require additional prompting to maintain format consistency.
<b>Multilingual and Domain Robustness</b>	<b>Good</b> - Handles multiple languages reasonably well. Strong performance on domain-specific content including technical, legal, and academic texts. Good with OCR-affected documents.	<b>Good</b> - Developed by Chinese company with strong multilingual capabilities, particularly for Asian languages. Good domain robustness across technical fields.	<b>Good</b> - Solid multilingual support across major languages. Reasonable domain robustness though may struggle with highly specialized or historical texts.
<b>Alignment, Safety, and Transparency</b>	<b>Excellent</b> - Trained using Constitutional AI (CAI) methodology. High focus on safety and alignment. Clear content policies and refusal behaviors. Moderate transparency about training approaches.	<b>Good</b> - Open-source nature provides model transparency. Safety measures implemented but less extensively documented than Claude. Being newer, safety track record is still developing.	<b>Good</b> - Uses RLHF for alignment. Established safety measures and content policies. Moderate transparency about training methods. Well-established track record but some concerns about consistency.



# Claude AI's Comparison of itself DeepSeek R1 and ChatGPT4 (at 1 June 2025)

Parameter	Claude Sonnet 4	DeepSeek R1	ChatGPT-4 (GPT-4o)
<b>Document Handling</b>	<b>Excellent</b> - Excels at handling long documents, processing up to 100K tokens in a single session. Superior at analyzing complex documents, PDFs, and structured text. Strong performance with academic papers, legal documents, and technical manuals.	<b>Good</b> - Handles documents well with strong analytical capabilities. Good at extracting information from structured documents. Less extensively tested on very long documents compared to Claude.	<b>Good</b> - Solid document processing capabilities. Can handle various file formats including PDFs and images. Good at extracting information but may struggle with very long documents.
<b>Ability to Summarize</b>	<b>Excellent</b> - Outstanding at creating concise, accurate summaries while preserving key information. Excellent at multi-level summarization (executive summaries, detailed summaries, bullet points). Strong at maintaining context across long documents.	<b>Very Good</b> - Strong summarization capabilities with logical structure. Good at identifying key points and maintaining coherence. Chain-of-thought approach helps in creating well-reasoned summaries.	<b>Good</b> - Decent summarization but can sometimes miss nuanced details or include irrelevant information. Generally reliable for standard summarization tasks.
<b>Named Entity Recognition (NER)</b>	<b>Excellent</b> - Superior performance in identifying and categorizing entities (persons, locations, organizations, dates, etc.). Excellent at following specific NER tagging formats like <PER>, <LOC>, <ORG>. Strong accuracy in complex, ambiguous contexts.	<b>Good</b> - Solid NER capabilities with good accuracy for standard entity types. Reasonable performance on specialized domains. May require more specific prompting for complex tagging schemes.	<b>Fair to Good</b> - Basic NER functionality works well for common entities. May struggle with specialized domains or complex entity relationships. Less consistent with specific formatting requirements.
<b>Precision in Following Task Instructions</b>	<b>Excellent</b> - Industry-leading performance in following complex, multi-step instructions precisely. Exceptional at maintaining formatting requirements, adhering to constraints, and following specific protocols without deviation.	<b>Very Good</b> - Strong instruction following with step-by-step reasoning approach. Good at complex tasks but may occasionally add unnecessary explanations or deviate slightly from strict formatting.	<b>Good</b> - Generally follows instructions but may require more explicit guidance for complex tasks. Can sometimes interpret instructions creatively rather than literally.
<b>Integration of Tools</b>	<b>Very Good</b> - Strong tool integration capabilities with function calling, web search, and document processing. Well-designed API for integration with external systems. Good at maintaining context across tool interactions.	<b>Limited</b> - While it lacks a dedicated plugin system, its open architecture allows developers to build custom tools around it. Open-source nature enables custom integrations but requires more development work.	<b>Excellent</b> - Extensive ecosystem with plugins, GPTs, and API integrations. Built-in web browsing, image generation (DALL-E), and code interpreter. Well-established third-party integrations.
<b>Reasoning</b>	<b>Excellent</b> - Advanced reasoning capabilities with strong logical consistency. Excellent at complex analysis, causal reasoning, and multi-step problem solving. Strong performance on reasoning benchmarks.	<b>Excellent</b> - Exhibits distinct strengths in addressing computational challenges and mathematical reasoning. Chain-of-thought approach provides transparent reasoning process. Strong performance on mathematical and logical reasoning tasks.	<b>Good</b> - Solid reasoning abilities for general tasks. May struggle with very complex logical chains or mathematical proofs. Reasoning can sometimes be inconsistent across similar problems.
<b>Response Times</b>	<b>Fast</b> - Excels at tasks demanding rapid responses, like knowledge retrieval. Generally fast response times for most queries. May be slightly slower for very complex reasoning tasks.	<b>Moderate</b> - Response times vary by model size. Larger models (671B) may be slower but more capable. Generally reasonable for most applications. Chain-of-thought reasoning may add processing time.	<b>Fast</b> - Generally quick response times. Opus delivers similar speeds to Claude 2 and 2.1, but with much higher levels of intelligence. Optimized for speed in conversational interactions.
<b>Multimodal Capabilities</b>	<b>Limited</b> - Claude 3 is multimodal: it can process image and audio content alongside text-based prompts, but Claude lacks ChatGPT's extensive multimodal capabilities. Although Claude has sufficient vision capabilities to analyze uploaded files, including images and PDFs, it does not support image generation, voice interaction or web browsing.	<b>Text-Only</b> - Primarily focused on text processing. Limited multimodal capabilities. Strong in text-based reasoning and analysis but lacks image or audio processing features.	<b>Excellent</b> - Comprehensive multimodal support including text, images, voice, and file processing. Can generate images via DALL-E, process audio, and handle various file formats. Built-in web browsing capabilities.

## ChatGPT's Comparison of itself Claude and DeepSeek R1 (at 1 June 2025)

Feature	ChatGPT-4 (OpenAI)	Claude (Opus, Anthropic)	DeepSeek (V2 / Instruct)
<b>Instruction Following</b>	Excellent: highly optimized for instruction compliance via RLHF	Excellent: excels with clear prompts, aided by Constitutional AI	Strong: well-tuned for structured prompts, sometimes verbose
<b>Factual Accuracy</b>	Very high, especially GPT-4-turbo with browsing	High, with strong reasoning and summarization	Good, but variable—some factual drift observed
<b>Architecture &amp; Size</b>	Proprietary; ~1.76T MoE (GPT-4-turbo, unconfirmed)	Proprietary; MoE system (~Claude 3 Opus likely 200B+ params)	Open-source; 16B–33B sizes; some models use dense + MoE hybrid
<b>Zero/Few-shot Performance</b>	Excellent: strong even without examples	Excellent: handles unseen tasks smoothly	Moderate–Good: needs more examples for niche domains
<b>Output Format Control</b>	High: consistent structure; good for code and tagging	High: predictable formatting; minimal noisy output	Medium: some variability in formatting; better in chat mode
<b>Multilingual &amp; Domain Robustness</b>	High: strong across languages and technical domains	High: performs well on multilingual and long texts	Medium: supports Chinese/English best; weaker on historical texts
<b>Alignment &amp; Safety</b>	RLHF-aligned; customizable via system messages	Constitutional AI: safe, honest, and transparent	Developing: safety filters improving; less mature than others

## ChatGPT's Comparison of itself Claude and DeepSeek R1 (at 1 June 2025)

Parameter	ChatGPT-4 (OpenAI)	Claude Opus (Anthropic)	DeepSeek (V2/Instruct)
<b>Document Handling</b>	Strong (100K+ tokens in GPT-4-turbo)	Exceptional (up to 200K tokens)	Limited (context size smaller; ~32K in top models)
<b>Summarization</b>	Very strong; handles dense and domain-	Excellent; known for clarity and coherence	Variable quality; summaries may be verbose or shallow
<b>Named Entity Recognition</b>	Strong with few-shot prompts; format stable	Cleanest outputs, especially for historical NER	Noisier outputs; format inconsistency
<b>Instruction Precision</b>	Very high; responds well to structured prompts	High fidelity to formatting and tone	Moderate; occasionally deviates from format
<b>Tool Integration</b>	Extensive: code interpreter, browsing, plugins	None publicly available	No tools; manual workflow
<b>Reasoning</b>	Excellent; good with chains of thought and	Strong; especially consistent in long-form reasoning	Decent; good in coding, weaker in abstract reasoning
<b>Response Time</b>	Moderate (especially with complex tasks or	Slower for large contexts but steady	Fastest among the three in most scenarios
<b>Multimodal Capabilities</b>	Yes (image input available in GPT-4-turbo)	Multimodal planned but not yet released	Text-only

# DeepSeek R1 's Comparison of itself Claude and ChatGPT (at 1 June 2025)

Parameter	DeepSeek-V3	Claude 3 (Opus/Sonnet)	ChatGPT-4 (GPT-4-turbo)
<b>Instruction Following</b>	Strong, follows structured formats well (e.g., NER tagging).	Excellent, excels in complex multi-step reasoning and nuanced prompts.	Very strong, but sometimes overly verbose or misses subtle formatting.
<b>Factual Accuracy</b>	High, but less verified than GPT-4/Claude. Fewer hallucinations.	Best-in-class, low hallucination rate, strong grounding in sources.	High but occasionally hallucinates on obscure topics.
<b>Architecture &amp; Size</b>	Unknown exact size (likely MoE-based), not open-source.	Mixture-of-Experts (MoE), ~tens of billions, partially efficient.	~1.8T parameters (rumored sparse MoE), closed-source.
<b>Zero/Few-shot Learning</b>	Strong, adapts quickly to examples.	Exceptional, excels in few-shot in-context learning.	Very good, but sometimes requires more explicit guidance.
<b>Output Structure Control</b>	Predictable, minimal noise in structured tasks.	Highly controllable, excels in JSON/XML/formatting.	Good, but can over-explain or deviate slightly.
<b>Multilingual &amp; Domain Robustness</b>	Strong in English & Chinese, decent in others.	Broad multilingual support, handles niche domains well.	Very robust, best for non-English among the three.
<b>Alignment &amp; Safety</b>	Conservative, avoids harmful content.	Constitutional AI, strong ethical alignment.	RLHF-tuned, but sometimes overly cautious.

# DeepSeek R1 's Comparison of itself Claude and ChatGPT (at 1 June 2025)

Parameter	DeepSeek-V3	Claude 3 (Opus/Sonnet)	ChatGPT-4 (GPT-4-turbo)
<b>Document Handling</b>	Strong, supports long documents (up to 128K tokens).	Best-in-class (200K context), excels at parsing large files.	Very good (128K context), but slightly weaker than Claude on very long docs.
<b>Ability to Summarize</b>	High-quality summaries, concise and accurate.	Exceptional at abstractive summaries, retains key details.	Strong, but can sometimes be verbose.
<b>Name Entity Recognition (NER)</b>	Accurate, follows tagging rules well.	Best at structured NER, minimal errors.	Good, but may occasionally miss strict formatting.
<b>Precision in Following Instructions</b>	Very good, adheres closely to prompts.	Best, extremely reliable for complex tasks.	Strong, but may over-explain or deviate slightly.
<b>Integration of Tools</b>	Supports some API integrations, but less mature than OpenAI.	Strong with tool use (e.g., web search, calculators).	Best, due to OpenAI's extensive plugin/API ecosystem.
<b>Reasoning</b>	Strong logical and analytical reasoning.	Top-tier, excels in multi-step and nuanced reasoning.	Excellent, but Claude 3 Opus is slightly better.
<b>Response Times</b>	Fast, optimized for efficiency.	Fast (Sonnet), Opus can be slower due to size.	Very fast (GPT-4-turbo optimized for speed).

# Takeaways ChatGPT-4 (OpenAI)

## ❖ Strengths:

- ❖ General-purpose performance across domains (NER, summarization, reasoning).
- ❖ Working with multilingual or historical contexts requiring flexibility (and do not have a specialist tool).
- ❖ You require consistent formatting for structured outputs (e.g., XML tags, JSON).
- ❖ You want plugins, browsing, or code execution (with GPT-4-turbo in ChatGPT Plus).
- ❖ Mature API ecosystem

## ❖ Weaknesses

- ❖ Limited transparency about architecture and training data.
- ❖ Can be more expensive than open models.
- ❖ Over-sanitizes or avoid sensitive content due to alignment filters.

# Takeaways DeepSeek

## ❖ Strengths:

- ❖ Transparent chain-of-thought reasoning process
- ❖ Open access, fast responses, or local model use but there is a tradeoff in precision and formatting.
- ❖ Exceptionally strong in terms of math, coding, and reasoning benchmarks.
- ❖ Cost efficiency (e.g., 2025 \$0.96 per 1M tokens).

## ❖ Weaknesses

- ❖ Less stable format control outputs may be noisy and require post-processing.
- ❖ Multilingual uses beyond English and Chinese (i.e., less effective for historical/low-resource language texts).
- ❖ Weaker factual accuracy and safety filters compared to GPT-4 or Claude.

# Takeaways Claude

## ❖ Strengths:

- ❖ Best for long documents, summarization, and high-precision NER in sensitive/historical contexts.
- ❖ Precise instruction following and structured outputs.
- ❖ Long-context reasoning of upto 200K tokens which support analysis of large docs).
- ❖ You value ethically aligned behavior and want models trained via Constitutional AI.
- ❖ In Coding Claude Sonnet 4 currently leads in programming capabilities.

## ❖ Weaknesses

- ❖ Slightly more verbose in some settings (e.g., may overexplain).
- ❖ No API fine-tuning, and less plugin support than OpenAI.
- ❖ Cost higher on average than ChatGPT or DeepSeek.

# Takeaways General

- ❖ ChatGPT-4: Good all-around document processing with multimodal file support
- ❖ ChatGPT-4: Most mature ecosystem with extensive plugin support and third-party integrations
- ❖ ChatGPT-4: for Creative & Multimodal (e.g., image generation, voice interaction)
- ❖ DeepSeek R1: Strong analytical capabilities but less tested on very long documents
- ❖ DeepSeek R1: for Development & Math (e.g., transparent reasoning, cost-effective)
- ❖ Claude: Unmatched precision in following complex instructions and formatting requirements
- ❖ Claude: for Research & Academia (e.g., document analysis, precise citations)

Use Case	Best Choice
Clean structured outputs for NER/summarization	<b>Claude Opus</b>
General multilingual and robust NLP tasks	<b>ChatGPT-4</b>
Open-source, cost-effective customization	<b>DeepSeek Instruct</b>
Long documents and context-sensitive reasoning	<b>Claude Opus</b>
Plugin integrations, API flexibility	<b>ChatGPT-4</b>
Chinese bilingual tasks or local deployment	<b>DeepSeek</b>

# Claude's Advice (1 June 2025)

The comparison table above provides a comprehensive analysis of these three leading AI models across the seven key features you specified. Here are some additional insights based on the recent search results:

## Recent Developments:

- DeepSeek R1 has shown significant improvements in benchmark performance, enhanced front-end capabilities, and reduced hallucinations. It rivals or even surpasses OpenAI o1 in math, coding, and reasoning benchmarks [DataCamp](#) [Bentoml](#)
- Claude Sonnet 4 is currently considered superior to GPT-4o in coding abilities according to programmer reviews [Claude vs. ChatGPT: What'...](#)
- DeepSeek R1 offers competitive performance at a significantly lower cost of \$0.96 per 1M tokens with a MMLU score of 0.849 [DeepSeek R1 0528 \(May '2...](#)

**Choosing the Right Model:** The choice ultimately depends on your specific needs. Claude Sonnet 4 excels in structured tasks and reliability, DeepSeek R1 offers excellent value and transparency as an open-source option, while ChatGPT-4 provides a mature ecosystem with multimodal capabilities. For tasks requiring precise instruction following and structured outputs, Claude tends to be the preferred choice, while DeepSeek R1 is particularly strong for mathematical and reasoning tasks at a lower cost.

# Activity Two, Part 3

## Letters of Mary Shelley

# Prompt Engineering & Meta-Comparison

# HNER Instruct-style LLMs and Historical Texts

- ❖ Investigates performance of open-access Instruct-style LLMs (e.g., LLaMA-2, LLaMA-3, Mixtral, Zephyr) on Historical Named Entity Recognition (HNER) on noisy, multilingual historical texts.
- ❖ Three public historical corpora: HIPE, NewsEye, and AJMC
- ❖ Compares two prompting strategies: deductive prompting (using annotation guidelines) and inductive prompting (few-shot examples)
- ❖ Models are evaluated on strict and fuzzy boundary matching for precision, recall, and F1-score; 128 experiments were conducted in total

Carlos-Emiliano González-Gallardo, Hanh Thi Hong Tran, Ahmed Hamdi, Antoine Doucet, Giorgio Maria Di Nunzio, Mickaël Coustaty, Francesco Gelati, et al. "Leveraging Open Large Language Models for Historical Named Entity Recognition." In *Linking Theory and Practice of Digital Libraries*, 15177:379–95. Switzerland: Springer, 2024. [https://doi.org/10.1007/978-3-031-72437-4\\_22](https://doi.org/10.1007/978-3-031-72437-4_22).



# Corpora Summary (1)

- ❖ AJMC (Ajax Multi-Commentary Corpus)
  - ❖ Content: 19th-century classical commentaries on Sophocles' Ajax
  - ❖ Languages: English, French, German (with Ancient Greek code-switching)
  - ❖ Entity Types: PER (person), LOC (location), ORG (organization), BIBLIO (bibliographic references to primary/secondary literature)
  - ❖ Focus: Entities tied to scholarly referencing and classical scholarship

González-Gallardo, et al 2024

# Corpora Summary (2)

- ❖ HIPE (CLEF-HIPE-2022 Historical Newspaper Corpus)
  - ❖ Content: Newspapers from Switzerland, Luxembourg, and the U.S. (19th–20th centuries)
  - ❖ Languages: English, French, German
  - ❖ Entity Types: PER, LOC, ORG, DATE (temporal expressions), HUMPROD (human productions like books or artworks) and Domain-specific subtypes (e.g., social roles embedded in PER)
  - ❖ Focus: General and cultural heritage NER, especially for OCR-impaired texts

González-Gallardo, et al 2024

# Corpora Summary (3)

- ❖ NewsEye
  - ❖ Content: Newspapers from France and Austria (1854–1946)
  - ❖ Languages used: French and German
  - ❖ Entity Types: PER, LOC, ORG, DATE & other domain-specific types based on article
  - ❖ Focus: Diverse press content from political, cultural, and social domains

# NER and HNER

## ❖ NER (Named Entity Recognition)

- ❖ Automatically identifies and classifies entities in text (e.g., people, places, organizations, dates).
- ❖ Works well on modern, clean, and well-edited texts (e.g., news articles, social media, websites).
- ❖ Contemporary language data using standard orthography and vocabulary.

## ❖ HNER (Historical Named Entity Recognition)

- ❖ Same core task as NER, but applied to historical documents.
- ❖ Unique challenges:
  - ❖ Spelling variation over time (e.g., "Britania" vs. "Britain"),
  - ❖ Outdated language and grammar,
  - ❖ Orthographic (e.g., the long s, written f; or capitalization or punctuation)
  - ❖ OCR errors from digitized scans of old texts
  - ❖ Named entity categories with historical context (e.g., titles “Archduke” or “Magistrate”)
  - ❖ Multilingual and code-switched texts, often in obsolete dialects

Feature	NER	HNER
Text type	Contemporary	Historical (19th c., 18th c., etc.)
Spelling consistency	High	Low (archaic forms, OCR noise)
Language variation	Standardized modern language	Outdated, dialectal, multilingual
Named entity scope	General	Often domain-specific (e.g., titles, roles)
Performance	Higher with modern models	Lower unless specifically adapted

# Inductive vs Deductive Prompting

- ❖ In Inductive Prompting the LLM is shown a few labeled examples. It is expected to generalize from them to new cases.
- ❖ In Deductive Prompting LLM given the same annotation guidelines that would be provided to human annotators
- ❖ Fundamental Difference:
  - ❖ Inductive reflects how LLMs learn from patterns (few-shot learning).
  - ❖ Deductive tests how well LLMs can follow instructions (think IKEA Instructions)

# Inductive Prompting

- ❖ How it works:
  - ❖ The prompt includes several annotated sentences (e.g., with <PER>, <LOC> tags).
  - ❖ Each example illustrates the entity types the model should extract.
  - ❖ The model is then given a new sentence and asked to tag entities in the same way.

# Inductive Prompting

- ❖ Strengths:
  - ❖ Leverages the LLMs pattern recognition abilities.
  - ❖ Requires no explanation of annotation rules—just examples.
- ❖ Weaknesses:
  - ❖ May fail if examples don't represent the full variety of the data.
  - ❖ More prone to errors in historical or noisy texts.
  - ❖ Dependent upon carefully selected examples.

# Inductive Prompting

You are a NER system. Tag the entities as <PER>, <LOC>, <ORG>, or <DATE>.

Example 1: "Queen <PER>Victoria</PER> visited <LOC>India</LOC> in <DATE>1875</DATE>."

Example 2: "<ORG>The Times</ORG> reported on <DATE>March 3, 1901</DATE>."

Now tag this sentence:

"President Kennedy met Queen Elizabeth in London on June 5, 1961."

"President <PER>Kennedy</PER> met <PER>Queen Elizabeth</PER> in <LOC>London</LOC> on <DATE>June 5, 1961</DATE>."

# Inductive Prompting

You are a NER system. Tag the entities as <PER>, <LOC>, <ORG>, or <DATE>.

Example 1: "Queen <PER>Victoria</PER> visited <LOC>India</LOC> in <DATE>1875</DATE>."

Example 2: "<ORG>The Times</ORG> reported on <DATE>March 3, 1901</DATE>."

Now tag this sentence:

"President Kennedy met Queen Elizabeth in London on June 5, 1961."

"President <PER>Kennedy</PER> met <PER>Queen Elizabeth</PER> in <LOC>London</LOC> on <DATE>June 5, 1961</DATE>."

is that correct you tagged Kennedy as a person without his title, but you tagged Elizabeth with her title.

Excellent observation—and you're right to question the inconsistency.

## ✓ Here's the issue:

In the sentence:

"President Kennedy met Queen Elizabeth in London on June 5, 1961."

- You **should either** include the titles consistently in both person tags or exclude them consistently—**depending on the annotation guidelines.**

## ✓ Corrected tagging (with consistent inclusion of titles):

"<PER>President Kennedy</PER> met <PER>Queen Elizabeth</PER> in <LOC>London</LOC> on <DATE>June 5, 1961</DATE>."

# Inductive Prompting

now tag the following:

"President John F. Kennedy met with Pope Paul VI in Rome at the Vatican on July 2, 1963"

Sure—using consistent annotation that includes titles and honorifics as part of the `<PER>` entity (as per the HIPE-style guidelines), here is the correctly tagged sentence:

```
"<PER>President John F. Kennedy</PER> met with <PER>Pope Paul VI</PER> in <LOC>Rome</LOC> at the <LOC>Vatican</LOC> on <DATE>July 2, 1963</DATE>."
```

# Inductive Prompting

Let's introduce a tag for "titles" <TITL>. Can you retag the sentence with this new tag as well.

Certainly. With the new <TITL> tag introduced to separately mark **titles and honorifics** (e.g., "President", "Pope"), we'll isolate them from the <PER> (person) tag.

Here is the revised tagging:

```
"<TITL>President</TITL> <PER>John F. Kennedy</PER> met with <TITL>Pope</TITL> <PER>Paul VI</PER> in <LOC>Rome</LOC> at the <LOC>Vatican</LOC> on <DATE>July 2, 1963</DATE>."
```

This reflects a more granular annotation, separating **role/title** from the **person's name**, which can be useful for downstream tasks like title normalization or role extraction.

# Inductive Prompting

Tag these sentences: As a Senator John F Kennedy meet Pope Pius XII on September 15, 1955 Castel Gandolfo. Kennedy had meet the Pope Pius XII earlier in 1937 and 1939 in Rome at the Vatican.

Using the tag set with `<PER>`, `<LOC>`, `<ORG>`, `<DATE>`, and the newly introduced `<TITL>` for titles and honorifics, here's the properly tagged version of the two sentences:

## Tagged Sentences:

As a `<TITL>Senator</TITL>` `<PER>John F. Kennedy</PER>` met `<TITL>Pope</TITL>` `<PER>Pius XII</PER>` on `<DATE>September 15, 1955</DATE>` `<LOC>Castel Gandolfo</LOC>`. `<PER>Kennedy</PER>` had met the `<TITL>Pope</TITL>` `<PER>Pius XII</PER>` earlier in `<DATE>1937</DATE>` and `<DATE>1939</DATE>` in `<LOC>Rome</LOC>` at the `<LOC>Vatican</LOC>`.

# Inductive Prompting

tag the following sentence: Two faculty from Vassar College Dr John Wilson and the Mary Watson Professor of History traveled to the Vatican Archives to study the papers of Pope Pius XII.

Using the tags <PER> (person), <TITL> (title/honorific), <ORG> (organization), and <LOC> (location), here's the annotated version of the sentence:

Two faculty from <ORG>Vassar College</ORG> <TITL>Dr</TITL> <PER>John Wilson</PER> and the <TITL>Mary Watson Professor of History</TITL> traveled to the <LOC>Vatican Archives</LOC> to study the papers of <TITL>Pope</TITL> <PER>Pius XII</PER>.

Is the Vatican Archives a place or an Organisation

"Two faculty from <ORG>Vassar College</ORG> <TITL>Dr</TITL> <PER>John Wilson</PER> and the <TITL>Mary Watson Professor of History</TITL> traveled to the <ORG>Vatican Archives</ORG> to study the papers of <TITL>Pope</TITL> <PER>Pius XII</PER>."

# Deductive Prompting

- ❖ How it works:
  - ❖ The prompt explains what each entity type means (e.g., "a person includes full name, role, and title").
  - ❖ It may include boundary rules (e.g., where entity begins and ends) and examples from the original annotation manual.
    - ❖ In an NER task Boundary Rules guide the handling of nested or overlapping entities, multi-word expressions (NYC), punctuation and parenthesis and so forth.
    - ❖ Strict matching vs fuzzy matching
  - ❖ The model uses these rules to decide which tokens to label and how.

# Boundary Rules

- ❖ If the boundary rules instructed the LLM to tag "person name" and "title or honorific" with a single <PER> tag then one would expect the following
- ❖ The result to be similar to what I annotated manually below:
  - ❖ "By virtue of a bill of sale issued by <PER>Tom Scarlett, Circuit Court Clerk for Putnam County</PER>..."
- ❖ In the case of the LLM if it labeled a Correct Prediction (Strict Match):
  - ❖ "By virtue of a bill of sale issued by <PER>Tom Scarlett, Circuit Court Clerk for Putnam County</PER>..."
  - ❖ It delivered a perfect match and counts under strict and fuzzy
- ❖ Imperfect Prediction (Fuzzy Match Only):
  - ❖ "By virtue of a bill of sale issued by <PER>Tom Scarlett</PER>, Circuit Court Clerk for Putnam County..."
  - ❖ It fails strict match but get points under fuzzy match for correct core, but incomplete span

# Deductive Prompting

- ❖ Strengths:
  - ❖ Aligns more closely with human annotation logic.
  - ❖ Encourages rule-based consistency in predictions.
- ❖ Weaknesses:
  - ❖ Long or complex guidelines can exceed token limits.
  - ❖ Models may not follow instructions precisely, especially in low-resource or noisy cases.

# Deductive Prompting

You are a Name Entity Recognition (NER) system. Follow these annotation rules:

- A PERSON includes full name.
- A TITLE includes role and honorifics.
- An ORGANIZATION includes political, governmental, and corporate entities.
- A LOCATION includes any geographic place.
- A DATE includes full or partial temporal expressions.

Understood. Please provide the sentence(s) you'd like annotated, and I will tag them using the specified NER categories: <PERSON>, <TITLE>, <ORGANIZATION>, <LOCATION>, and <DATE>.

# Deductive Prompting

"President Kennedy met Queen Elizabeth in London on June 5, 1961."

"<TITLE>President</TITLE> <PERSON>Kennedy</PERSON> met <TITLE>Queen</TITLE>  
<PERSON>Elizabeth</PERSON> in <LOCATION>London</LOCATION> on <DATE>June 5, 1961</  
DATE>."

# Deductive Prompting

"President John F. Kennedy met with Pope Paul VI in Rome at the Vatican on July 2, 1963"

"<TITLE>President</TITLE> <PERSON>John F. Kennedy</PERSON> met with <TITLE>Pope</TITLE> <PERSON>Paul VI</PERSON> in <LOCATION>Rome</LOCATION> at the <ORGANIZATION>Vatican</ORGANIZATION> on <DATE>July 2, 1963</DATE>"

Is the Vatican an organisation or a location

The Vatican can be annotated as either an <ORGANIZATION> or a <LOCATION>, depending on the context:

- Use <ORGANIZATION> when referring to **the governing body**, such as the Holy See, the papacy, or the Catholic Church's central administration.
  - Example: "He worked for the Vatican on matters of diplomacy." → <ORGANIZATION>Vatican</ORGANIZATION>
- Use <LOCATION> when referring to **the physical place**, i.e., Vatican City.
  - Example: "They visited the Vatican on a tour of Rome." → <LOCATION>Vatican</LOCATION>

In your sentence:

"President John F. Kennedy met with Pope Paul VI in Rome at the Vatican on July 2, 1963"

"The Vatican" refers to the **physical place** where the meeting occurred. Therefore, the correct annotation is:  
<LOCATION>Vatican</LOCATION>.



correct the annotation please

Here is the corrected annotation based on your rules and the clarification:

"<TITLE>President</TITLE> <PERSON>John F. Kennedy</PERSON> met with <TITLE>Pope</TITLE>  
<PERSON>Paul VI</PERSON> in <LOCATION>Rome</LOCATION> at the <LOCATION>Vatican</  
LOCATION> on <DATE>July 2, 1963</DATE>"



# Deductive Prompting

"As a Senator John F Kennedy meet Pope Pius XII on September 15, 1955 Castel Gandolfo. Kennedy had meet the Pius XII earlier in 1937 and 1939 in Rome at the Vatican."

Here is the annotated version of your sentence according to the specified NER rules:

"As a <TITLE>Senator</TITLE> <PERSON>John F. Kennedy</PERSON> met <TITLE>Pope</TITLE> <PERSON>Pius XII</PERSON> on <DATE>September 15, 1955</DATE> at <LOCATION>Castel Gandolfo</LOCATION>. <PERSON>Kennedy</PERSON> had met <TITLE>Pope</TITLE> <PERSON>Pius XII</PERSON> earlier in <DATE>1937</DATE> and <DATE>1939</DATE> in <LOCATION>Rome</LOCATION> at the <LOCATION>Vatican</LOCATION>."

# Deductive Prompting

"Two faculty from Vassar College Dr John Wilson and the Mary Watson Professor of History traveled to the Vatican Archives to study the papers of Pope Pius XII."

Here is the annotated version of the sentence:

"Two faculty from <ORGANIZATION>Vassar College</ORGANIZATION> <TITLE>Dr</TITLE>  
<PERSON>John Wilson</PERSON> and the <TITLE>Mary Watson Professor of History</TITLE> traveled  
to the <ORGANIZATION>Vatican Archives</ORGANIZATION> to study the papers of <TITLE>Pope</  
TITLE> <PERSON>Pius XII</PERSON>."

# Deductive Prompting with DEEPSEEK

## <https://chat.deepseek.com/>

1. "President [PERSON Kennedy] met Queen [PERSON Elizabeth] in [LOCATION London] on [DATE June 5, 1961]."
2. "President [PERSON John F. Kennedy] met with [PERSON Pope Paul VI] in [LOCATION Rome] at the [LOCATION Vatican] on [DATE July 2, 1963]."
3. "As a [TITLE Senator] [PERSON John F Kennedy] meet [PERSON Pope Pius XII] on [DATE September 15, 1955] [LOCATION Castel Gandolfo]. He had meet the pope earlier in [DATE 1937] and [DATE 1939] in [LOCATION Rome] at the [LOCATION Vatican]."
4. "Two faculty from [ORGANIZATION Vassar College] [PERSON Dr John Wilson] and the [TITLE Mary Watson Professor of History] traveled to the [LOCATION Vatican Archives] to study the papers of [PERSON Pope Pius XII]."

# Deductive Prompting with DEEPSEEK

## <https://chat.deepseek.com/>



Here are the annotated sentences based on the given rules:

1. "<TITLE>President</TITLE> <PERSON>Kennedy</PERSON> met <TITLE>Queen</TITLE> <PERSON>Elizabeth</PERSON> in <LOCATION>London</LOCATION> on <DATE>June 5, 1961</DATE>."
2. "<TITLE>President</TITLE> <PERSON>John F. Kennedy</PERSON> met with <TITLE>Pope</TITLE> <PERSON>Paul VI</PERSON> in <LOCATION>Rome</LOCATION> at the <ORGANIZATION>Vatican</ORGANIZATION> on <DATE>July 2, 1963</DATE>."
3. "As a <TITLE>Senator</TITLE> <PERSON>John F Kennedy</PERSON> meet <TITLE>Pope</TITLE> <PERSON>Pius XII</PERSON> on <DATE>September 15, 1955</DATE> <LOCATION>Castel Gandolfo</LOCATION>. He had meet the pope earlier in <DATE>1937</DATE> and <DATE>1939</DATE> in <LOCATION>Rome</LOCATION> at the <ORGANIZATION>Vatican</ORGANIZATION>."
4. "Two faculty from <ORGANIZATION>Vassar College</ORGANIZATION> <TITLE>Dr</TITLE> <PERSON>John Wilson</PERSON> and the <TITLE>Mary Watson Professor of History</TITLE> traveled to the <ORGANIZATION>Vatican Archives</ORGANIZATION> to study the papers of <TITLE>Pope</TITLE> <PERSON>Pius XII</PERSON>."

# Deductive Prompting with Claude.ai

Here are the annotated sentences with named entity recognition:

**Sentence 1:** "[TITLE: President] [PERSON: Kennedy] met [TITLE: Queen] [PERSON: Elizabeth] in [LOCATION: London] on [DATE: June 5, 1961]."

**Sentence 2:** "[TITLE: President] [PERSON: John F. Kennedy] met with [TITLE: Pope] [PERSON: Paul VI] in [LOCATION: Rome] at [LOCATION: the Vatican] on [DATE: July 2, 1963]"

**Sentence 3:** "As a [TITLE: Senator] [PERSON: John F Kennedy] meet [TITLE: Pope] [PERSON: Pius XII] on [DATE: September 15, 1955] [LOCATION: Castel Gandolfo]. He had meet the [TITLE: pope] earlier in [DATE: 1937] and [DATE: 1939] in [LOCATION: Rome] at [LOCATION: the Vatican]."

**Sentence 4:** "Two faculty from [ORGANIZATION: Vassar College] [TITLE: Dr] [PERSON: John Wilson] and the [PERSON: Mary Watson] [TITLE: Professor of History] traveled to [LOCATION: the Vatican Archives] to study the papers of [TITLE: Pope] [PERSON: Pius XII]."

# Deductive Prompting with Claude.ai

**S** Could you do boundary tagging in <> and also in Sentence 4 the Vatican Archives is an organisation and not a location

Here are the annotated sentences with boundary tagging in <> and the correction for Sentence 4:

**Sentence 1:** "<TITLE>President</TITLE> <PERSON>Kennedy</PERSON> met <TITLE>Queen</TITLE> <PERSON>Elizabeth</PERSON> in <LOCATION>London</LOCATION> on <DATE>June 5, 1961</DATE>."

**Sentence 2:** "<TITLE>President</TITLE> <PERSON>John F. Kennedy</PERSON> met with <TITLE>Pope</TITLE> <PERSON>Paul VI</PERSON> in <LOCATION>Rome</LOCATION> at <LOCATION>the Vatican</LOCATION> on <DATE>July 2, 1963</DATE>"

**Sentence 3:** "As a <TITLE>Senator</TITLE> <PERSON>John F Kennedy</PERSON> meet <TITLE>Pope</TITLE> <PERSON>Pius XII</PERSON> on <DATE>September 15, 1955</DATE> <LOCATION>Castel Gandolfo</LOCATION>. He had meet the <TITLE>pope</TITLE> earlier in <DATE>1937</DATE> and <DATE>1939</DATE> in <LOCATION>Rome</LOCATION> at <LOCATION>the Vatican</LOCATION>."

**Sentence 4:** "Two faculty from <ORGANIZATION>Vassar College</ORGANIZATION> <TITLE>Dr</TITLE> <PERSON>John Wilson</PERSON> and the <PERSON>Mary Watson</PERSON> <TITLE>Professor of History</TITLE> traveled to <ORGANIZATION>the Vatican Archives</ORGANIZATION> to study the papers of <TITLE>Pope</TITLE> <PERSON>Pius XII</PERSON>."

# Noisy Outputs

- ❖ noisy outputs refer to the inconsistencies, deviations, or malformed responses produced by LLMs when generating responses for HNER tasks..
- ❖ “the generative nature of the Instruct models impacts the predictions that are produced,” often requiring post-processing to correct structural and content-related noise (Section 3.4).
- ❖ Noisy outputs are not just formatting errors--they arise from the challenges of controlling LLM behaviour especially when applied to OCR-impaired, orthographically variable historical texts.
- ❖ While Instruct models can identify many named entities, their generative tendencies often introduce noise that reduces precision, recall, and evaluation reliability.
- ❖ Post-Processing as a strategy. Historian-in-the-loop.

González-Gallardo, et al 2024

# Noisy Outputs Types

- ❖ Format Violations
  - ❖ Despite prompts specifying exact formats (e.g., tagging entities in-line with <PER>, <LOC>), LLMs may instead restate the sentence followed by a separate list of entities or add explanatory notes, making the output harder to align with gold standards.
- ❖ Incorrect or Inconsistent Entity Classes
  - ❖ Models may generate entity types not present in the guidelines (e.g., animals ideology, currency)
  - ❖ “LLMs tend to propose named entity classes that were not explicitly included in the prompts,” the authors observe (Section 4).
- ❖ Partial or Misaligned Entities
  - ❖ LLMs often recognize part of an entity, such as tagging only the person’s name but omitting their role, title, or separate first and last name. These boundary errors lower performance under strict evaluation but may be partially credited under fuzzy matching.
- ❖ Altered Input
  - ❖ Some outputs omit, reorder, or change parts of the original sentence, despite prompts requesting preservation of input structure.

# To recap, they found, Noisy results as

- ❖ Fails to follow the expected format (e.g., doesn't use <PER>, <LOC> tags as instructed)
- ❖ Includes extra commentary or explanations from the model
- ❖ Mislabeled entity types (e.g., uses 'production' instead of the expected 'humprod')
- ❖ Adds or drops parts of the input sentence
- ❖ Omits named entities
- ❖ Invents named entities

González-Gallardo, et al 2024

# Key Findings

- ❖ LLMs underperform compared to fine-tuned neural NER systems.
- ❖ Inductive prompting slightly outperforms deductive prompting ( $\sim 5\%$  gain).
- ❖ Multi-turn interaction (i.e., conversation style prompting strategy) reduces noisy outputs but doesn't increase accuracy.
- ❖ LLMs struggle with OCR errors, historical spelling variation, orthographic variation (e.g., historic spelling variation, capitalization), and inconsistent output formatting—especially with boundary accuracy and class adherence.
- ❖ Despite limitations, open LLMs can assist human annotators in historical NER workflows,
- ❖ Authors provide a reproducible framework and open-source toolkit for future experimentation.



# Activity Two, Part 4

## Letters of Mary Shelley

# **NER Extraction & Relationship Mapping**

Activity Two, Part 5  
Letters of Mary Shelley  
**Bias & Hallucination Detection Discussion**

# Challenges of Language

# LLMs and English

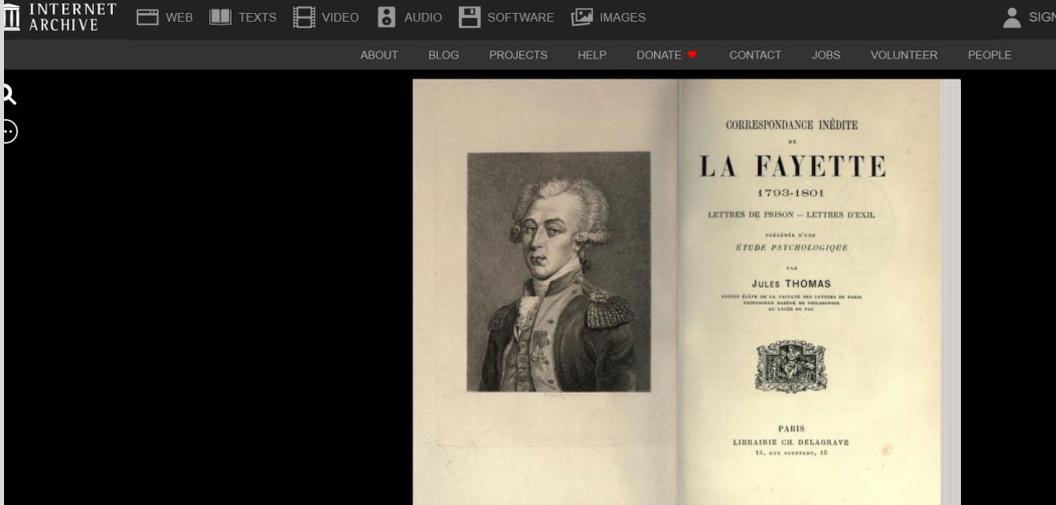
- ❖ Trained mainly on English Texts
  - ❖ Source availability of training data (e.g., news, Wikipedia, books)
  - ❖ Research bias arising from English-speaking team dominance in development of foundational LLMs
  - ❖ Market focus arising from individual and corporate demand.

# Consequences of Training

- ❖ Better performance in English in terms of fluency, accuracy, and nuance.
- ❖ Weaker performance in other languages: Especially in low-resource language.
- ❖ Bias propagation: English-language cultural, political, and historical viewpoints as a result often dominates dominate model outputs.

# LLMs Multilingual Models

mBERT, XLM-R, and GPT-4 have multilingual training, but as English disproportionate percentage of the training data model performance often declines less-represented languages.



## lettres de prison - lettres d'exil, précédée d'une Etude psychologique par Jules Thomas

by Lafayette, Marie Joseph Paul Yves Roch Gilbert Du Motier, marquis de, 1757-1834; Thomas, Jules, 1856-

Publication date [1903]  
Publisher Paris C. Delagrave  
Collection robarts; toronto; university\_of\_toronto  
Contributor Robarts - University of Toronto  
Language French  
Item Size 675.1M

26

Addeddate 2008-06-16 13:51:06  
Associated-names Thomas, Jules, 1856-  
Call number AFQ-2295  
Camera 1Ds  
Copyright-evidence Evidence reported by andrea@archive.org for item correspondancein00lafauoft on June 16, 2008: no visible notice of copyright; exact publication date unknown.  
Copyright-evidence-date 20080616135034  
Copyright-evidence-operator andrea@archive.org  
Copyright-region US  
Curatestate approved  
External-identifier urn:oclc:record:1042554929  
Foldoutcount 0  
Identifier correspondancein00lafauoft

[SHOW MORE](#)

Full catalog record [MARCXML](#)

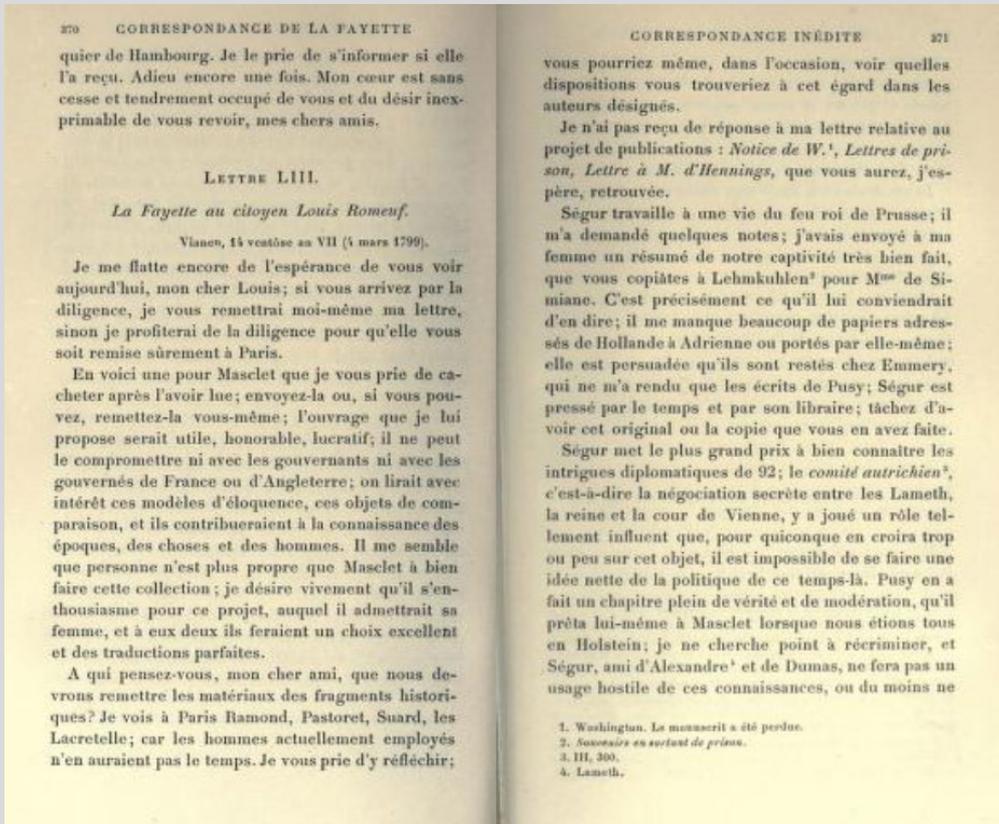
923 Views

1 Favorite

### DOWNLOAD OPTIONS

<a href="#">ABBY GZ</a>	1 file
<a href="#">CHOCR</a>	1 file
<a href="#">DAISY</a>	Temporarily Unavailable For users with print-disabilities
<a href="#">EPUB</a>	Temporarily Unavailable
<a href="#">FULL TEXT</a>	1 file
<a href="#">HOCR</a>	1 file
<a href="#">ITEM TILE</a>	1 file
<a href="#">OCR PAGE INDEX</a>	1 file
<a href="#">OCR SEARCH TEXT</a>	1 file
<a href="#">PAGE NUMBERS JSON</a>	1 file
<a href="#">PDF</a>	1 file
<a href="#">SCRIBE SCANDATA ZIP</a>	1 file
<a href="#">SINGLE PAGE PROCESSED JP2 ZIP</a>	1 file

# Lettre 53, a Louis Romeuf (Vianen, 14 ventôse an VII / 4 mars 1799)\*.



Letter from Marquis de Lafayette to Louis Romeuf, his trusted secretary and confidant, pushing forward the writing and publishing of ideas that emerged in the context of the French Revolution.

\* *Lafayette, Gilbert du Motier, marquis de. La Vie politique et militaire du général Lafayette. Correspondance*, édité par [éditeur non précisé]. Paris : Félix Alcan, 1903

Vianen, 14 ventôse an VII (4 mars 1799).

Je me flatte encore de l'espérance de vous voir aujourd'hui, mon cher Louis; si vous arrivez par la diligence, je vous remettrai moi-même ma lettre, sinon je profiterai de la diligence pour qu'elle vous soit remise sûrement à Paris.

En voici une pour Masclat que je vous prie de cacheter après l'avoir lue; envoyez-la ou, si vous pouvez, remettez-la vous-même; l'ouvrage que je lui propose serait utile, honorable, lucratif; il ne peut le compromettre ni avec les gouvernants ni avec les gouvernés de France ou d'Angleterre; on lirait avec intérêt ces modèles d'éloquence, ces objets de comparaison, et ils contribueraient à la connaissance des époques, des choses et des hommes. Il me semble que personne n'est plus propre que Masclat à bien faire cette collection; je désire vivement qu'il s'enthousiasme pour ce projet, auquel il admettrait sa femme, et à eux deux ils feraient un choix excellent et des traductions parfaites.

A qui pensez-vous, mon cher ami, que nous devons remettre les matériaux des fragments historiques? Je vois à Paris Ramond, Pastoret, Suard, les Lacre telle; car les hommes actuellement employés n'en auraient pas le temps. Je vous prie d'y réfléchir; VOUS pourriez même, dans l'occasion, voir quelles dispositions vous trouveriez à cet égard dans les auteurs désignés.

Je n'ai pas reçu de réponse à ma lettre relative au projet de publications: Notice de W. Lettres de prison, Lettre à M. d'Hennings, que vous aurez, j'espère, retrouvée.

Séjour travaille à une vie du feu roi de Prusse; il m'a demandé quelques notes; j'avais envoyé à ma femme un résumé de notre captivité très bien fait, que vous copiâtes à Lelinkuhlen<sup>^</sup> pour M<sup>\*\*\*</sup> de Simiane. C'est précisément ce qu'il lui conviendrait

d'en dire; il me manque beaucoup de papiers adressés de Hollande à Adrienne ou portés par elle-même; elle est persuadée qu'ils sont restés chez Emmerly,

qui ne m'a rendu que les écrits de Pusy; Séjour est pressé par le temps et par son libraire; tâchez d'avoir cet original ou la copie que vous en avez faite.

Séjour met le plus grand prix à bien connaître les intrigues diplomatiques de 92; le comité autrichien<sup>^</sup>, c'est-à-dire la négociation secrète entre les Lameth,

la reine et la cour de Vienne, y a joué un rôle tellement influent que, pour quiconque en croira trop ou peu sur cet objet, il est impossible de se faire une idée nette de la politique de ce temps-là. Pusy en a fait un chapitre plein de vérité et de modération, qu'il prêta lui-même à Masclat lorsque nous étions tous en Ilolstein; je ne cherche point à récriminer, et Séjour, ami d'Alexandre<sup>^</sup> et de Dumas, ne fera pas un usage hostile de ces connaissances, ou du moins ne lira que ce qui est nécessaire à l'histoire. Cependant je ne voudrais pas que cet écrit fut donné par vous-même à Séjour; il serait remis par moi si je vais bientôt à Paris, ou par Georges avec un billet confidentiel; ce que je vous demande, mon cher Louis, c'est de retrouver cet écrit chez Masclat.

Je voudrais bien avoir une copie de la lettre que j'écrivais pour mes compagnons au Directoire<sup>^</sup> ainsi que notre déclaration de Rochefort<sup>^</sup>; on m'a perdu tout cela, et ces deux pièces sont dans ce moment essentielles.

Il serait peut-être bien fait de donner la seconde à Séjour, dont l'ouvrage aura sûrement beaucoup de lecteurs dans tous les pays.

Je ne vous écris ce matin que sur ces objets, et vous embrasse de tout mon cœur. M<sup>\*\*</sup> de Simiane ne m'a pas accusé la réception d'un projet de note anglaise.

# Process

- ❖ Assessing the manuscript itself
- ❖ Transcription of the document
- ❖ Validating the transcription
- ❖ OCRing of the transcription
- ❖ Validating the OCRing potential  
correcting for mis-registered features

# ChatGPT 4 Summarize the letter -French

## Résumé de la lettre :

Lafayette informe Romeuf qu'il espère le voir bientôt et lui confie une lettre à transmettre à Masclat, contenant une proposition de projet littéraire : une anthologie de modèles d'éloquence et d'objets de comparaison, destinée à enrichir la compréhension des époques, des événements et des hommes. Il insiste sur le fait que ce travail serait utile, honorable et sans risque politique, et que Masclat et sa femme sont les mieux placés pour le mener à bien.

Il demande également à Romeuf de réfléchir à qui confier les matériaux des fragments historiques, évoquant plusieurs figures parisiennes comme Ramond, Pastoret, Suard et les Lacretelle. Lafayette revient sur un projet de publication non encore commenté (Notice de W., lettres de prison, lettre à M. d'Hennings), et mentionne que Ségur prépare une biographie du feu roi de Prusse. Il souhaite lui fournir des notes, notamment un résumé de sa captivité, mais certains papiers manquent — peut-être restés chez Emmerly.

Il souligne l'importance des intrigues diplomatiques de 1792 (le "comité autrichien") et évoque un chapitre écrit par Pusy à ce sujet, prêté à Masclat. Lafayette demande à Romeuf de le retrouver, mais précise qu'il ne doit pas le transmettre directement à Ségur, préférant le faire lui-même ou par l'intermédiaire de Georges.

Enfin, il réclame des copies d'une lettre adressée au Directoire au nom de ses compagnons de captivité, ainsi que de la déclaration de Rochefort, toutes deux perdues, et envisage d'en transmettre une à Ségur. Il termine en exprimant son affection et note que Madame de Simiane n'a pas accusé réception d'un projet de note anglaise.

# ChatGPT 4 Summarize the letter English

## Summary of the Letter:

### 1. Coordination of a Letter Delivery:

Lafayette expresses the hope that he might see Romeuf in person, or otherwise ensure that a letter is securely delivered to him in Paris.

### 2. Proposal for Masclat:

Lafayette includes a letter for Masclat (likely Félix-Jacques-Antoine Masclat, a journalist and translator), proposing a literary project involving the collection and translation of eloquent historical texts. He sees this work as useful, honorable, and profitable, and notes it would not compromise Masclat politically. He believes Masclat and his wife are well-suited for the task.

### 3. Historical Fragments Project:

Lafayette discusses the organization of historical materials and wonders who in Paris might be entrusted with them—mentioning Ramond, Pastoret, Suard, and the Lacretable brothers. He notes that those currently in government roles would lack the time for this.

### 4. Unanswered Letter:

He mentions a lack of response to a previous proposal involving the Notice of W., letters from prison, and a letter to M. d'Hennings—suggesting ongoing publishing plans.

### 5. Contribution to Ségur's Book:

Lafayette says that Ségur (likely Louis Philippe, comte de Ségur) is writing a biography of Frederick the Great and requested notes. Lafayette refers to a summary of his own captivity, copied by Romeuf at Lelinkuhlen, as potentially useful. However, some documents remain missing, likely left with Emmerly.

### 6. Diplomatic Intrigues of 1792:

Ségur is particularly interested in the secret negotiations between the Lameths, Queen Marie-Antoinette, and the Viennese court (called the "Austrian Committee"). Lafayette affirms that Pusy (Jean-Xavier Bureau de Pusy) wrote a balanced account of this, once lent to Masclat. Lafayette wants Romeuf to retrieve this text—but insists it should not be handed to Ségur directly by Romeuf, but via Lafayette or Georges, to ensure discretion.

### 7. Missing Declarations:

Lafayette wants a copy of a letter he wrote to the French Directory on behalf of his fellow prisoners, and a declaration from Rochefort—both now lost but crucial for his current work.

### 8. Note to Ségur:

He considers giving Ségur the Rochefort declaration, since the book will likely be widely read and this could enhance the historical account.

### 9. Closing:

He sends affectionate regards and notes that Madame de Simiane has not acknowledged receipt of an English-language note he had sent.

# DeepSeek Summary



## Summary of the Letter (200–300 words):

In this letter dated March 4, 1799, La Fayette writes to his friend Louis Romeuf, expressing hope to see him soon. If Romeuf arrives by coach, La Fayette will hand him the letter personally; otherwise, he will send it securely to Paris.

La Fayette encloses a letter for Masclet, urging Romeuf to seal it after reading. He proposes a literary project to Masclet—a collection of eloquent historical texts that would serve as valuable comparisons without offending French or English authorities. La Fayette believes Masclet, possibly assisted by his wife, is ideally suited for this task due to his skill in translation and selection.

He then asks Romeuf to consider who might best handle the compilation of historical fragments, suggesting figures like Ramond, Pastoret, Suard, or Lacretelle, as current officials lack the time. La Fayette also mentions unresolved matters, such as the lack of response regarding planned publications (*Notice de W., Lettres de prison, etc.*).

The letter touches on Ségur's ongoing biography of the late Prussian king, for which La Fayette provided notes. He seeks missing documents, possibly held by Emmerly or Masclet, detailing diplomatic intrigues of 1792, particularly involving the "Austrian Committee." While La Fayette trusts Ségur to use this information fairly, he prefers the material be delivered discreetly, either by himself or a confidant.

Finally, La Fayette requests copies of his past writings, including a letter to the Directory and the Rochefort declaration, which are now lost but crucial. He suggests Ségur's widely read work might be a suitable platform for the latter. Closing warmly, he notes that Mme de Simiane has not acknowledged receiving an English draft note.

(Word count: ~250)



# Claude Summary

**S** Please provide an english summary of the following French letter in about 200 words.

This letter from the Marquis de Lafayette to citizen Louis Romeuf, dated March 4, 1799, during Lafayette's exile in the Netherlands, discusses several literary and historical projects.

Lafayette asks Romeuf to deliver a letter to Masplet, proposing a collection project that would be "useful, honorable, and profitable" - likely involving translations of eloquent works that would serve as models for comparison and contribute to understanding different historical periods. He believes Masplet and his wife would be ideal collaborators for creating excellent selections and perfect translations.

Lafayette seeks advice on whom to entrust with materials for historical fragments, mentioning several potential candidates in Paris including Ramond, Pastoret, and Suard. He notes that currently employed officials wouldn't have time for such work.

The letter also discusses Ségur's biographical work on the late King of Prussia, for which Lafayette had provided notes. He requests help locating a summary of their captivity that he had written, which Romeuf had previously copied. Lafayette mentions missing papers that may be with Emmery.

Significantly, Lafayette discusses the sensitive topic of the "Austrian committee" - secret negotiations between the Lameth brothers, the Queen, and Vienna's court in 1792. He describes this as crucial for understanding the politics of that era, noting that Pusy had written a truthful chapter on the subject.

The letter concludes with requests for copies of his letter to the Directory and their Rochefort declaration.

# Motivation and Intent

## Claude Analysis of

S Provide an interpretation of motivation and intent

Lafayette exile 1799 Netherlands historical context

10 results ▾

Based on the historical context and the content of this letter, Lafayette's motivations and intentions reveal a complex strategic mindset of a political exile working to shape historical narratives and secure his legacy during a precarious period.

### Primary Motivations:

**Intellectual Legacy Management:** Having been imprisoned for over five years after fleeing France in 1792 (Gilbert du Motier, Marquis ...), Lafayette is deeply concerned with controlling how history will remember his role and the events he witnessed. His emphasis on finding the "right" people to handle historical materials suggests he wants sympathetic, competent voices to tell these stories.

**Financial Necessity:** The repeated emphasis on projects being "lucrative" indicates Lafayette, despite his noble status, faces financial pressures during his exile. The various publishing ventures represent both income opportunities and platforms for influence.

**Political Rehabilitation:** By 1799, Lafayette is positioning himself for eventual return to French political life. His careful navigation around sensitive topics (like the Austrian committee negotiations) shows someone aware that his words could be scrutinized by both French and foreign governments.

**Strategic Information Control:** Lafayette's handling of the "Austrian committee" materials reveals sophisticated political calculation. He acknowledges this information's explosive potential but wants it released through carefully chosen intermediaries (Ségur rather than himself) to avoid direct association with potentially damaging revelations.

**Network Preservation:** The letter demonstrates Lafayette's effort to maintain his intellectual and political network during exile, using trusted intermediaries like Romeuf to coordinate various projects and maintain relationships with influential figures in Paris.

His ultimate intent appears to be establishing a favorable historical record while generating income and maintaining political relevance for his eventual return to France.

# ChatGPT 4

## Interpretation:

This letter reveals Lafayette's **deep involvement in shaping historical memory** after the Revolution—through writing, selecting trusted collaborators, and cautiously managing sensitive materials. It shows his desire to **control the narrative** of key events like his captivity and 1792 diplomacy, and his concern with accuracy, discretion, and legacy.

# DeepSeek R1

Can you provide an interpretation in about 50 words.



This letter reveals La Fayette's efforts to shape historical narratives through trusted allies like Masclet and Romeuf. He emphasizes discretion in handling sensitive political documents, particularly regarding the 1792 diplomatic intrigues, while supporting Ségur's biography. His urgency in recovering lost writings underscores his desire to preserve his legacy amid post-revolutionary turmoil.

(50 words)

# Claude

**S** Provide an interpretation in 50 words.

Lafayette, in exile, strategically manages his historical legacy while seeking income through publishing projects. He carefully controls sensitive political information about the Austrian committee negotiations, positioning himself for eventual return to France by maintaining networks and ensuring sympathetic voices shape historical narratives of his controversial past.

**You are a Name Entity Recognition (NER) system.**  
 Follow these annotation rules:  
 - A PERSON includes full name.  
 - A TITLE includes role and honorifics.  
 - An ORGANIZATION includes political, governmental, and corporate entities.  
 - A LOCATION includes any geographic place.  
 - A DATE includes full or partial temporal expressions.  
 Use <-> for NER boundaries to mark the beginning and end of the entities  
 Annotate the letter.

Lettre L11L <PERSON>La Fayette</PERSON> au <TITLE>citoyens</TITLE> <PERSON>Louis  
 Rameau</PERSON>.

<LOCATION>Vianen</LOCATION>, <DATE>14 ventôse an VII</DATE> (<DATE>4 mars 1799/  
 DATE).  
 Je me flatte encore de l'espérance de vous voir aujourd'hui, mon cher <PERSON>Louis/  
 PERSON>; si vous arrivez par la diligence, je vous remettrai moi-même ma lettre, sinon je  
 profiterai de la diligence pour qu'elle vous soit remise sincèrement à <LOCATION>Paris</  
 LOCATION>.

En voici une pour <PERSON>Masclet</PERSON> que je vous prie de cacher après l'avoir lu; je  
 croyez-la ou, si vous pouvez, remettre-la vous-même; l'ouvrage que je lui propose serait  
 utile, honorable, lucratif; il ne peut le compromettre ni avec les gouvernements ni avec les  
 gouvernés de <LOCATION>France</LOCATION> ou de <LOCATION>Angleterre</LOCATION>;  
 on lirait avec intérêt ces modèles d'éloquence, ces objets de comparaison, et ils  
 contribueraient à la connaissance des époques, des choses et des hommes. Il me semble que  
 personne n'est plus propre que <PERSON>Masclet</PERSON> à bien faire cette collection; je  
 désire vivement qu'il s'enthousiasme pour ce projet, auquel il admettrait sa femme, et à eux  
 deux ils feraient un choix excellent et des traductions parfaites.

A qui pensez-vous, mon cher ami, que nous devons remettre les matériaux des fragments  
 historiques? Je vois à <LOCATION>Paris</LOCATION> <PERSON>Rameau</PERSON>, <PERSON>Pastoret</PERSON>, <PERSON>Pastoret</PERSON>, <PERSON>Staud</PERSON>, les <PERSON>Lacretelle</  
 PERSON>; car les hommes actuellement employés n'en auraient pas le temps. Je vous prie d'y  
 réfléchir; vous pourriez même, dans l'occasion, voir quelles dispositions vous trouveriez à cet  
 égard dans les auteurs désignés.

Je n'ai pas reçu de réponse à ma lettre relative au projet de publications; Notice de  
 <PERSON>R</PERSON> <PERSON>Lettres de prison, Lettre à <TITLE>M</TITLE>  
 <PERSON>d'Yckninge</PERSON>, que vous auez, j'espère, retrouvée.

<PERSON>Ségur</PERSON> travaille à une vie du feu <TITLE>nois</TITLE> de  
 <LOCATION>Brusse</LOCATION>; il m'a demandé quelques notes; j'avais envoyé à ma femme  
 un résumé de notre captivité très bien fait, que vous copiez à <LOCATION>Lindenhof</  
 LOCATION> pour <TITLE>M</TITLE> <PERSON>me</PERSON> de <PERSON>Simiane</  
 PERSON>. C'est précisément ce qu'il lui conviendrait d'en dire; il me manque beaucoup de  
 papiers adressés de <LOCATION>Hollande</LOCATION> à <PERSON>Adrienne</PERSON> ou  
 portés par elle-même; elle est persuadée qu'ils sont restés chez <PERSON>Emmery</  
 PERSON>, qui ne m'a rendu que les écrits de <PERSON>Paup</PERSON>; <PERSON>Ségur</  
 PERSON> est pressé par le temps et par son libraire; j'ai peur d'avoir eu original ou la copie que  
 vous en avez faite.

<PERSON>Ségur</PERSON> met le plus grand prix à bien connaître les intrigues  
 diplomatiques de <DATE>92</DATE> <ORGANIZATION>le comité autrichien</  
 ORGANIZATION>, c'est à dire la négociation secrète entre les <PERSON>Lamelle</PERSON>, <TITLE>la reine</TITLE> et <ORGANIZATION>la cour de Vienne</ORGANIZATION>; y a joué  
 un rôle tellement influent que, pour quelque chose en croire trop ou peu sur cet objet, il est  
 impossible de se faire une idée nette de la politique de ce temps-là. <PERSON>Paup</  
 PERSON> en a fait un chapitre plein de vérité et de modération, qu'il prêta lui-même à  
 <PERSON>Masclet</PERSON> lorsque nous étions tous en <LOCATION>Hollande</  
 LOCATION>; je ne cherche point à récriminer, et <PERSON>Ségur</PERSON>, ami  
 d'<PERSON>Alexandre</PERSON> et de <PERSON>Dumas</PERSON>, ne fera pas un usage  
 hostile de ces connaissances, ou du moins ne dira que ce qui est nécessaire à l'histoire.  
 Cependant je ne voudrais pas que cet écrit fut donné par vous-même à <PERSON>Ségur</  
 PERSON>; il serait remis par moi si je vais bientôt à <LOCATION>Paris</LOCATION>, ou par  
 <PERSON>Georges</PERSON> avec un billet confidentiel; ce que je vous demande, mon cher  
 <PERSON>Louis</PERSON>, c'est de retrouver cet écrit chez <PERSON>Masclet</PERSON>.

Je voudrais bien avoir une copie de la lettre que j'écrivais pour mes compagnons au  
 <ORGANIZATION>Directoire</ORGANIZATION> ainsi que notre déclaration de  
 <LOCATION>Rochebeuf</LOCATION>; on m'a perdu tout cela, et ces deux pièces sont dans ce  
 moment essentielles.

Il serait peut-être bien fait de donner la seconde à <PERSON>Ségur</PERSON>, dont l'ouvrage  
 aura sincèrement beaucoup de lecteurs dans tous les pays.

Je ne vous écris ce matin que sur ces objets, et vous embrasse de tout mon cœur. <TITLE>M</  
 TITLE> <PERSON>me</PERSON> de <PERSON>Simiane</PERSON> ne m'a pas accusé la  
 réception d'un projet de note anglaise.



# Another Tagging Scheme

You are an NER can you tag this letter for Entity Types and Definitions

- \* SENDER includes the full name or identifier of the letter writer
- \* RECIPIENT includes the full name or identifier of the letter addressee
- \* SIGNATURE includes names, titles, or identifiers used in letter closings
- \* ADDRESS includes any postal addresses, street names, or mailing locations
- \* REFERENCE includes document numbers, case files, or citation identifiers
- \* MONETARY includes currency amounts, financial figures, or payment terms
- \* CORRESPONDENCE\_DATE includes dates related to letter writing, sending, or receiving
- \* DEADLINE includes due dates, response timeframes, or time-sensitive requirements

Annotation Format

Use <> brackets to mark entity boundaries at the beginning and end of each identified entity.

5 You are an NER, can you tag this letter for Entity Types and Definitions

- \* **SENDER** includes the full name or identifier of the letter writer
- \* **RECIPIENT** includes the full name or identifier of the letter addressee
- \* **SIGNATURE** includes names, titles, or identifiers used in letter closings
- \* **ADDRESS** includes any postal addresses, street names, or mailing locations
- \* **REFERENCE** includes document numbers, case files, or citation identifiers
- \* **MONETARY** includes currency amounts, financial figures, or payment terms
- \* **CORRESPONDENCE\_DATE** includes dates related to letter writing, sending, or receiving
- \* **DEADLINE** includes due dates, response timeframes, or time-sensitive requirements

Annotation Format

Use `<>` brackets to mark entity boundaries at the beginning and end of each identified entity.

Lettre <REFERENCE>LII</REFERENCE>, <SENDER>La Fayette</SENDER> au citoyen  
 <RECIPIENT>Louis Rameau</RECIPIENT>.

Vianen, <CORRESPONDENCE\_DATE>14 ventôse an VII (4 mars 1799)</  
 CORRESPONDENCE\_DATE>.

Je me flatte encore de l'espérance de vous voir aujourd'hui, mon cher Louis; si vous arrivez  
 par la diligence, je vous remettrai moi-même ma lettre, sinon je profiterai de la diligence pour  
 qu'elle vous soit remise sûrement à <ADDRESS>Paris</ADDRESS>.

En voici une pour Masclet que je vous prie de cacheter après l'avoir lue; envoyez-la ou, si vous  
 pouvez, remettez-la vous-même; l'ouvrage que je lui propose serait utile, honorable, lucratif;  
 il ne peut le compromettre ni avec les gouvernans ni avec les gouvernés de France ou  
 d'Angleterre; on lit avec intérêt ces modèles d'éloquence, ces objets de comparaison, et ils  
 contribueraient à la connaissance des époques, des choses et des hommes. Il me semble que  
 personne n'est plus propre que Masclet à bien faire cette collection; je désire vivement qu'il  
 s'enthousiasme pour ce projet, auquel il admettrait sa femme, et à eux deux ils feraient un  
 choix excellent et des traductions parfaites.

A qui pensez-vous, mon cher ami, que nous devons remettre les matériaux des fragments  
 historiques? Je vois à <ADDRESS>Paris</ADDRESS> Ramond, Pastoret, Staud, les Laçretelle;  
 car les hommes actuellement employés n'en aiment pas le temps. Je vous prie d'y réfléchir;  
 vous pourriez même, dans l'occasion, voir quelles dispositions vous trouveriez à cet égard  
 dans les auteurs désignés.

Je n'ai pas reçu de réponse à ma lettre relative au projet de publications: Notice de W. Lettres  
 de prison, Lettre à M. d'Yfenning, que vous avez, j'espère, retrouvée.

Ségar travaille à une vie du feu roi de Prusse; il m'a demandé quelques notes; j'avais envoyé à  
 ma femme un résumé de notre captivité très bien fait, que vous copiez à Lehmkuhlen pour  
 Mme de Simiane. C'est précisément ce qu'il lui conviendrait d'en dire; il me manque  
 beaucoup de papiers adressés de Hollande à Adrienne ou portés par elle-même; elle est  
 persuadée qu'ils sont restés chez Emmary, qui ne m'a rendu que les écrits de Puy; Ségar est  
 pressé par le temps et par son libraire; tâchez d'avoir cet original ou la copie que vous en avez  
 faite.

Ségar met le plus grand prix à bien connaître les intrigues diplomatiques de 92; le comité  
 autrichien, c'est-à-dire la négociation secrète entre les Lameth, la reine et la cour de Vienne, y  
 a joué un rôle tellement influent, que pour quiconque en croira trop ou peu sur cet objet, il est  
 impossible de se faire une idée nette de la politique de ce temps-là. Puy en a fait un chapitre  
 plein de vérité et de modération, qu'il prêta lui-même à Masclet lorsque nous étions tous en  
 l'histoire; je ne cherche point à dénigrer, et Ségar, ami d'Alexandre et de Dumas, ne fera pas  
 un usage hostile de ces connaissances, ou du moins ne dira que ce qui est nécessaire à  
 l'histoire. Cependant je ne voudrais pas que cet écrit fut donné par vous-même à Ségar; il  
 serait remis par moi si je vais bientôt à <ADDRESS>Paris</ADDRESS>, ou par Georges avec un  
 billet confidentiel; ce que je vous demande, mon cher Louis, c'est de retrouver cet écrit chez  
 Masclet.

Je voudrais bien avoir une copie de la lettre que j'écrivais pour mes compagnons au Directoire  
 ainsi que notre déclaration de Rochefort; on m'a perdu tout cela, et ces deux pièces sont dans  
 ce moment essentielles.

Il serait peut-être bien fait de donner la seconde à Ségar, dont l'ouvrage aura sûrement  
 beaucoup de lecteurs dans tous les pays.

Je ne vous écris ce matin que sur ces objets, et vous embrasse de tout mon cœur. Mme de  
 Simiane ne m'a pas accusé la réception d'un projet de note anglaise.



# Examples of Possible Analyses

- ❖ Translate the letter
- ❖ Summarize the letter
- ❖ Provide an interpretation of the letter
- ❖ Identify Entities in the letter or all the letters, persons, places, concepts etc.

# Risks of LLMs to analyze non-English Texts

- ❖ Oversimplifying or distorting key ideas.
- ❖ Misinterpreting rhetorical nuance and tone.
- ❖ Ignoring culturally dependent specific intellectual references.
- ❖ Would using just a French language trained LLM be sufficient or do we require an LLM trained specifically on French Enlightenment and revolutionary-era corpora

- ❖ *“On lirait avec intérêt ces modèles d'éloquence, ces objets de comparaison, et ils contribueraient à la connaissance des époques, des choses et des hommes.”*  
(“One would read with interest these models of eloquence, these objects of comparison, and they would contribute to the understanding of eras, of events, and of people.”)
- ❖ *Summarization often flattens meaning, losing the intellectual weight of phrases or rhetorical terms grounded in Enlightenment-era ideals.*
- ❖ *Emotional tone can be misjudged. When I presented the sentence to ChatGPT it responded that tone was “...passionate and promotional”, but really it is measured, diplomatic, and idealistic.*
- ❖ *The English LLM may fail to recognize conceptual entities even when it successfully identifies Entities such as persons, places, dates, objects.*

NLU bypass

NATURAL LANGUAGE UNDERSTANDING

NLU pipelines

Pre-processing >

Classifiers >

Vectorizers >

TF-IDF

FastText

FlauBERT

Entity extraction

Language support

Designing Intents

Designing Entities

SECURITY

Overview

Hosting

Activity logs

Automatic screen lock

# FlauBERT

Suggest Edits

## FlauBERT (French Language Understanding BERT)

FlauBERT is designed for specialized chatbots that require deep, nuanced language understanding and, while it demands extensive training data and computational resources, it provides unparalleled performance in complex domains.

### Use Case in Chatbots:

FlauBERT is best suited for sophisticated chatbot applications that demand a high level of language understanding. This is particularly useful for customer service chatbots, or those deployed in complex domains like healthcare or legal advice.

### Training Data:

FlauBERT, being a BERT-based model, usually requires extensive data for fine-tuning—around 100 or more samples per intent are recommended for optimal performance. The model was initially trained on a large corpus of French web data, giving it a robust understanding of the language.

### Training Time:

Training FlauBERT can be computationally intensive and time-consuming, especially if you're fine-tuning the model for specific tasks. However, the performance gains often justify the increased training time.

Updated over 1 year ago

← FastText

Entity extraction →

<https://docs.smartly.ai/docs/flaubert>

# FlauBERT French-language LLM

- ❖ From Facebook AI, not a web-based or online service.
- ❖ Access requires programming tools (e.g., PyTorch or TensorFlow).
- ❖ Typically integrated into other tools
- ❖ Download the model and interact with it via code in a compatible environment.

# González-Gallardo, et al 2024 is Repeatable

- ❖ The study uses open-source Instruct models such as: LLaMA-2 (70B), LLaMA-3 (70B), Mixtral (8×7B MoE), Zephyr (7B).
- ❖ Models accessible via HuggingFace's Inference API
- ❖ Public Datasets, AJMC, HIPE, and NewsEye freely available and well-documented.
- ❖ The code, prompt templates, and evaluation scripts on GitHub: [https://github.com/cic4k/LLMs\\_for\\_historical\\_NER](https://github.com/cic4k/LLMs_for_historical_NER)
- ❖ Four-stage process (model selection, prompt design, interaction mode, post-processing) is clearly explained
- ❖ The use of few-shot prompting reduces the need for training infrastructure—no fine-tuning required, just prompt engineering.



# Activity Three, Slave Narratives

# Final Discussion Future of History

# Questions

Also you can contact me at:  
[seamus.ross@utoronto.ca](mailto:seamus.ross@utoronto.ca)