



Midjourney: dreamy tornado of documents, computer, flow of structured data, by Caspar David Friedrich, 19th century scholar, landscape --ar 16:9 --zoom 1.25
Zoom out: data sucking skyscraper-sized screen, dreamy tornado of documents --chaos 5 --ar 16:9 --zoom 1.65



New Ways of Creating Research Data: Conversion of Unstructured Text to TEI XML using GPT on the Correspondence of Hugo Schuchardt with a Web Prototype for Prompt Engineering

FORGE 2023, Tübingen

Christopher Pollin

Zentrum für Informationsmodellierung, Universität Graz

Christian Steiner

Digital Humanities Craft OG

Constantin Zach

Independent Software Developer





**Prompt
Engineering &
Fine-Tuning**

**Erzeugen von
Forschungsdaten**

**Workflow
Unstrukturierter
Text → TEI XML**

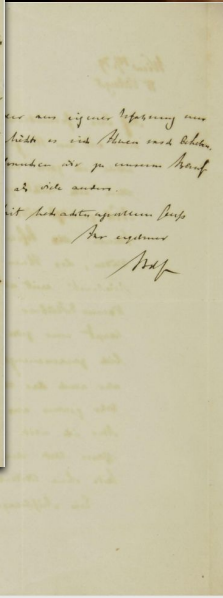
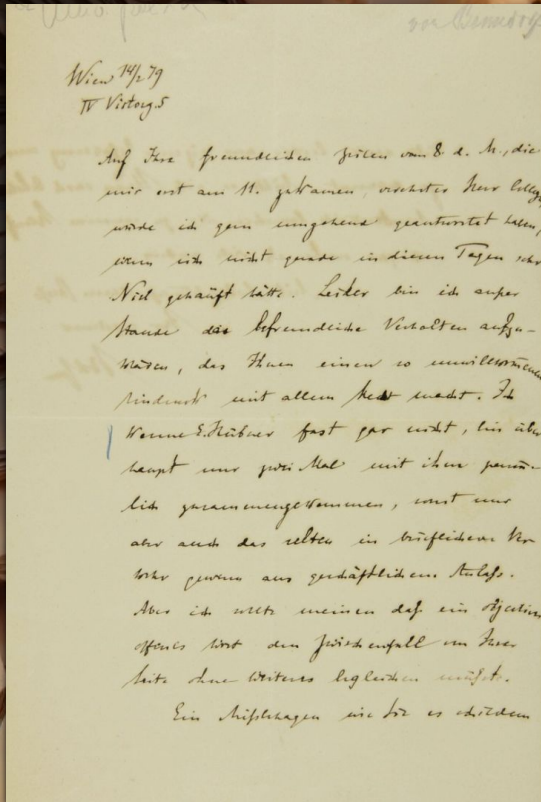
Web Prototyp

**Korrespondenz
von Hugo
Schuchardt**

Agenda

HUGO SCHUCHARDT ARCHIV

<https://schuchardt.uni-graz.at>



Auf den freundlichen Willen von E. A. Schick
 mit und am 11. geborenen Kinder. Das Kind
 wurde ich ganz ungeschadet gezeugt, aber
 man ich nicht gerade in diesem Tagen so
 viel gereizt war. Leider bin ich außer
 Stande die folgenden Wochen auf-
 zuhalten, das Klein einem so unwillkür-
 lich! mit allem Besten. Ich
 komme Schicksal fast gar nicht, bis ich
 klagt um ein paar Mal mit dem gezei-
 gten Zusammenhang, und um
 das auch das Leben in tiefsteren bei
 den ganzen aus gedrücktem Tiefs.
 Aber ich will meinen das ein Apokal-
 ypse hat den Frieden mit im Inn-
 ren den letzten Gedanken nicht.
 Ein Auftragen wie das so schiden

Auf Ihre freundlichen Zeilen vom 8. d. M., die mir erst am 11. zukamen, verehrter Herr College, würde ich gern umgehend geantwortet haben, wenn sich nicht gerade in diesen Tagen sehr Viel gehäuft hätte. Leider bin ich ausser Stande das befremdliche Verhalten aufzuklären, das Ihnen einen so unwillkommenen Eindruck mit allem Recht macht. Ich kenne [E. Hübner](#)³ fast gar nicht, bin überhaupt nur zwei Mal mit ihm persönlich zusammengekommen, sonst nur aber auch das selten in brieflichem Verkehr gewesen aus geschäftlichem Anlass.⁴ Aber ich sollte meinen dass ein objectives offenes Wort den Zwischenfall von Ihrer Seite ohne Weiteres begleichen müsste.

Ein Missbehagen wie Sie es schildern [2] ist mir leider aus eigener Erfahrung nur zu bekannt. Möchte es sich Ihnen rasch beheben. Munterkeit brauchen wir zu unserm Beruf so gar mehr als viele andere.

Mit hochachtungsvollem Gruss

Ihr ergebener

Bdf

Friedrich August Otto Benndorf an Hugo Schuchardt (02-00932). Wien, 14. 02. 1879. Hrsg. von Hubert Szemethy (2022). In: Bernhard Hurch (Hrsg.): Hugo Schuchardt Archiv. Online unter <https://gams.uni-graz.at/o:hsa.letter.7711>, abgerufen am 07. 06. 2023. Handle: hdl.handle.net/11471/518.10.1.7711.

Begriffe

LLM - Large Language Models

GPT, Orca, LLaMa, Mistral, ...

Fine Tuning

Prompt Engineering

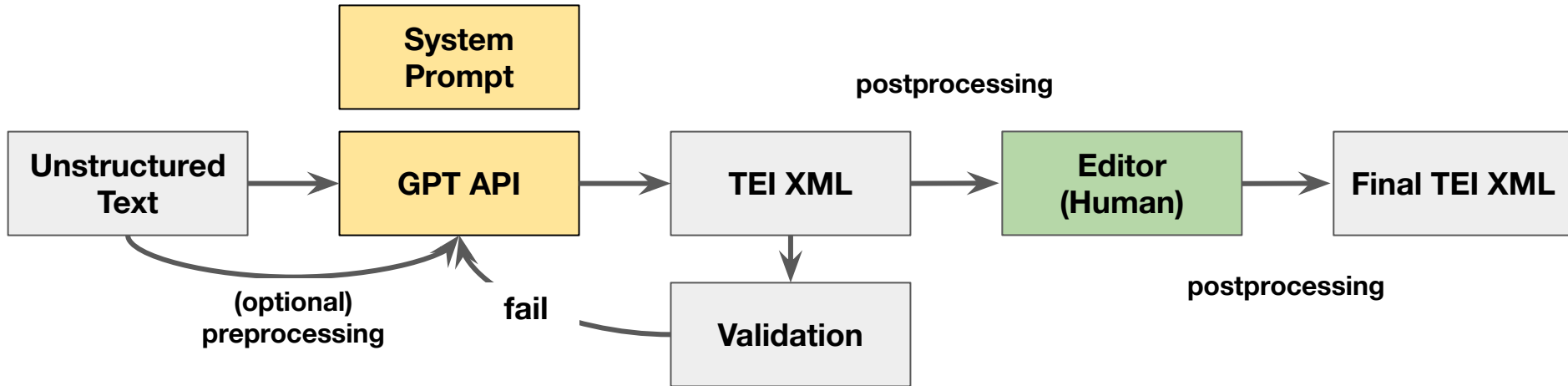
GPT API vs ChatGPT

Interne Begriffe

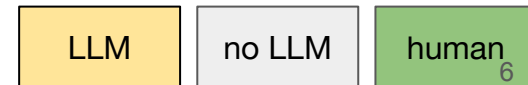
Original-TEI: “echtes” / menschlich generiertes TEI aus dem HSA Projekt

AI-TEI: In unserem Workflow generiertes TEI

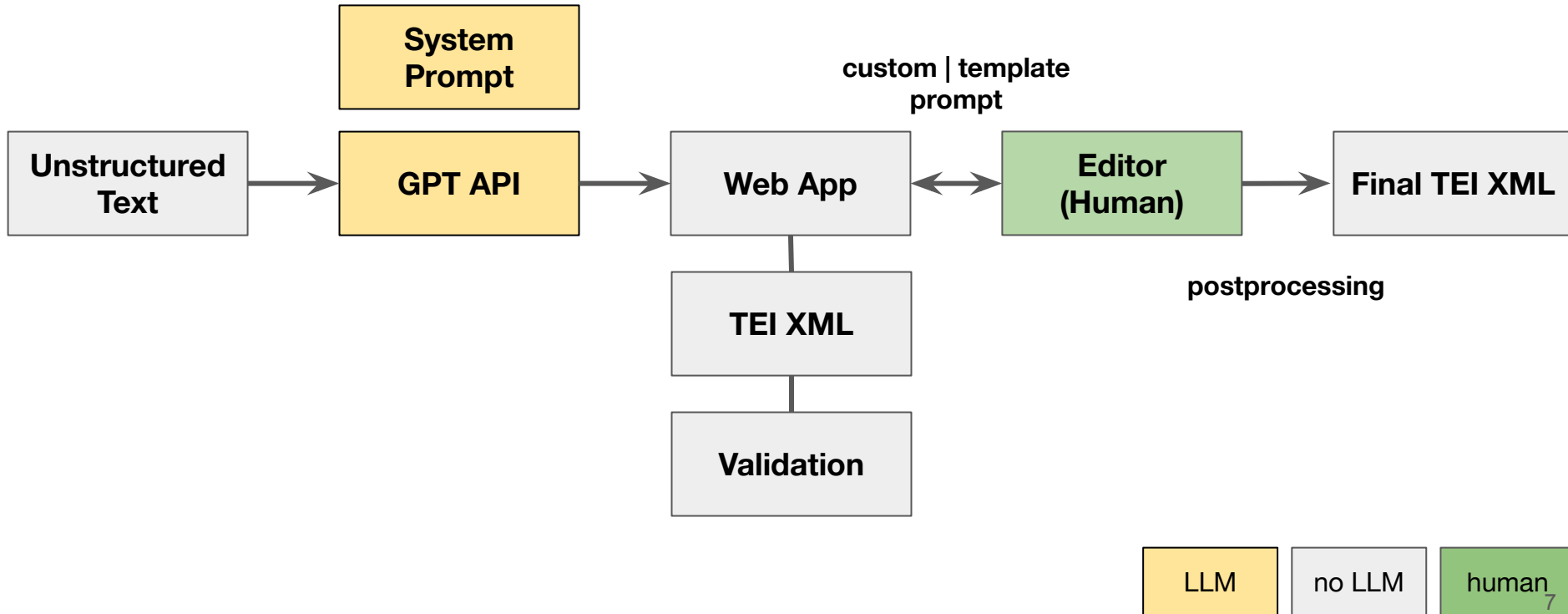
(1) Bulk-Transformation-Workflow: Plaintext to TEI XML (ohne Webapp)



n Briefe werden mittels eines Python Scripts nach TEI XML transformiert



(2) Single-Transformation-Workflow: Plaintext to TEI XML (mit Webapp)



Web-Prototyp-Interface (GPT-3.5)

Collection



Input files

[sa.letter.4495](#)

[sa.letter.1000](#)

[sa.letter.17.txt](#)

[sa.letter.1.txt](#)

Download Input files

Result files

Download Result files

Input Text

Port-Louis, 20 Janvier 1885.

Mon cher Monsieur Schuchardt,

Ma Doudou vous envoie une petite brochure jaune (note 1) à laquelle vous ferez bon accueil, j'en suis sûr.

Nous sommes anxieux l'un et l'autre d'avoir de vos nouvelles, et voudrions bien savoir que ce sont vos travaux, et non la maladie, qui nous privent de toute lettre de vous. Portez-vous bien, c'est le plus sincère de nos vœux, et croyez-moi, mon cher Monsieur Schuchardt,

Votre bien cordialement dévoué

C Baissac

_____1 Welche Broschüre hier gemeint ist, kann nicht mehr ausfindig gemacht werden.

Upload file(s)

Convert current file

TEI XML Preview

```
2 <div xmlns="http://www.tei-c.org/ns/1.0">
3   <dateline>
4     <date>Port-Louis, 20 Janvier 1885</date>.
5   </dateline>
6   <p>
7     Mon cher Monsieur <persName
ref="#schuchardt">Schuchardt</persName>,
8   </p>
9   <p>
10    Ma Doudou vous envoie une petite brochure jaune
(<note>1</note>) à laquelle vous ferez bon accueil, j'en suis
sûr.
11  </p>
12  <p>
13    Nous sommes anxieux l'un et l'autre d'avoir de vos
nouvelles, et voudrions bien savoir que ce sont vos travaux, et
non la maladie, qui nous privent de toute lettre de vous.
14  </p>
15  <p>
16    Portez-vous bien, c'est le plus sincère de nos vœux, et
```

Download Result

I am a fine-tuned model for [project-name]. Let me know what I can do for you.

annotate all persons using <name>

Your request has been processed with the ID: 1006

Strategien zur Überprüfung von Ergebnissen

- Plaintext vergleich Input und Output: weicht ein Text in den Textknoten ab, muss die gesamte Transformation erneut durchgeführt werden.
- Schema: Validierung nach einem sehr spezifischen Schema. Gibt es eine Abweichung, dann muss die Transformation erneut durchgeführt werden: Verwendung von Elementen, die nicht erlaubt sind.
- XPath:
 - Grundsätzliche Regeln überprüfen: “mind 1 <div>”
 - Für unser Experiment: gibts es die selbe Anzahl an Elemente in AI-TEI und Original-TEI
- Experimentell: GPT-4 Review der Ergebnisse
- Menschliches Feedback mit der Webapp bzw. im Workflow nach Bulk-Transformation

[1]

Wien 14/2 79

IV Victorg. 5

Auf Ihre freundlichen Zeilen vom 8. d. M., die mir erst am 11. zukamen, verehrter Herr College, würde ich gern umgehend geantwortet haben, wenn sich nicht gerade in diesen Tagen sehr Viel gehäuft hätte. Leider bin ich ausser Stande das befremdliche Verhalten aufzuklären, das Ihnen einen so unwillkommenen Eindruck mit allem Recht macht. Ich kenne E. Hübner fast gar nicht, bin überhaupt nur zwei Mal mit ihm persönlich zusammengekommen, sonst nur aber auch das selten in brieflichem Verkehr gewesen aus geschäftlichem Anlass. Aber ich sollte meinen dass ein objectives offenes Wort den Zwischenfall von Ihrer Seite ohne Weiteres begleichen müsste.

Ein Missbehagen wie Sie es schildern [2] ist mir leider aus eigener Erfahrung nur zu bekannt. Möchte es sich Ihnen rasch beheben. Munterkeit brauchen wir zu unserm Beruf so gar mehr als viele andere.

Mit hochachtungsvollem Gruss

Ihr ergebener

Bdf

Input Plaintext



You will act as a skilled expert automaton that is proficient in transforming unstructured text, specifically multilingual letters from or to Hugo Schuchardt (1842-1927), 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
- * <pb> Marks page breaks e.g. "|{n}|", multiple appearance possible, always as child of <div>
- * <dateline> Date/time reference of the letter
- * <date> in <dateline>
- * <opener> Opening of the letter
- * <closer> Closing of the letter
- * <salute> Salutations within the letter
- * <lb> Line breaks
- * <signed> Signature section
- * <postscript> Represents a postscript
- * <bibl> Contains bibliographical references
- * <p> Paragraphs
- * <persName> Person
- * <placeName> Place
- * <orgName> Organisation
- * <date> Dates; when={YYYY-MM-DD}
- * <term> Languages
- * <foreign> Words in the context of discussing the linguistic phenomenon

Guidelines:

- * Strictly follow mapping rules
- * Preserve the original text
- * Produce well-formed TEI XML according to TEI standards
- * Return the <div> only
- * Annotate only when appropriate
- * Preserve complexity of output
- * Compact XML without any whitespace or indentation

System Prompt

```

<div subtype="original" type="letter" xml:id="L.7711" xml:lang="de">
  <pb facs="#IMG.1" n="1"/>
    <dateline>
      Wien 14/2
      79<lb/> IV Victorg. 5
    </dateline>
    <p>Auf Ihre freundlichen Zeilen vom 8. d. M., die mir erst am 11. zukamen, verehrter Herr
      College, würde ich gern umgehend geantwortet haben, wenn sich nicht gerade in diesen Tagen
      sehr Viel gehäuft hätte. Leider bin ich ausser Stande das befremdliche Verhalten
      aufzuklären, das Ihnen einen so unwillkommenen Eindruck mit allem Recht macht. Ich kenne
        <persName ref="https://gams.uni-graz.at/o:hsa.persons#P.1805">E. Hübner</persName>
      fast gar nicht, bin überhaupt nur
      zwei Mal mit ihm persönlich zusammengekommen, sonst nur aber auch das selten in
      brieflichem Verkehr gewesen aus geschäftlichem Anlass. Aber ich sollte meinen dass ein objectives offenes Wort
      den Zwischenfall von Ihrer Seite ohne Weiteres begleichen müsste.</p>
    <p>Ein Missbehagen wie Sie es schildern <pb facs="#IMG.2" n="2"/> ist mir leider aus eigener
      Erfahrung nur zu bekannt. Möchte es sich Ihnen rasch beheben. Munterkeit brauchen wir zu
      unserm Beruf so gar mehr als viele andere.</p>
    <closer>
      <salute>Mit hochachtungsvollem Gruss</salute>
      <lb/>
      <salute>Ihr ergebener</salute>
      <lb/>
      <signed>
        <persName ref="https://gams.uni-graz.at/o:hsa.persons#P.1127">Bdf</persName>
      </signed>
    </closer>
  </div>

```

Original-TEI

<note> gekürzt

```

<div>
  <pb n="1"/>
  <dateline>
    <lb/>Wien 14/2 79<lb/>IV Victorg. 5<lb/>
  </dateline>
  <opener>Auf Ihre freundlichen Zeilen vom 8. d. M., die mir erst am 11. zukamen,
verehrter Herr College, würde ich gern umgehend geantwortet haben, wenn sich nicht
gerade in diesen Tagen sehr Viel gehäuft hätte. Leider bin ich ausser Stande das
befremdliche Verhalten aufzuklären, das Ihnen einen so unwillkommenen Eindruck mit
allem Recht macht. Ich kenne E. Hübner fast gar nicht, bin überhaupt nur zwei Mal
mit ihm persönlich zusammengekommen, sonst nur aber auch das selten in brieflichem
Verkehr gewesen aus geschäftlichem Anlass. Aber ich sollte meinen dass ein
objectives offenes Wort den Zwischenfall von Ihrer Seite ohne Weiteres begleichen
müsste.</opener>
    <p>Ein Missbehagen wie Sie es schildern <pb n="2"/> ist mir leider aus
eigener Erfahrung nur zu bekannt. Möchte es sich Ihnen rasch beheben. Munterkeit
brauchen wir zu unserm Beruf so gar mehr als viele andere.</p>
    <closer>
      <salute>Mit hochachtungsvollem Gruss</salute>
      <lb/><salute>Ihr ergebener</salute>
      <lb/><signed>
        <persName ref="[authorityLink]">Bdf</persName>
      </signed>
    </closer>
</div>

```

AI-TEI (fine-tuned GPT-3.5)

ohne <note>
Temperature: 0

- Fine-tuned GPT-3.5 mit 470 .txt und TEI XML trainiert.
 - OpenAI empfiehlt mindestens 50 Datensätze.
- Temperature: 0
stabile Ergebnisse.
- <lb>, <pb>, <persName>, <signed>, <closer> etc. sind korrekt.
- Alles, wofür GPT nicht benötigt wird, wird programmatisch gemacht: Attribute in <div> z.B.
- <opener> ist falsch.
- @ref="[authorityLink]" haben wir inkorrekt im fine-tuning "antrainiert"
- "E. Hübner" hat kein persName bekommen.


```

<div>
  <pb n="1"/>
  <dateline><placeName>Wien</placeName> <date when="1879-02-14">14/2
79</date></dateline>
  <opener>
    <lb/>IV Victorg. 5
    <lb/>
    <salute>Auf Ihre freundlichen Zeilen vom 8. d. M., die mir erst am 11.
zukamen, verehrter Herr
      College,</salute>
  </opener>
  <p>würde ich gern umgehend geantwortet haben, wenn sich nicht gerade in
diesen Tagen sehr Viel gehäuft hätte. Leider bin ich ausser Stande das befremdliche
Verhalten aufzuklären, das Ihnen einen so unwillkommenen Eindruck mit allem Recht
macht. Ich kenne <persName>E. Hübner</persName> fast gar nicht, bin überhaupt nur
zwei Mal mit ihm persönlich zusammengekommen, sonst nur aber auch das selten in
brieflichem Verkehr gewesen aus geschäftlichem Anlass. Aber ich sollte meinen dass
ein objectives offenes Wort den Zwischenfall von Ihrer Seite ohne Weiteres
begleichen müsste.</p>
  <pb n="2"/>
  <p>Ein Missbehagen wie Sie es schildern ist mir leider aus eigener Erfahrung
nur zu bekannt. Möchtees sich Ihnen rasch beheben. Munterkeit brauchen wir zu unserm
Beruf so gar mehr als viele andere.</p>
  <closer>
    <lb/>Mit hochachtungsvollem Gruss
    <lb/><signed>Ihr ergebener
    <lb/>Bdf</signed>
  </closer>
</div>

```

AI-TEI (GPT-4)

ohne <note>
Temperature: 0

- Temperature: 0
stabile Ergebnisse.
- macht <opener> richtig
- <salute> fehlt

Fine-tuned GPT-3.5:

“Das habe ich so gelernt und
das mache ich jetzt immer so”
GPT-4: Reasoning des Textes
→ findet passendere
Annotationen

- Wien ist annotiert und das
Datum normalisiert, was im
Original TEI nicht der Fall ist.
- “E.Hübner” hat ein
<persName>, aber “Bdf” nicht
mehr

```

<div>
  <pb n="1"/>
  <dateline>
    <placeName>Wien</placeName>
    <date when="1879-02-14">14/2 79</date>
  </dateline>
  <opener>
    <address>
      <addrLine>IV Victorg. 5</addrLine>
    </address>
  </opener>
  <p>Auf Ihre freundlichen Zeilen vom <date when="1879-02-08">8. d. M.</date>, die mir
erst am <date when="1879-08-11">11.</date> zukamen, verehrter Herr College, würde
ich gern umgehend geantwortet haben, wenn sich nicht gerade in diesen Tagen sehr
Viel gehäuft hätte. Leider bin ich ausser Stande das befremdliche Verhalten
aufzuklären, das Ihnen einen so unwillkommenen Eindruck mit allem Recht macht. Ich
kenne <persName>E. Hübner</persName> fast gar nicht, bin überhaupt nur zwei Mal mit
ihm persönlich zusammengekommen, sonst nur aber auch das selten in brieflichem
Verkehr gewesen aus geschäftlichem Anlass. Aber ich sollte meinen dass ein
objectives offenes Wort den Zwischenfall von Ihrer Seite ohne Weiteres begleichen
müsste.</p>
  <pb n="2"/>
  <p>Ein Missbehagen wie Sie es schildern ist mir leider aus eigener Erfahrung nur zu
bekannt. Möchte es sich Ihnen rasch beheben. Munterkeit brauchen wir zu unserm Beruf
so gar mehr als viele andere.</p>
  <closer>
  <salute>Mit hochachtungsvollem Gruss</salute>
  <signed>Ihr ergebener Bdf</signed>
  </closer>
</div>

```

AI-TEI (GPT-4)

ohne <note>

Temperature: 1.1

- Mit steigender Temperatur wird GPT-4 kreativer. Kreativität führt zu mehr korrekten Annotationen, aber auch zu mehr ungewollten Annotationen:
<adress>, <addrLine>
- Spannend!:
 - "8. d. M"
 - "11."wurden als Datum erkannt und korrekt normalisiert. Aber es war sehr instabil und ich konnte es nicht reproduzieren.

Wir haben also ein Werkzeug von dem wir wissen, dass wir damit "gutes" (komplexeres) TEI erzeugen können, aber wir wissen noch nicht, wie wir es stabilisieren können. Die entsprechenden Workflows müssen noch entwickelt werden.

Backend: Single-Transformation-Workflow mit Webapp

Rust

axum ...

web application
framework

Postgres

**RESTful API,
OpenAPI-Standard,
Swagger**

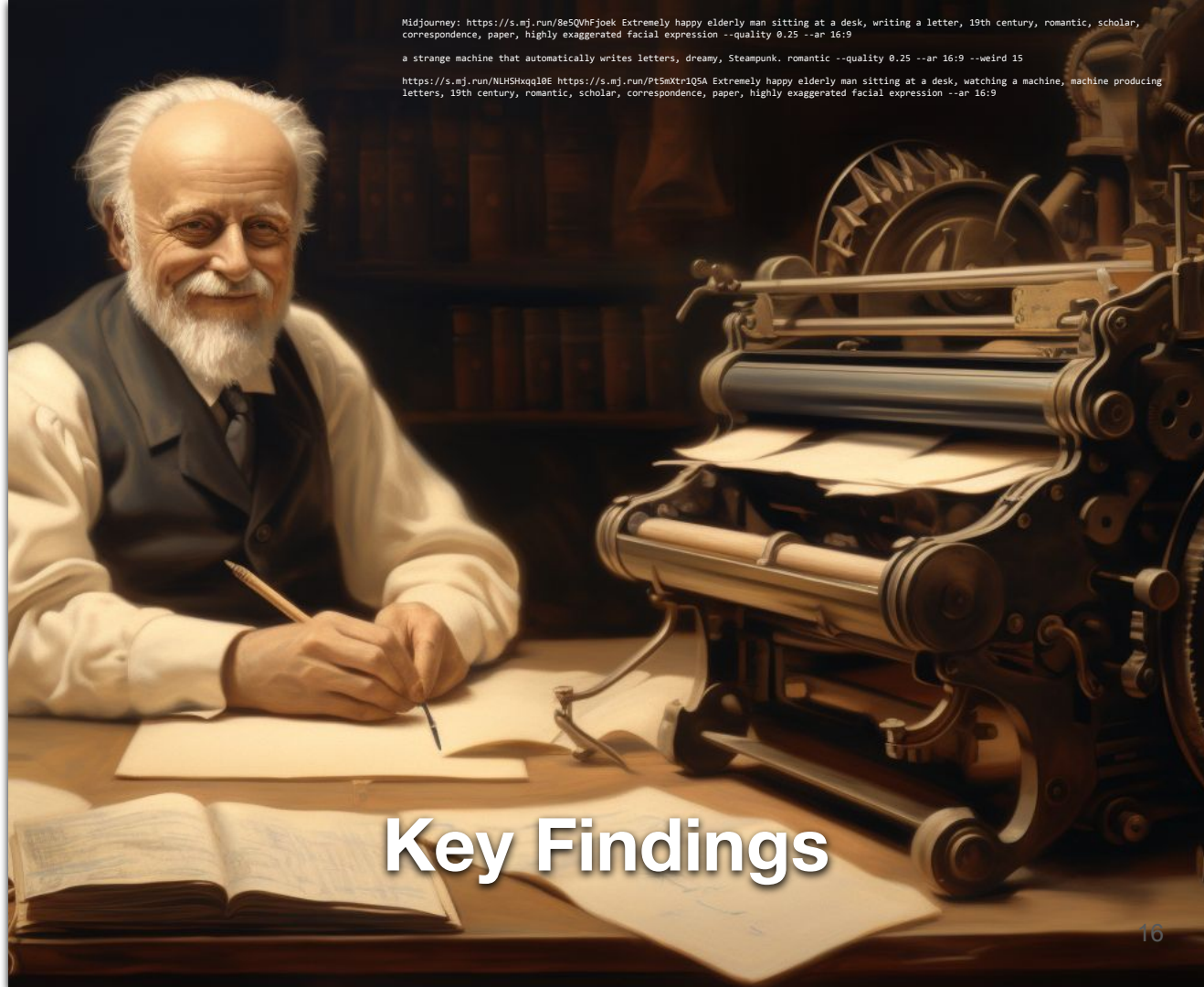
llm-chain

Letter Manage Letters		^
GET	/corpus/{corpus_id}/letter List all Letters	v
GET	/corpus/{corpus_id}/letter/{letter_id} Get single Letter	v
GET	/corpus/{corpus_id}/letter/{letter_id}/txt Get the unstructured text from the Letter	v
POST	/corpus/{corpus_id}/upload Upload one or multiple Letters to a Corpus	v
POST	/corpus/{corpus_id}/letter/{letter_id}/convert Perform initial conversion of unstructured text to TEI.	v
TEI Manage TEIs		^
GET	/corpus/{corpus_id}/letter/{letter_id}/tei List all TEI versions of the given Letter	v
POST	/corpus/{corpus_id}/letter/{letter_id}/tei/upload Upload a new TEI version	v
GET	/corpus/{corpus_id}/letter/{letter_id}/tei/{tei_id} Get the TEI by id	v
GET	/corpus/{corpus_id}/letter/{letter_id}/tei/{tei_id}/xml Get the structured text by id from a Letter	v
GET	/corpus/{corpus_id}/letter/{letter_id}/tei/{tei_id}/prompt Get the prompt with which this TEI was generated	v
Chat Iterative refinement of TEI		^
POST	/corpus/{corpus_id}/letter/{letter_id}/chat Creates a blocking AI model response for the given chat conversation	v
POST	/corpus/{corpus_id}/letter/{letter_id}/chat_stream Creates a streaming AI model response for the given chat conversation	v

- Prompt Engineering: Der Prompt ist wichtig
- Token sparen
- Temperature vs. Stabilität
- fine-tuned GPT-3.5 vs. GPT-4 vs. fine-tuned GPT-4 (?)
- Umsichtige Verifikationsstrategie erforderlich
- Menschliche Bewertung wird wichtiger!
- Kostenpunkt API

AI-Unterstützung und keine vollkommene Automatisierung!

Ziel: Teile des Workflows werden effizienter, aber menschliche Bearbeitung unausweichlich.



Midjourney: <https://s.mj.run/8e5QVnFjoek> Extremely happy elderly man sitting at a desk, writing a letter, 19th century, romantic, scholar, correspondence, paper, highly exaggerated facial expression --quality 0.25 --ar 16:9

<https://s.mj.run/NLHSHxqql0E> <https://s.mj.run/Pt5mXtr1Q5A> a strange machine that automatically writes letters, dreamy, Steampunk, romantic --quality 0.25 --ar 16:9 --weird 15

<https://s.mj.run/NLHSHxqql0E> <https://s.mj.run/Pt5mXtr1Q5A> Extremely happy elderly man sitting at a desk, watching a machine, machine producing letters, 19th century, romantic, scholar, correspondence, paper, highly exaggerated facial expression --ar 16:9

Key Findings

Anhang

Verbesserungen

- Function Calling (im Fine Tuning coming soon)
 - Wikidata Reconciliation
- GPT-4 Fine Tuning (coming soon)
- Named entity recognition
- Complex Prompt Engineering Templates Chain of Thought Prompting
- Pseudo-Markup
- Vision information in the digitized image as context
- Open Source LLMs statt OpenAI
- ...

You will act as a skilled expert automaton that is proficient in transforming unstructured text, specifically multilingual letters from or to Hugo Schuchardt (1842-1927), 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
- * <pb> Marks page breaks e.g. "`{n}`", multiple appearance possible, always as child of <div>
- * <dateline> Date/time reference of the letter
- * <date> in <dateline>
- * <opener> Opening of the letter
- * <closer> Closing of the letter
- * <salute> Salutations within the letter
- * <lb> Line breaks

Prompt 1/2

- * <signed> Signature section
- * <postscript> Represents a postscript
- * <bibl> Contains bibliographical references
- * <p> Paragraphs
- * <persName> Person
- * <placeName> Place
- * <orgName> Organisation
- * <date> Dates; when={YYYY-MM-DD}
- * <term> Languages
- * <foreign> Words in the context of discussing the linguistic phenomenon

Guidelines:

- * Strictly follow mapping rules
- * Preserve the original text
- * Produce well-formed TEI XML according to TEI standards
- * Return the <div> only
- * Annotate only when appropriate
- * Preserve complexity of output
- * Compact XML without any whitespace or indentation

Collection



Input files

[sa.letter.4495](#)
[sa.letter.1000](#)
[sa.letter.17.txt](#)
[sa.letter.1.txt](#)

Download Input files

Result files

Download Result files

Input Text

Port-Louis, 20 Janvier 1885.

Mon cher Monsieur Schuchardt,

Ma Doudou vous envoie une petite brochure jaune (note 1) à laquelle vous ferez bon accueil, j'en suis sûr.

Nous sommes anxieux l'un et l'autre d'avoir de vos nouvelles, et voudrions bien savoir que ce sont vos travaux, et non la maladie, qui nous privent de toute lettre de vous. Portez-vous bien, c'est le plus sincère de nos vœux, et croyez-moi, mon cher Monsieur Schuchardt,

Votre bien cordialement dévoué

C Baissac

_____1 Welche Broschüre hier gemeint ist, kann nicht mehr ausfindig gemacht werden.

Upload file(s)

Convert current file

TEI XML Preview

```
2 <div xmlns="http://www.tei-c.org/ns/1.0">
3   <dateline>
4     <date>Port-Louis, 20 Janvier 1885</date>.
5   </dateline>
6   <p>
7     Mon cher Monsieur <persName
8       ref="#schuchardt">Schuchardt</persName>,
9   </p>
10  <p>
11    Ma Doudou vous envoie une petite brochure jaune
12    (<note>1</note>) à laquelle vous ferez bon accueil, j'en suis
13    sûr.
14  </p>
15  <p>
16    Nous sommes anxieux l'un et l'autre d'avoir de vos
17    nouvelles, et voudrions bien savoir que ce sont vos travaux, et
18    non la maladie, qui nous privent de toute lettre de vous.
19  </p>
20  <p>
21    Portez-vous bien, c'est le plus sincère de nos vœux, et
```

Download Result

I am a fine-tuned model for [project-name]. Let me know what I can do for you.

annotate all persons using <name>

Your request has been processed with the ID: 1006