



# Empowering Text Encoding with Large Language Models: Benefits and Challenges

Martina Scholger, Christopher Pollin, Sabrina Strutz  
Department of Digital Humanities, University of Graz

9 Oct 2024: TEI 2024. Textos, lenguas y comunidades

# AI Integration in Digital Scholarly Editing Workflows

**Current state:** Transitioning from Word/PDF to a fully digital edition.



**Workflow:** Automated process from Word -> TEI Garage -> TEI.  
This results in a “Gold Standard” for further evaluation.

**Continuous expansion:** New letters are regularly added and existing ones have to be refined.

**Core Question:** Can Large Language Models be employed to assist in specific tasks within the scholarly editing process?

# Prompting a Digital Scholarly Edition interface

- AI can already do a lot
- Example: Website of a DSE based on Hamlet, entirely prompt-based
- But: quality of the data is the most important factor for the quality of the results of the prompting process.

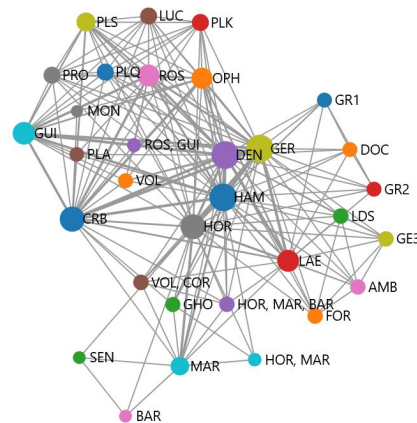
Philological Analysis of Shakespeare's Texts

Home Compare Editions Character Interactions Compare Interactions Word Frequency Concordance Lexical  
Dispersion Compare Word Frequencies

Search

Search

Character Interactions in ham-1603-22275x-bli-c01



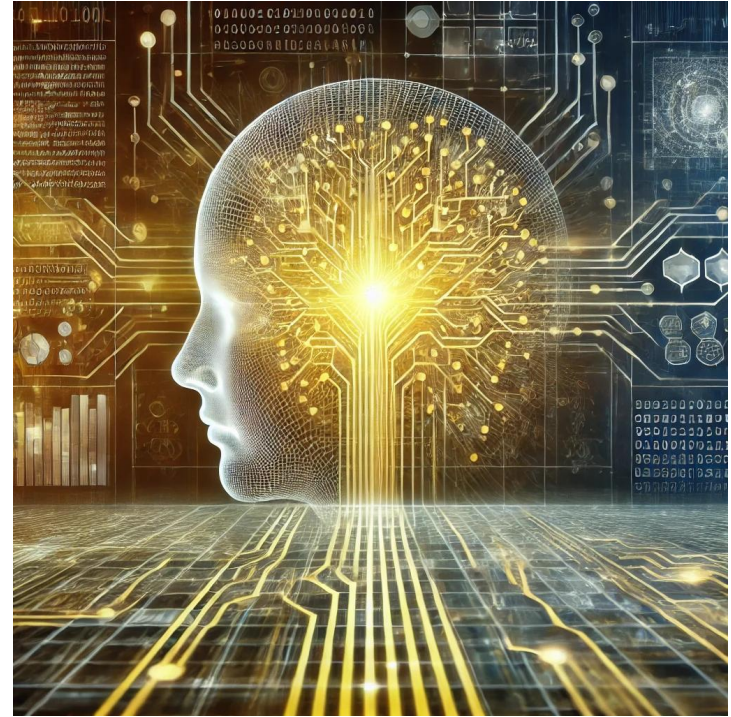
# LLMs & Generative AI

## Generative AI

- artificial intelligence models that create new content
- text, images, video, or audio
- learning patterns from existing data
- create unique outputs

## Large Language Models

- AI models designed to understand and generate natural language.
- Text generation, classification, translation, etc.



# Methods: Prompting & RAG

# Prompting



GPT4 / GPT-4o



# Prompt Engineering

**Prompt:** A natural language input provided to a model, which the model then responds to. This input can range from questions and statements to instructions or scenarios that guide the model's response.

**Prompt Engineering:** Prompt engineering is the process of formulating effective and precise prompts for AI systems to enhance the system's performance and achieve desired outputs and behaviors.

Ethan Mollick. Captain's log: the irreducible weirdness of prompting AIs.  
<https://www.oneusefulting.org/p/captains-log-the-irreducible-weirdness>







# Whisker Wisdom

## A Cat's Guide to Prompt Engineering

### 3.3 Design Principles

In this study, a number of guiding principles are used and instructions to elicit high-quality responses from the model or lead to irrelevant responses. Thus, the unnecessary information that does not contribute to guide the model. This is the basic principle of

**Conciseness and Clarity:** Generally, overly verbose prompts can lead to irrelevant responses. Thus, the unnecessary information that does not contribute to guide the model. This is the basic principle of

**Contextual Relevance:** The prompt must provide enough background and domain of the task, terminology, or situational descriptions can add context. We highlight this design philosophy

**Task Alignment:** The prompt should be clearly indicate the language and structure that clearly indicate the task. Involve phrasing the prompt as a question, that fits the task's expected input and output

**Example Demonstrations:** For more complex tasks, the prompt can demonstrate the desired format by showing input-output pairs, especially in the model due to its training data.

**Avoiding Bias:** Prompts should be designed to avoid ethical implications, especially for sensitive tasks.

**Incremental Prompting:** For tasks that require a series of prompts, the prompt should be structured to guide the model through the task. Also, prompts should be adjustable based on feedback, i.e., it needs to be well-tailored to the model's behavior. Moreover, the prompt should be well-tailored to the model's behavior.

Finally, more advanced prompts can be used for complex tasks. For instance, the prompt can include pseudo-code within the prompt

- Ensure specificity and clarity
- Allow time for system processing
- Provide context and examples
- Take an iterative approach
- Set explicit constraints
- Avoid excessive complexity
- Prompt in English



# Prompting methods

- **Zero-Shot Learning:** No examples provided; the model uses only the instruction.
- **In-Context Learning:** Enables models to understand and perform tasks based on contextual examples and instructions.
  - **One-Shot Learning:** One example (input-output pair) is given to guide the model.
  - **Few-Shot Learning:** Multiple examples (input-output pairs) are provided to establish a pattern for the model to follow.

# Zero-shot

You are an expert in encoding plain text letters using TEI (Text Encoding Initiative).

PERSONA

Your task is to encode the given letter into TEI-compliant XML.

TASK

Leave the TEI-Header out and encode the letter text only.

INSTRUCTION

# One-shot

You are an expert in encoding plain text letters using TEI (Text Encoding Initiative).

PERSONA

Your task is to encode the given letter into TEI-compliant XML.

TASK

Leave the TEI-Header out and encode the letter text only.

INSTRUCTION

*\*Plain Text Letter:\**

Bei [Fn: PvTh D.2.1.14.] der bestehenden Ordnung kann in diese des Supplikanten Bitte zwar nicht eingewilligt werden; in dessen bleibt immer richtig, daß auf Verdienst und Wohlverhalten Rücksicht genommen und nach Beschaffenheit der Umstände auch jedweder auf eine sonst vergnügliche Art solle Gerechtigkeit geleistet werden; welches Supplikant nach seiner Einsicht selbst wohl zu verlangen nicht außer acht setzen wird.

v[on] Jenisch

Von der Kanzley den 20. Hornung [1]794

---

*\*TEI encoded letter:\**

```
<div type="entireLetter"><div type="letter" subtype="original"><p>Bei<note place="bottom"><p>PvTh  
D.2.1.14.</p></note> der bestehenden Ordnung kann in diese des Supplikanten Bitte zwar nicht eingewilligt werden; in  
dessen bleibt immer richtig, daß auf Verdienst und Wohlverhalten Rücksicht genommen und nach Beschaffenheit der  
Umstände auch jedweder auf eine sonst vergnügliche Art solle Gerechtigkeit geleistet werden; welches Supplikant nach  
seiner Einsicht selbst wohl zu verlangen nicht außer acht setzen wird.</p><closer><dateline>Von der Kanzley den 20.  
Hornung [1]794</dateline><signed>v[on] Jenisch</signed></closer></div></div>
```

EXAMPLE

# Few-shot

You are an expert in encoding plain text letters using TEI (Text Encoding Initiative).

Your task is to encode the given letter into TEI-compliant XML.

Leave the TEI-Header out and encode the letter text only.

\*Plain text letter 1:\*

...  
---

\*TEI encoded letter 1r:\*

```
<div type="entireLetter"><div type="letter" subtype="original">...</div></div>
```

---

\*Plain text letter 2:\*

...  
---

\*TEI encoded letter 2:\*

```
<div type="entireLetter"><div type="letter" subtype="original">...</div></div>
```

PERSONA

TASK

INSTRUCTION

EXAMPLE

EXAMPLE

You will act as a skilled expert automaton that is proficient in transforming unstructured text, specifically letters from or to the Austrian orientalist Joseph von Hammer-Purgstall (1774-1856), into well-formed TEI XML. Analyze the provided text based on the mapping rules I have shared and then execute the transformation to produce TEI XML, ensuring you adhere to the guidelines and only annotate if certain.

Mapping rules:

- \* <div> Entire letter
- \* <dateline> Date/time reference of the letter
- \* <date> in <dateline>
- \* <opener> Opening of the letter
- \* <closer> Closing of the letter
- [...]

Guidelines:

- \* Strictly follow mapping rules
- \* Preserve the original text in its entirety
- \* Produce well-formed TEI XML according to the TEI standard
- \* Return the <div> only

Letter to encode:

```
``` {plain text} ```
```

Encoding examples:

```
``` xml example 1, example 2```
```

This is very important for my career!

Persona
Task
Context
Specifications
Instructions
Examples
Emotional Prompting

Prompting



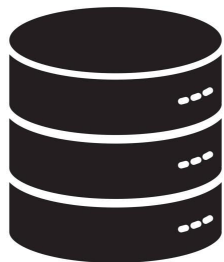
GPT4 / GPT-4o

Prompt

LLM

Response

RAG



Vector DB



Retrieved  
documents



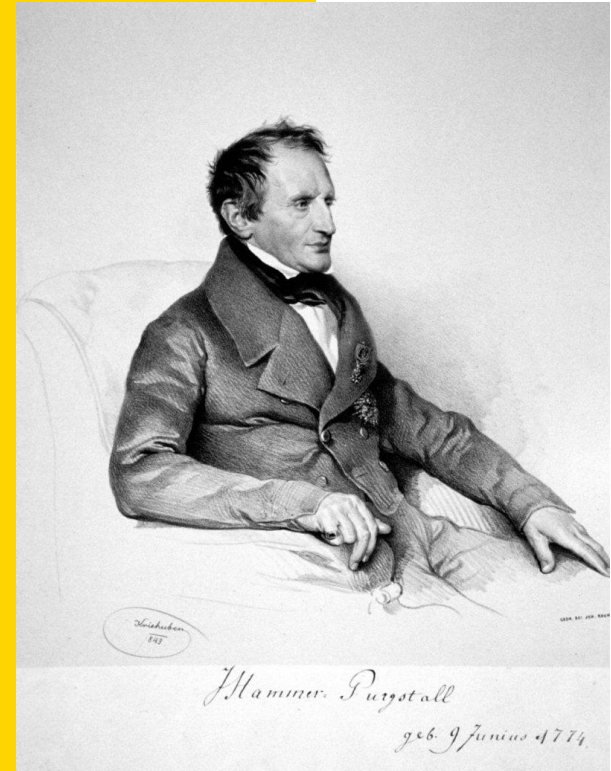
# RAG (Retrieval Augmented Generation)

- combines retrieval-based methods with generative language models
- integrates relevant information from a knowledge base
- increases the accuracy and relevance of text output

What is RAG? (Retrieval Augmented Generation) by Don Woodlock

<https://www.youtube.com/watch?v=u47GtXwePms>

# Digital letter edition of Joseph von Hammer-Purgstall



1774-1856

# Material

Joseph von Hammer-Purgstall (1774-1856)  
Austrian Orientalist, Historian, and Diplomat  
Pioneer in Middle Eastern Studies  
Vast network of academic contacts  
> Letters, memory notes, ...

- 3.461 TEI encoded letters
- ~4.000 additional letters exist and are waiting for processing
- Available in print and online (as PDF);  
<http://gams.uni-graz.at/hp>

5000  
pages

1794

1. 2000 HPP/Jenisch 1794 II 20 [Wien]

Euer Gnaden  
Unterzeichneter bittet um die gnädige Erlaubnis, künftigen Sonntag die Redoute<sup>1</sup> besuchen zu dürfen; weil er  
1.) im letzten und sechsten Jahre seiner Studien und bereits ein Jahr länger an der Akademie ist als es seine Vorfahren waren, da ihnen diese Erlaubnis erteilt wurde.  
2.) auf das ihm jederzeit von S[ei]ner H[och]vohrliebe dem Herrn Direktor erteilte Zeugnis seines guten Verhaltens und sittlichen Betragens stolz ist, und sich daher schmeichelt, dass die Besorgnis einiger daraus entstehenden üblen Folgen bei ihm nicht statt haben dürfte.  
3.) die Vorschrift einer für diesen Fall geforderten anständigen und sicheren Begleitung nicht besser erfüllt haben könnte, indem er in der Gesellschaft der Frau Schwester des Herrn Direktors sich befindet wird.

Josef HAMMER

2. 366 Jenisch/HPP 1794 II 20 [Wien]

Bei der bestehenden Ordnung kann in diese des Supplikanten Bitte zwar nicht eingewilligt werden; in dessen bleibet immer richtig, daß auf Verdienst und Wohlverhalten Rücksicht genommen und nach Beschaffenheit der Umstände auch jedweder auf eine sonst vernünftige Art solle Gerechtigkeit geleistet werden; welches Supplikant nach seiner Einsicht selbst wohl

1. Eine Tauschveranlassung oder auch ein dergleichen es sich hier handelt, ist nicht festzulegen.  
2. Die Erlaubnis durch H[och]vohrliebe, das nachfolgende vorangehende Bitte H[och]vohrliebe mit dem Vermerk

PUBLIKATIONEN AUS DEM ARCHIV DER UNIVERSITÄT GRAZ

BAND 41\*

JOSEPH VON HAMMER-PURGSTALL  
ERINNERUNGEN UND BRIEFE

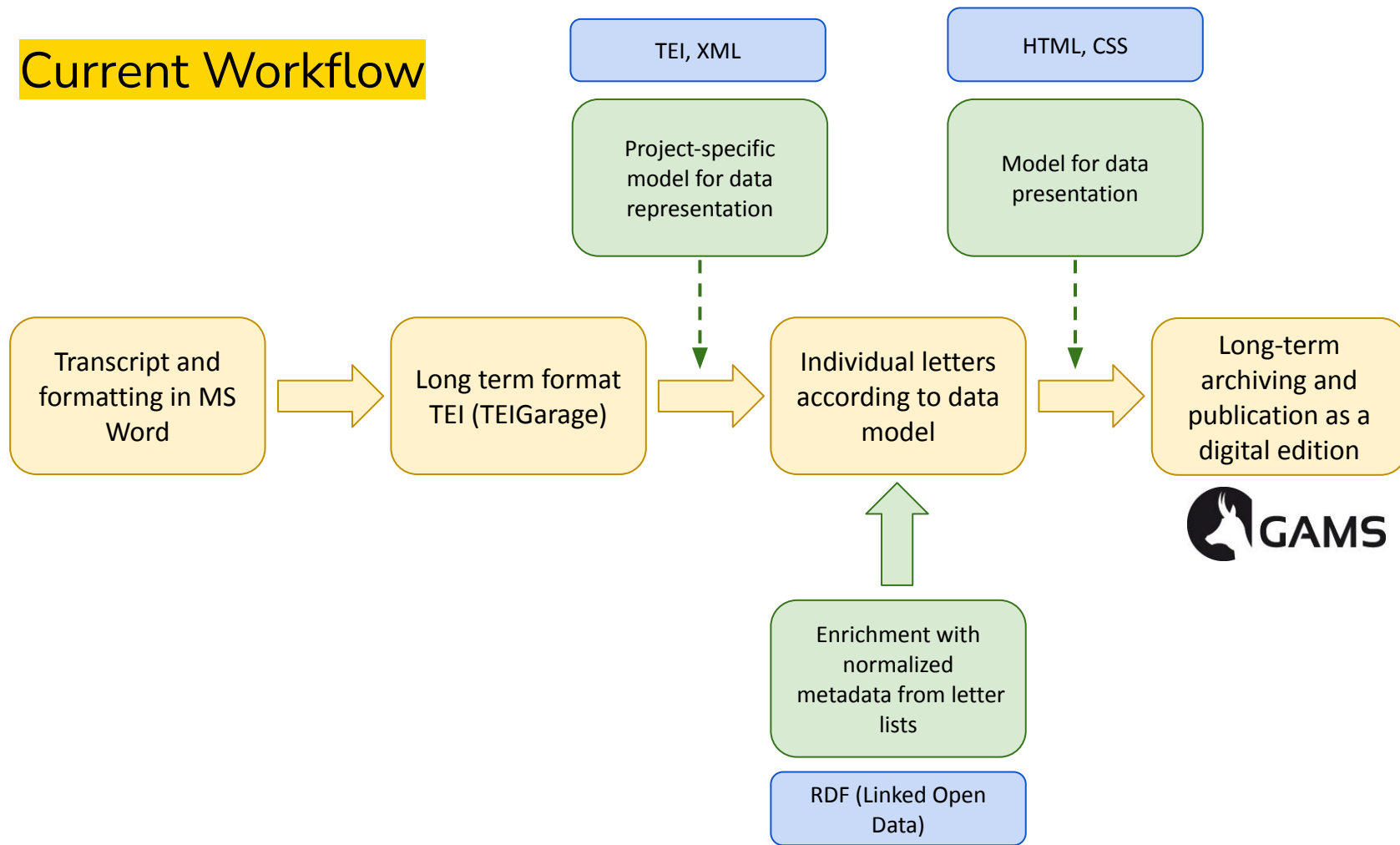
Version 1 201107: Briefe von 1790 bis Ende 1819

Herausgegeben von  
Walter Höflechner und Alexandra Wagner  
unter Heranziehung der Arbeiten von  
Herbert König, Gerit Koltz-Aiko, Alexandra Marics,  
Gustav Mittelbach, T. Thomas Walling u. a.

Band 1



# Current Workflow



# Machine-readable transcript

Formatvorlagen

HPE:Begrüßung

HPE:Briefabschluss

HPE:Briefüberschrift

HPE:Briefunterschrift

HPE:Datumszeile

HPE:Jahresüberschrift

HPE:Kommentar

HPE:Nicht bearbeitet

HPE:Postskriptum

HTML Adresse

HTML Akronym

HTML Beispiel

HTML Code

HTML Schreibmaschine

HTML Variable

HTML Vorformatiert

iast

Vorschau anzeigen

Verknüpfte Formatvorlagen deaktivieren

A<sub>+</sub>

A<sub>Q</sub>

A<sub>✓</sub>

Optionen...

1794

1. 2000 HP/Jenisch 1794 II 20/[Wien]

Euer Gnaden,

Unterzeichneter<sup>1</sup> bittet um die gnädige Erlaubnis, künftigen Sonntag die Redoute<sup>2</sup> besuchen zu dürfen<sup>3</sup>, weil er

1.) im letzten und sechsten Jahre seiner Studien und bereits ein Jahr länger an der Akademie ist als es seine Vorfahren waren, da ihnen diese Erlaubnis erteilt wurde.

2.) auf das ihm jederzeit von S[einer] H[ochwohlgeboren] dem Herrn Direktor erteilte Zeugnis seines guten Verhaltens und sittlichen Betragen stolz ist, und sich daher schmeichelt, dass die Besorgnis einiger daraus entstehenden üblen Folgen bei ihm nicht statt haben dürfte.

3.) die Vorschrift einer für diesen Fall geforderten anständigen und sicheren Begleitung nicht besser erfüllt haben könnte, indem er in der Gesellschaft der Frau Schwester des Herrn Direktors sich befinden wird.

Josef HAMMER

Den 20ten Hornung 1794

2. 366 Jenisch/HP 1794 II 20/Wien

Bei<sup>4</sup> der bestehenden Ordnung kann in diese des Supplikanten Bitte zwar nicht

Eine Tanzveranstaltung oder i  
eleganter Maskenball. Um we  
Veranstaltung es sich hier har  
feststellbar.

19

# TEI letter model

```
<div type="letter" subtype="original">
  <opener><salute>Euer Gnaden</salute>,</opener>
  <p>Unterzeichneter<note type="comment"><p> PvTh D.2.1.14.</p></note> bittet um
    die gnädige Erlaubnis, künftigen Sonntag die Redoute<note place="bottom"><p>
      [...] </note> besuchen zu dürfen<note type="comment">[...]</p></note>, weil er</p>
  <list>
    <item n="1.)">im letzten und sechsten Jahre seiner Studien [...] und bereits ein
      Jahr länger an der Akademie ist als es seine Vorfahren waren, da ihnen
      diese Erlaubnis erteilt wurde.</item>
    [...]
  </list>
  <closer>
    <signed>Joseph <persName>Hammer</persName></signed>
    <dateline>Den 20ten Hornung 1794</dateline>
  </closer>
</div>
```



# Presentation as digital scholarly edition

Euer Gnaden,

Unterzeichneter<sup>1</sup> bittet um die gnädige Erlaubnis, künftigen Sonntag die Redoute<sup>2</sup> besuchen zu dürfen<sup>3</sup>, weil er

1.) im letzten und sechsten Jahre seiner Studien und bereits ein Jahr länger an der Akademie ist als es seine Vorfahren waren, da ihnen diese Erlaubnis erteilt wurde.

2.) auf das ihm jederzeit von S[einer] H[ochwohlgeboren] dem Herrn Direktor erteilte Zeugnis seines guten Verhaltens und sittlichen Betragen stolz ist, und sich daher schmeichelt, dass die Besorgnis einiger daraus entstehenden üblen Folgen bei ihm nicht statt haben dürfte.

3.) die Vorschrift einer für diesen Fall geforderten anständigen und sicheren Begleitung nicht besser erfüllt haben könnte, indem er in der Gesellschaft der Frau Schwester des Herrn Direktors sich befinden wird.

Joseph HAMMER

Den 20ten Hornung 1794

<sup>1</sup> PvTh D.2.1.14.

<sup>2</sup> Eine Tanzveranstaltung oder auch ein eleganter Maskenball. Um welche derartige Veranstaltung es sich hier handelt, ist nicht feststellbar.

<sup>3</sup> Die Erledigung durch JENISCH s. das nachfolgende Stück, findet sich auf demselben Blatt wie die vorangegangene Bitte HPs mit dem Vermerk „Von der Kanzlei den 20. Hornung 1794“.

## Joseph von Hammer-Purgstall an Bernhard Freiherr von Jenisch. [Wien], 20.2.1794

 TEI-Source

AbsenderIn: Joseph von Hammer-Purgstall

EmpfängerIn: Bernhard Freiherr von Jenisch

Absendeort: [Wien]

Datum: 20.2.1794

Sprache: Deutsch

Quelle: PvTh D.2.1.14 - Nachlass Rudolf Payer von Thurn

Zitiervorschlag: Joseph von Hammer-Purgstall an Bernhard Freiherr von Jenisch. [Wien], 20.2.1794. In: Walter Höflechner unter Mitarbeit von Alexandra Wagner, Gerit Koitz-Arko und Sylvia Kowatsch (Hrsg.): Joseph von Hammer-Purgstall. Briefe. Online unter <http://glossa.uni-graz.at/o:hpe.00001>

What tasks can generative AI take on?

# Generative AI in the Editing workflow

(incomplete suggestions)

- Transcription
- OCR Cleanup
- TEI encoding
- Normalization
- Named Entity Recognition
- Translation & text summary

See workshop material from DHd 2024: <https://doi.org/10.5281/zenodo.10893761>

## Text recognition

2.

Wien, den 4. Oktober. 1798.

Ich habe das Glück gehabt, den großen Orientalisten, den Frater Paulinus aus Rom, hier persönlich kennen zu lernen, und mich eben so herzlich zu freuen, daß er ein geborner Oesterreicher ist, als er selbst über das Wiedersehn seines lieben Vaterlandes erfreut ist. Er geht nun als Bibliothekär nach Padua, wo er zugleich das Sekretariat der emigrirten Propaganda übernehmen wird. Diese Stadt, von jeher ein ehrwürdiger Sitz der  
Ges

# Text recognition

You are an AI expert automaton with specialized skills in analyzing printed and handwritten texts from historical documents. Your current task involves processing and interpreting letters written by the Austrian diplomat and orientalist Joseph von Hammer-Purgstall (1774-1856).

The input for this task are images of Hammer-Purgstall's letters. Your approach involves the following detailed steps:

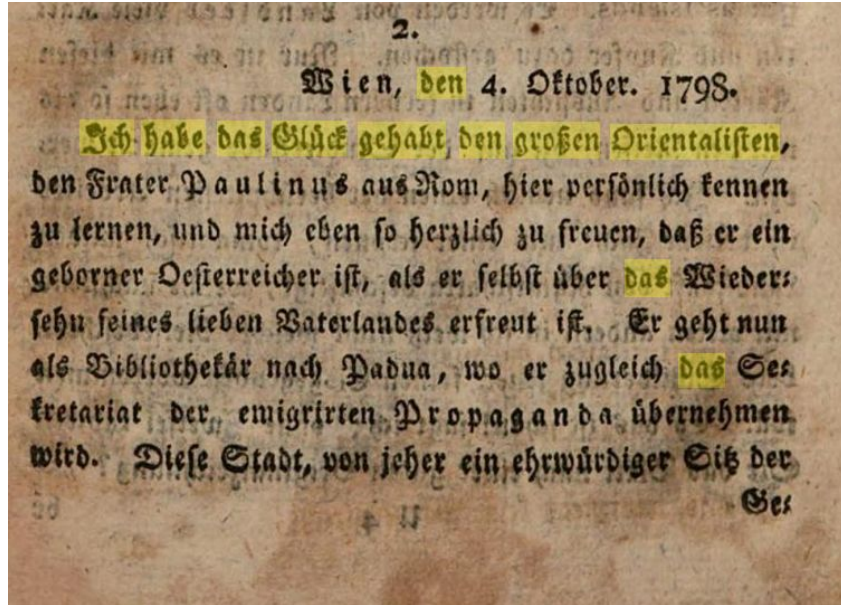
1. **\*\*Image Analysis\*\***: Thoroughly examine the image of each letter, focusing on details such as ink color, pen type, and paper quality.
2. **\*\*Text Reading\*\***: Carefully read the entire text of the letter.
3. **\*\*Context Analysis\*\***: Conduct a comprehensive context analysis of the text, considering the historical and cultural background of Hammer-Purgstall's era.
4. **\*\*Rereading for Accuracy\*\***: Review the text once more to ensure complete accuracy in understanding and interpretation.

Further Rules:

- \* Return in Code text with format Plain Text.
- \* If you can not read something write [\*] for every character.
- \* Make "\n" for every new line.
- \* First read this system prompt very carefully and ask me for uploading the image.
- \* Do not use Code Interpreter or OCR. Use the Vision of GPT-Vision.

Your goal is to extract and interpret the text with high precision, considering all the above aspects of Hammer-Purgstall's handwriting style. Let's work on this step by step. This is very important for my career!

# Text recognition



Wien, den 4. Oktober. 1798.

Ich habe das Glück gehabt, den großen Orientalisten, **ben** Frater Paulinus aus Rom, hier persönlich kennen zu lernen, und mich eben so herzlich zu freuen, daß er ein geborner **Desterreicher** ist, als er selbst über das Wiedersehn seines lieben **Baterlandes** erfreut ist. Er geht nun als Bibliothekar nach Padua, wo er zugleich das Sekretariat der emigrirten Propaganda übernehmen wird. Diese Stadt, von jeher ein ehrwürdiger **Sitz** der Ges



# OCR corrections

**Wien, den 4. Oktober. 1798.**

Ich habe das Glück gehabt, den großen Orientalisten, den Frater Paulinus aus Rom, hier persönlich kennen zu lernen, und mich ebenso herzlich zu freuen, daß er ein geborner Österreicher ist, als er selbst über das Wiedersehen seines lieben Vaterlandes erfreut ist. Er geht nun als Bibliothekar nach Padua, wo er zugleich das Sekretariat der emigrierten Propaganda übernehmen wird. Diese Stadt, von jeher ein ehrwürdiger Sitz der...

- risk of over-correction (emigirten -> emigrierten; eben so -> ebenso)

You will act as a skilled expert automaton that is proficient in transforming unstructured text, specifically letters from or to the Austrian orientalist Joseph von Hammer-Purgstall (1774-1856), into well-formed TEI XML. Analyze the provided text based on the mapping rules I have shared and then execute the transformation to produce TEI XML, ensuring you adhere to the guidelines and only annotate if certain.

#### Mapping rules:

- \* <div> Entire letter
- \* <dateline> Date/time reference of the letter
- \* <date> in <dateline>
- \* <opener> Opening of the letter
- \* <closer> Closing of the letter
- [...]

#### Guidelines:

- \* Strictly follow mapping rules
- \* Preserve the original text in its entirety
- \* Produce well-formed TEI XML according to the TEI standard
- \* Return the <div> only

#### Letter to encode:

```
``` {plain text} ```
```

#### Encoding examples:

```
``` xml example 1, example 2```
```

This is very important for my career!

Persona

Task

Specifications

Instructions

Examples

Emotional  
Prompting

# Prompting the TEI letter structure

You will act as a skilled expert automaton that is proficient in transforming unstructured text, specifically letters from or to the Austrian orientalist Joseph von Hammer-Purgstall (1774-1856), into well-formed TEI XML. Analyze the provided text and execute the transformation to produce TEI XML, ensuring you adhere to the guidelines and only annotate if certain.

## Guidelines:

- \* Strictly follow the encoding rules of the TEI for letter encoding
- \* Embed the footnotes provided at the end of the document directly to the text they belong to
- \* Normalize the date using the @when attribute on <date>
- \* Preserve the original text in its entirety
- \* ...

Letter from Hammer-Purgstall to Jenisch:

''' {text} '''

This is very important for my career!

PERSONA

TASK

CONTEXT

INSTRUCTION

INPUT

EMOTIONAL  
PROMPTING

# Plain text input

Euer Gnaden,

Unterzeichneter [Fn: PvTh D.2.1.14.] bittet um die gnädige Erlaubnis, künftigen Sonntag die Redoute [Fn: Eine Tanzveranstaltung oder auch ein eleganter Maskenball. Um welche derartige Veranstaltung es sich hier handelt, ist nicht feststellbar.] besuchen zu dürfen [Fn: Die Erledigung durch Jenisch, s. das nachfolgende Stück, findet sich auf demselben Blatt wie die vorangegangene Bitte HPs mit dem Vermerk „Von der Kanzlei den 20. Hornung 1794“.], weil er

- 1.) im letzten und sechsten Jahre seiner Studien und bereits ein Jahr länger an der Akademie ist als es seine Vorfahren waren, da ihnen diese Erlaubnis erteilt wurde.
- 2.) auf das ihm jederzeit von S[einer] H[ochwohlgeboren] dem Herrn Direktor erteilte Zeugnis seines guten Verhaltens und sittlichen Betragen stolz ist, und sich daher schmeichelt, dass die Besorgnis einiger daraus entstehenden üblen Folgen bei ihm nicht statt haben dürfte.
- 3.) die Vorschrift einer für diesen Fall geforderten anständigen und sicheren Begleitung nicht besser erfüllt haben könnte, indem er in der Gesellschaft der Frau Schwester des Herrn Direktors sich befinden wird.

Joseph HAMMER

Den 20ten Hornung 1794

## TEI letter structure - XML output

```
<div>
  <opener>
    <salute>Euer Gnaden,</salute>
  </opener>
  <p>Unterzeichneter<note place="inline">PvTh D.2.1.14.</note> bittet um die
gnädige Erlaubnis, künftigen Sonntag die <mentioned
rend="term">Redoute</mentioned><note place="inline">...</note> besuchen zu
dürfen<note place="inline">...</note>, weil er</p>
  <list>
    <item>im letzten...</item>
    <item>auf das ihm jederzeit von <hi rend="supplied">S[einer]</hi>
      <hi rend="supplied">H[ochwohlgeboren]</hi>...</item>
    <item>die Vorschrift einer...</item>
  </list>
  <signed>Joseph HAMMER</signed>
  <dateline>
    <date when="1794-02-20">Den 20ten Hornung 1794</date>
  </dateline>
</div>
```

# TEI letter structure - Addition of mapping rules

- \* enclose the letter text in a <div> element with the @type-attribute 'letter';
- \* find opener <opener>, closer <closer>, signatures <signed>, salutations <salute>, paragraphs <p> and other textual structure;
- \* place the notes as embedded elements <note type="comment">, not at the end of the document;
- \* encode person names as <persName>;
- \* put the reference to norm data (e.g. GND or VIAF) into a @ref-attribute only if you are sure that it is the correct reference
- \* normalize the date using the @when attribute in <date>
- \* resolve abbreviations that are marked with square brackets and use the <choice> mechanism with <abbr> and <expan>. Encode every word separately in an <choice> element, put the character(s) outside the square brackets into <abbr> and the character(s) within the square brackets into <expan>. Remove the square brackets.
- \* make sure that punctuation marks or articles are not part of salutations, dates, or person names.
- \* address only the steps that are in the instructions. do not add more.

## SPECIFICATIONS



```

<body>
  <div type="letter">
    <opener>
      <salute>Euer Gnaden</salute>,
    </opener>
    <p>Unterzeichneter<note type="comment" n="1">PvTh D.2.1.14.</note> bittet um die
gnädige Erlaubnis, künftigen Sonntag die...</p>
    <list>
      <item>
        <p>im letzten und sechsten Jahre seiner Studien und bereits ein Jahr länger an
der Akademie ist als es seine Vorfahren waren, da ihnen diese Erlaubnis erteilt
wurde.</p>
      </item>
      <item>
        <p>auf das ihm jederzeit von <choice><abbr>S</abbr><expn>einer</expn></choice>
<choice><abbr>H</abbr><expn>ochwohlgeboren</expn></choice> ...</p>
      </item>
      ...
    </list>
    <closer>
      <date when="1794-02-20">Den 20ten Hornung 1794</date>
      <signed>
        <persName>Joseph Hammer</persName>
      </signed>
    </closer>
  </div>
</body>

```

# Tagging of different phenomena by []

Orthography, additions by editors

H[ochwohlgeboren]

```
<choice>  
  <abbr>H</abbr>  
  <expansion>ochwohlgeboren</expansion>  
</choice>
```

emporschwin[gen]des

```
<supplied reason="reconstructed">gen</supplied>
```

welches[sic]

```
<choice>  
  <sic>welches</sic>  
  <correction>welcher</correction>  
</choice>
```

# Named Entity Recognition

- RAG approach: Inclusion of additional lists as supplementary knowledge bases

```
persons_list = [  
    {"name": "Hammer-Purgstall, Joseph von", "reference":  
"https://www.wikidata.org/wiki/Q89546", "description": "Scholar and orientalist"},  
    {"name": "C. A. Böttiger", "reference": "https://www.wikidata.org/wiki/Q213870",  
"description": "German archaeologist and historian"}, ...  
]  
  
placenames_list = [  
    {"name": "Zürich", "reference": "https://www.geonames.org/2657896"},  
    {"name": "Weimar", "reference": "https://www.geonames.org/2812482"}, ...  
]
```

# Named Entity Recognition

```
augmented_prompt = f"""
```

```
You are an expert in TEI encoding for historical correspondences.
```

```
Encode the following letter into TEI-compliant XML, ensuring that it reflects the appropriate structure for a letter, including elements like <opener>, <closer>, <signed>, and <dateline>.
```

```
Pay special attention to the encoding of names of persons and places using the <persName> and <placeName> elements with appropriate references.
```

```
Here are the related persons that may be referenced in the encoding:
```

```
{related_info}
```

```
Here are the related placenames that may be referenced in the encoding:
```

```
{related_placenames_info}
```

- Full input text should be included as the body of the letter.
- Maintain the order of the text exactly as it appears.

```
Include a <standOff> section with all the available information about the persons and placenames listed above, structured as follows:
```

```
{standoff_section}
```

```
The expected output structure should adhere to TEI standards:
```

```
{output_example}
```

```
Now, encode the letter:
```

```
{input_text}
```

```
"""
```

# Named Entity Recognition (Input)

Weimar d[en] 1ten July 1796

Euer Hochwohlgeboren

Erhalten hier zwei Aushängebogen vom neuesten Stück des Merkurs , wozu Sie durch Herrn Hofrat Müller einen in jeder Rücksicht merkwürdigen und auserlesenen Aufsatz eingesendet haben. Da, wie wir hier aus den Zeitungen wissen, jener würdige Mann in wichtigen, Gott gebe für ganz Deutschland! wohltätigen Aufträgen nach London gegangen ist, so habe ich dem Verlangen des Herrn H[of]R[ates] gemäß Ihnen selbst das zuschicken wollen, was unser Vater Wieland gewiß mit einem Brief begleiten würde, wenn er hier wäre. Er ist aber seit einem Monat in Zürich bei seinem Schwiegersohn, dem Buchhändler Gesner, und kommt erst in 6 Wochen von da zurück.

Wie sehr freue ich mich, daß mir, der ich unterdessen die Besorgung des Merkur übernommen habe, dadurch das Vergnügen zu teil wird, Ihnen in aller Namen, die hier Ihre treffliche Probe schon gelesen haben, für diese neuen Aussichten auf den reichsten Blumengarten des Orients danken und besonders in Herders Namen die Bitte dringend wiederholen zu können, daß Sie den Merkur doch mit mehreren Beiträgen aus Ihrem Frucht- und Füllhorn zieren mögen. Sie sollen sogleich die Ehrenstelle darinnen erhalten, die sie verdienen.

So viel nur vorläufig. Kann Ihnen an einer weiteren Korrespondenz etwas gelegen sein, so beehren Sie in Abwesenheit meines Freundes Wieland damit Ihren gehorsamsten Diener

C. A. Böttiger

Oberconsistorialrat

# Named Entity Recognition (Output)

```
<div type="letter">
  <opener>
    <dateline><placeName ref="#place002">Weimar</placeName> d[en] 1ten July
      1796</dateline>
    <salute>Euer Hochwohlgeboren</salute>
  </opener>
  <p>Erhalten hier zwei Aushängebogen vom neuesten Stück des Merkurs , wozu Sie durch
    Herrn Hofrat <persName ref="#person006">Müller</persName> einen in jeder
    Rücksicht merkwürdigen und auserlesenen Aufsatz eingesendet haben. Da, wie wir
    hier aus den Zeitungen wissen, jener würdige Mann in wichtigen, Gott gebe für
    ganz Deutschland! wohltätigen Aufträgen nach London gegangen ist, so habe ich
    dem Verlangen des Herrn H[of]R[ates] gemäß Ihnen selbst das zuschicken wollen,
    was unser Vater <persName ref="#person008">Wieland</persName> gewiß mit
    einem Brief begleiten würde, wenn er hier wäre. Er ist aber seit einem Monat in
    <placeName ref="#place001">Zürich</placeName> bei seinem Schwiegersohn,
    dem Buchhändler <persName ref="#person009">Gesner</persName> , und kommt
    erst in 6 Wochen von da zurück.</p>
    ...
  <p>So viel nur vorläufig. Kann Ihnen an einer weiteren Korrespondenz etwas gelegen
    sein, so beehren Sie in Abwesenheit meines Freundes <persName
      ref="#person008">Wieland</persName> damit Ihren gehorsamsten Diener</p>
  <closer>
    <signed><persName ref="#person005">C. A. Böttiger</persName></signed>
    <salute>Oberconsistorialrat</salute>
  </closer>
</div>
```

```
<standOff>
  <listPerson>
    <person xml:id="person001">
      <persName>Hammer-Purgstall, Joseph von</persName>
      <idno
type="wikidata">https://www.wikidata.org/wiki/Q89546</idno>
    </person>
    <person xml:id="person002">
      <persName>C. A. Böttiger</persName>
      <idno
type="wikidata">https://www.wikidata.org/wiki/Q213870</idno>
    </person>
    ...
  </listPerson>
  <listPlace>
    <place xml:id="place001">
      <placeName>Zürich</placeName>
      <idno type="wikidata">https://www.geonames.org/2657896</idno>
    </place>
    <place xml:id="place002">
      <placeName>Weimar</placeName>
      <idno
type="wikidata">https://www.geonames.org/2812482</idno>
    </place>
  </listPlace>
</standOff>
```

# Translation [FR -> DE]

## ORIGINAL

De retour d'une petite course à Hambourg, où Mad Reimarus m'a beaucoup parlé de vous d'après les lettres de sa fille, je trouve l'heureux avis de Breitkopf, de pouvoir enfin vous faire parvenir un paquet, que j'avais fermé, il y a je crois deux mois.

## Traducción de los/as editores/as

Zurück von einer kleinen Reise nach Hamburg, wo Frau Reimarus mir viel von Ihnen nach den Briefen ihrer Tochter erzählt hat, finde ich die positive Mitteilung von Breitkopf, dass er Ihnen endlich ein Paket zukommen lassen kann, das ich, glaube ich, vor zwei Monaten zugemacht habe.

ChatGPT 4o

As an expert in translating historical texts, provide a translation of the 1807 letter from French to German. Maintain the original style and historical language, avoiding any modernization, and stay as faithful as possible to the source.

De retour d'une petite course à Ha  
< [ ] >



Here is the translation following your optimized prompt:

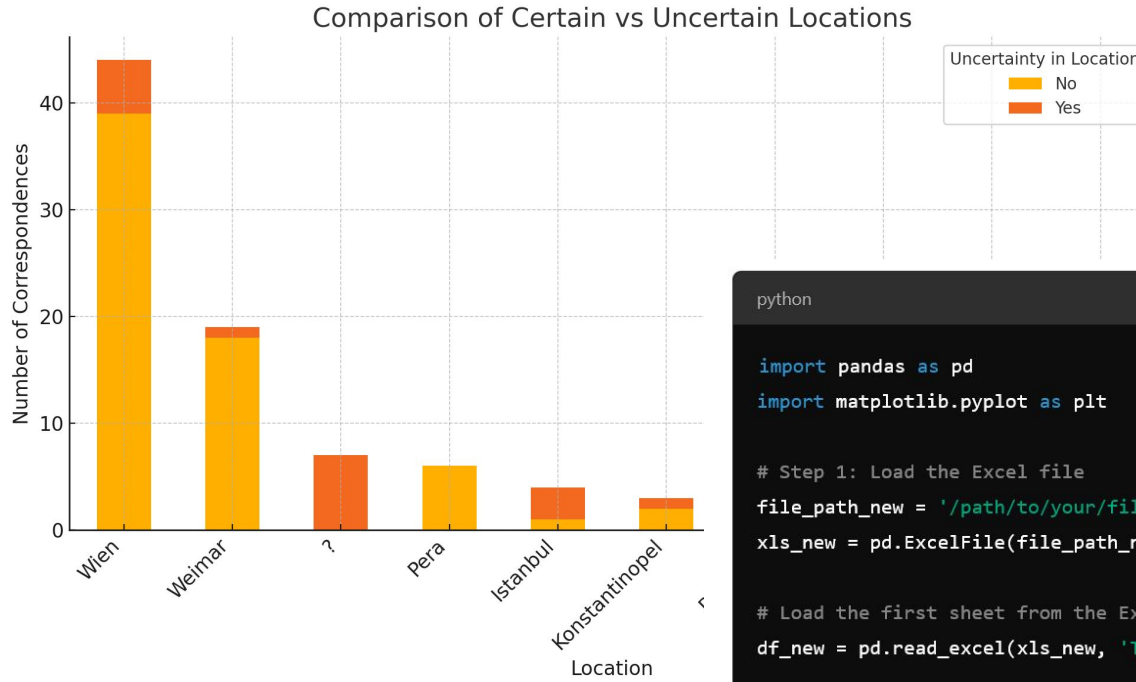
Zurückgekehrt von einem kleinen Ausflug nach Hamburg, wo Madame Reimarus mir viel von Ihnen erzählte, nach den Briefen ihrer Tochter, finde ich die erfreuliche Nachricht von Breitkopf vor, Ihnen endlich ein Paket zukommen lassen zu können, das ich, glaube ich, vor zwei Monaten verschlossen hatte.

# Visualisation of metadata: ADA (Advanced Data Analysis)

Datei Start Einfügen Seitenlayout Formeln Daten Überprüfen Ansicht Hilfe Was möchten Sie tun?											
Einfügen		Schriftart		Ausrichtung		Zahl		Formatvorlagen		Zellen	
M5											
	A	B	C	D	E	F	G	H	I	J	K
1	Identifikationsnr	Schreiber	Adressat		Jahr		Monat		Tag		Ort
2	2000	Hammer-Purgstall, Joseph von	Jenisch, Bernhard Freiherr von	1794	1794	2	2	20	20		[Wien]
3	366	Jenisch, Bernhard Freiherr von	Hammer-Purgstall, Joseph von	1794	1794	2	2	20	20		Wien
4	654	Rudolphi, Caroline	Hammer-Purgstall, Joseph von	1794	1794	5	5	1	1		Hamburg
5	91	Böttiger, Karl August	Hammer-Purgstall, Joseph von	1796	1796	7	7	1	1		Weimar
6	91	Böttiger, Karl August	Hammer-Purgstall, Joseph von	1796	1796	7	7	26	26		Weimar



# Visualisation of the metadata & Python code



python

[Code kopieren](#)

```
import pandas as pd
import matplotlib.pyplot as plt

# Step 1: Load the Excel file
file_path_new = '/path/to/your/file.xlsx' # Update this to the correct path
xls_new = pd.ExcelFile(file_path_new)

# Load the first sheet from the Excel file
df_new = pd.read_excel(xls_new, 'Tabelle1')

# Step 2: Clean the data and rename the columns for easier access
df_cleaned_new = df_new.rename(columns={
    'Schreiber': 'Sender',
```

# Evaluation of the results

- Human-in-the-Loop
- ODD + Schematron
- Comparison with “Gold Standard” from automated workflow
- Comparison of number of tokens and string comparison (possible “correction” of historical spelling)
- Comparison of number and order of elements used
- Future research issues:
  - Benchmark tests for DH data and DH tasks
  - Error categorization and analysis of typical errors for DH data

# Bias and ethical risks

## Training data

- Linguistic, social / cultural / ... / biases and prejudices from training data are reproduced and perpetuated by the AI
- Transparency, traceability and reproducibility remain problematic issues
- Potential for inclusion and equal opportunities, but also risk of exclusion (bias in data or access to systems)

# Conclusion and Outlook

- **Potential for integrating AI into the edition workflow**  
Automation and optimization of processes
- **Domain knowledge is essential**  
Expectations: "AI must be better than humans?"  
Human-in-the-loop for evaluation of results
- **The challenge of evaluating humanities data**  
Lack of metrics and gold standards
- **Use of open source models necessary**  
Promotes transparency and reproducibility of results

# References

- Höflechner, Walter unter Mitarbeit von Alexandra Wagner, Gerit Koitz-Arko und Sylvia Kowatsch. 2018. Joseph von Hammer-Purgstall. Briefe, Erinnerungen, Materialien.  
<http://gams.uni-graz.at/context:hp>
- IBM Technology. 22.3.2023. What are Generative AI models?  
[https://youtu.be/hflUstzHs9A?si=p\\_79GQpMzZYmvbny](https://youtu.be/hflUstzHs9A?si=p_79GQpMzZYmvbny)
- IBM Technology. 5.8.2024. AI, Machine Learning, Deep Learning and Generative AI Explained  
<https://www.youtube.com/watch?v=qYNweeDHiyU>
- Mollick, Ethan. Captain's log: the irreducible weirdness of prompting AIs.  
<https://www.oneusefulthing.org/p/captains-log-the-irreducible-weirdness>
- Pierazzo, Elena, Digital Scholarly Editing. Theories, Models and Methods (Surrey: Ashgate, 2015).
- Pollin, Christopher, Scholger, Martina, Sahle, Patrick, Vogeler, Georg, Schaßan, Torsten, Dumont, Stefan, Fischer, Franz, Henny-Krahmer, Ulrike, Fritze, Christiane, Roeder, Torsten, & Czmiel, Alexander. (2024). Workshop Generative KI, LLMs und GPT bei digitalen Editionen. DHd 2024 Quo Vadis DH (DHd2024), Passau, Deutschland. Zenodo. <https://doi.org/10.5281/zenodo.10893761>

## References

- Pollin, C. (2024). Workshopreihe "Angewandte Generative KI in den (digitalen) Geisteswissenschaften" (v1.1.0). Zenodo. <https://doi.org/10.5281/zenodo.10647754>
- Sahle, Patrick, Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels, 3 Bände, Schriften des Instituts für Dokumentologie und Editorik 7-9 (Norderstedt: BOD, 2013).
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." arXiv, 2017. <https://doi.org/10.48550/arXiv.1706.03762>.