



Rechtsquellenstiftung  
Fondation des sources du droit  
Fondazione per le fonti giuridiche

des Schweizerischen Juristenvereins  
de la Société suisse des juristes  
della Società svizzera dei giuristi



Universität St.Gallen

## Anwendung generativer KI zur Digitalisierung gedruckter Editionen am Beispiel der Sammlung Schweizerischer Rechtsquellen

W8: Generative KI, LLMs und GPT bei digitalen Editionen.  
DHd2024, Passau 26.02.2024–01.03.2024

Bastian Politycki · [bastian.politycki@unisg.ch](mailto:bastian.politycki@unisg.ch)

Kontext: Die Sammlung Schweizerischer Rechtsquellen

Ein Experiment zur Anwendung generativer KI

Ausgangslage: Rechtsquellen des Kantons Zug

Erste Gehversuche

Prototyp einer möglichen Pipeline

Ausblick

Quellcode, Trainingsdaten, Dokumentation und Auswertung:

- Github: <https://github.com/SSRQ-SDS-FDS/ssrq-retro-lab>
  - v1-Branch: <https://github.com/SSRQ-SDS-FDS/ssrq-retro-lab/tree/v1-ocr-and-classification>
  - v2-Release: <https://github.com/SSRQ-SDS-FDS/ssrq-retro-lab/releases/tag/v2.0.0>
- Zenodo: <https://doi.org/10.5281/zenodo.10683209>

# **Kontext: Die Sammlung Schweizerischer Rechtsquellen**

---

- Erschliessung und Edition von Rechtsquellen der gesamten Schweiz vom Mittelalter bis zur Helvetik (1798)
- (Wissens- und Informations-)Sammlung als Beitrag zur Grundlagenforschung in der Rechts-, Verfassungs-, Wirtschafts-, Sozial-, Orts- und Regionalgeschichte, Kirchengeschichte sowie historischen Sprachforschung
- Mehr als 140 Editionseinheiten
  - Beschluss zur (Retro)Digitalisierung 2009
  - Alle Bände vollständig gescannt
  - Im Web als PDF frei zugänglich (gilt auch für neu erschienene Bände)
  - Geführter Zugriff auf die Sammlung über die ‚Blättermaschine‘
  - Metadaten in einem proprietären XML-Format

- Anwendungen / Infrastruktur
  - Projektwebsite
  - SSRQ online (retrodigitalisierte Sammlung)
  - editio (born-digital Edition)
  - Dokumentation und Schema
  - verschiedene Entitätsdatenbanken
  - SSRQ Index
  - ...
- Technologien
  - Datenbanken etc.: PostgreSQL, eXist-DB, Fuseki
  - Software / Frameworks: Bootstrap, jQuery, TEIPublisher, XRX, ...
  - Programmiersprachen: JavaScript, OCaml, Perl, Python, SQL, SPARQL, XQuery

### Aufgabenfelder:

- Korrektur des OCR-Ergebnisses
- Strukturierte Auszeichnung in TEI-XML
- Annotation von Daten und Named-Entities
- Integration / Verknüpfung mit bestehenden Datensätzen
- Integration in digitale Edition (editio)

## Wieso nicht ...?

- Neuedition („from scratch“)
- manuelle Aufarbeitung
- programmatische Aufarbeitung („Scripting“ in Kombination mit klassischen NLP-Methoden)

## Wieso nicht ...?

- Neuedition („from scratch“)
- manuelle Aufarbeitung
- programmatische Aufarbeitung („Scripting“ in Kombination mit klassischen NLP-Methoden)

→ Kosten, Zeit, Ressourcen & in der Vergangenheit wenig erfolgreich

**Einsatz von LLMs:** niedrighschwelliger Zugang zu KI-Methoden, wenig Trainingsmaterial notwendig, schnelle Ergebnisse, flexible Einsatzmöglichkeiten

# Ein Experiment zur Anwendung generativer KI

---

- 3 Bände: 2 Editionsبände und ein Registerband
- Bearbeitet von Eugen Guber und Peter Stotz (Register)
- Zeugnisse der Verwaltungs- und Rechtsgeschichte von 858 bis 1798
- Sprachen: Latein, Deutsch
- Volltranskription und / oder Regest (keine Faksimiles oder Abbildungen)



378 II. Stadt und Amt

1637 Mai 11.

584. *Stadt- und Amtrat, Beginn*

*Der Stadt- und Amtrat beginnt im Sommer um 9 Uhr, im Winter um 10 Uhr. Der Ratsherr ist bei seinem Eide zu pünktlichem Erscheinen verpflichtet. Wegbleiben ohne begründete Ursache ist strafbar.*

5 *StAR 1, 79. – Vgl. auch ebd. 1, 122. Unberechtigtes Wegbleiben wird mit 2 lb geahnd. – Vgl. ebenso 2, 342; 3, 1, 39, 364; 4, 260.*

1644 Juli 28.

585. *Amtsgeheimnis*

10 *Ratsherren und Diener sollen das Amtsgeheimnis wahren, nichts aus dem Rat und von den Malefizanten im Turm ausschwatzen, ansonst sie für drei Monate des Amtes und des Dienstes enthoben sind und 50 lb Buße zu erstatten haben. Und wan auch sonderbar hersonen of myn herren argwonen wurden und in solches verwysen tetten, sollen umb*

15 *60 lb oder 35 kronen abgebißt werden.*

*StAR 3, 22. – Vgl. auch ebd. 6, 86.*

1671 Juni 26.

586. *Gebühr für gekauften Stadt- und Amtrat*

*Wer fürderhin Stadt- und Amtrat begehrt, soll zuvor 24 Kronen erlegen.*

20 *StAR 6, 221.*

1691 März 23.

587. *Die Instruktionen für die Ehrengesandten der Tagsatzung gehen von ihren Gemeinden aus.*

*StAR 11, 114.*

25 *1695 August 29.*

588. *Standesvertretung bei fürstlichen Ritten*

*EA VI, 2, S. 566, Nr. 308 f; dem S. 579, Nr. 312 b; S. 590, Nr. 318 q; S. 610, Nr. 330 g; S. 624, Nr. 335 nnn; S. 909, Nr. 493 a.*

1702 Januar 2.

589. *Stipendien müssen in Stadt und Amt zur freien Bewerbung ausgekündigt werden.*

*StAR 14, 265.*

1722 Mai 11.

590. *Wegen unanständigem Tabakschmauchen während der Ratszeit vor der*

35 *Ratsstube auf dem vordern, hintern und untern Estrich sowohl der*

B. Die städtischen Vogteien · 1554–1556

853

lange nach urwüßit undt zughdith. Ist der kauff geschehen umb ein hundert und zwenzig guldin. *BAHünenberg, Gemeinde-Libell 1664, S. 2.*

<sup>1</sup> *AG.* <sup>2</sup> *Loswald zu Meisterswil Hlg.* <sup>3</sup> *Farb.* abgegangen. <sup>4</sup> *Urmis.* abgegangen. <sup>5</sup> *Hlg.*

<sup>6</sup> *Chamau Hlg.* <sup>7</sup> *Die Herren von Schwarzenberg, Seitenlinie der Freien von Eschenbach-Schnabelburg.*

1415 Juni 27.

1555. *Kauf von Hünenberg*

*Die ganze Gemeinde zu Hünenberg, der obere und der untere Teil, erwirbt von Gütz von Hünenberg käuflich Twing und Bann und 10 Gerichte zu Hünenberg, ebenso die Leute und Güter, welche dorthin gehören, sowie alle Rechtung an Holz und Feld, Wunn und Weid, Steg und Weg, Wasser und Wasserfluß.*

*Or. im Korp. A. Zug, als Decke verwendet zum Marchodol von 1534; es wird ein Gerichtenscheid aus der Zeit des Hermann Bluntschi, 1539 Obervogt zu Hünenberg, wiedergegeben und darin auf ein nicht mehr vorhandenes Original-Kaufinstrument hingewiesen. – Reg.: UBZG Nr. 2447 (546 a).*

1416 Januar 21.

1556. *Anschluß an Zug*

*Die drei Brüder Hans, Ulrich und Heinrich Bätler von Hünenberg und ihr Vetter, Welti Bätler von Stadelmatt, haben laut Kaufbrief vom 17. Januar 1414 die Güter und Rechte des Junkers Hartmann von Hünenberg an sich gebracht und auch den anderen Leuten des Twings zu Hünenberg Anteil an den Rechtsamen und Nutzungen zuerkannt. Auf ihre Bitte regeln Ammann und Rat von Zug die Angelegenheit. [1] Was von dem teil, so sy köfft habent von junkher Hartman von Hünenberg, als der köffbrief wist, nutzes vallet, es sy von holts oder von velde, mit wunn, mit weide, wie sich der nutz erfunde, nit usgenomen, da sol man jeglichem, so vor und hienach geschriben stat, oder iren erben, so vil nutzes lassen volgen nach markzal, als sy den ir güt ufgeben habent an disen köff. [2] Were, daz die vorgeant oder diß nachgeschriben oder ir erben deheinst verköften uß disem teil, es were holts oder felde oder ander nutz, so davon viel, so sölt aber jeglichem werden, so vil als er usgeben hetti, ouch nach markzal. [3] Was ouch nutzes vallet ze Hünenberg in dem teil, so köfft ist von junkherr Hartman, es sy von gerichtten oder sust, die selben nutz sol ein vogt, der in zû geben ist von unser stat, teilen, und die vier, so gesetz[x] sint under innen ze den urteilen vil den eyden, so sy getâm habent, aber jeglichem in dem twing ufze-*

# Ausgangslage: Rechtsquellen des Kantons Zug

B. Die städtischen Vogesen - 1554-1556 853

lange nach überliß nach reglich. In der kauf geschien und ein besiedet und zwenig gütlin. *St. Hiltensberg, Gemündel-Lied 1564, S. 2.*

\* *St.* \* *Land* in *Münster* *St.* \* *Fast.* *abgegeben.* \* *Ums.* *abgegeben.* \* *Stg.* \* *Chausse* *Stg.* \* *Die Herren von Schwyzberg, Substituten der Herren von Entschelberg.*

1415 Juni 27. ←

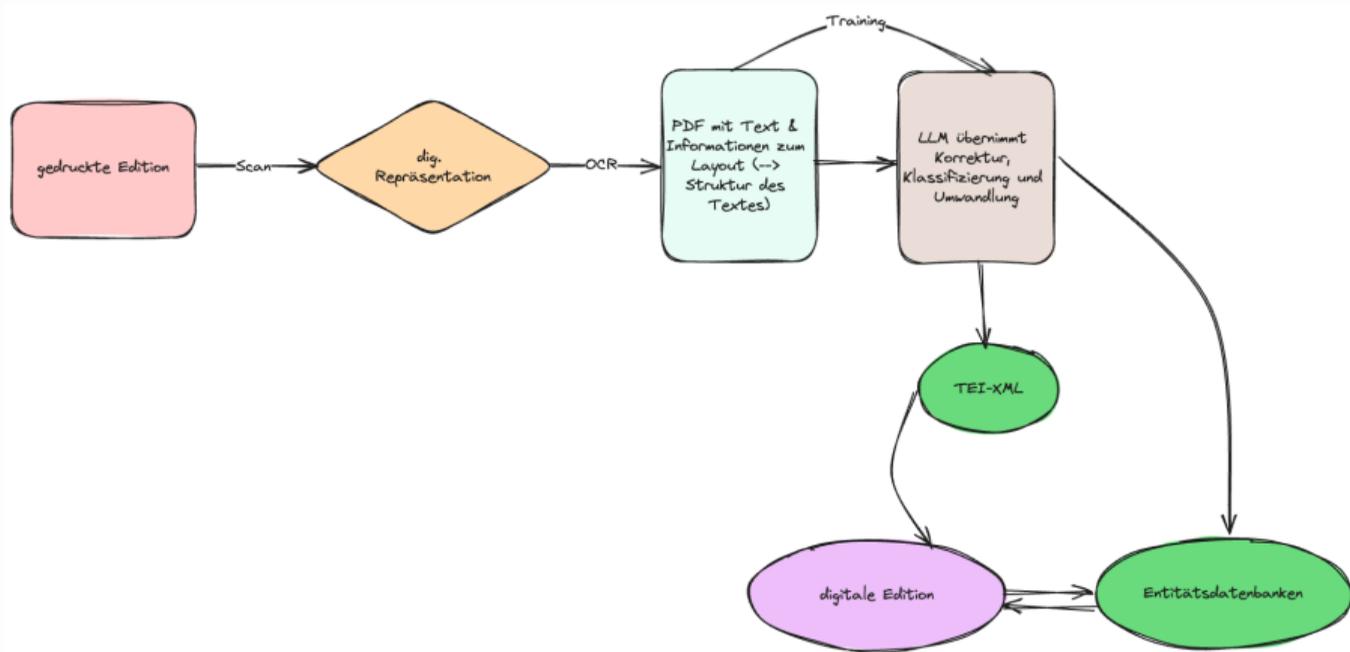
1555. ← **Kauf von Hünenberg**  
Die ganze Gemeinde in Hünenberg, der obere und der untere Teil, erwirbt von Gütern von Hünenberg nämlich Tüwing und Haux und die Gerichte zu Hünenberg, ebenso die Lehen und Güter, welche dorthin gehören, sowie alle Herbergung an Holz und Feld, Wägen und Weid, Stug und Wigg, Wasser und Wasserfisch.

1556. ← **1416 Januar 21. Anschlag an Zug**  
Die drei Brüder Hans, Ulrich und Heinrich Bähler von Hünenberg und ihr Vater, Wolt Bähler von Stadelbach, haben laut Kaufbrief vom 17. Januar 1414 die Güter und Rechte des Junkhens Hartmann von Hünenberg an sich gebracht und auch den anderen Leuten die Tüwing an Hünenberg Anteil an den Rechten und Nutzungen anerkannt. Auf ihre Bitte regeln Ammann und Rat von Zug die Angelegenheit. [1] Was von dem teil, so sy köhlt lobent von junkher Hartman von Hünenberg, als der kühnbrief wirt, nitzen vallet, so sy von holtz oder von veldt, mit wunn, mit weide, wie sich der nitzen erkunde, mit wagnemen, da wil man jeglichem, so vor und hiemach geschriben stat, oder iren erben, so vil nitzen lassen volgen nach marktal, als sy den ir güt ufgeben lobent an dem küh. [2] Were, das die vorgenant oder dñ nachgeschriben oder ir erben dabeint verköfen vil nitzen teil, so were holtz oder feldt oder ander nitzen, so darvon vil, so silte aber jeglichem werden, so vil als er negeben hett, auch nach marktal. [3] Was auch nitzen vallet an Hünenberg in dem teil, so köhlt ist von junkhert Hartman, so sy von gerichtten oder nitzen, die silten nitzen vil ein vugt, der in si geben ist von nitzen nitzen, talen, und die vore, so gontet[?] nitzen nitzen in dem nitzen by dem eyden, so sy gonten lobent, aber jeglichem in dem twing nitzen.

Durchlaufende Kopfzeile  
Quellnachweis  
Datumszeile  
Titel  
Regest  
Regest  
Quellnachweis  
Regest  
Transkript

Stück-zähler

# Idealtypischer Workflow mithilfe von LLMs



## Correcting Swiss Law Text

January 8, 2024



**Anonymous**

Correct the following text snippet, which was extracted via OCR from the Swiss Law Sources – don't modernize:

[2] Wan ouch die weibeU iren eyden, was sy ein aman oder statthalter heißen wurde, nit gnuog thuon wurden und, was in gebotten, nit statt thätten, so soll ein jeder gmeind rathsherren den selbigen weibei, der ancklaggt wirt, lassen gefenklich annemen und inne 25 gan Zug in thurn füren lassen.



**ChatGPT**

Certainly! Here's the corrected version of the text snippet from the Swiss Law Sources, without modernizing:

[2] Wan ouch die weibei iren eyden, was sy ein amman oder statthalter heißen wurde, nit gnuog thuon wurden und, was in gebotten, nit statt thätten, so soll ein jeder gemeind rathsherren den selbigen weibei, der ancklaggt wirt, lassen gefenklich annemen und in den Zug in thurn füren lassen.



**Anonymous**

Identify persons, places and dates.



**ChatGPT**

From the provided text snippet, here are the identifications of persons, places, and dates:

1. **Persons:**

- \* "weibel" (potentially an official or a representative role in this context)
- \* "amman" (likely a variant spelling of "ammann," a title for a community leader or official)

Get started with ChatGPT

- Prompt Engineering (Zero-Shot, Few-Shot, Chain of Thought, ...)
- Wissensanreicherung („Chat with your Documents“-Ansatz)
  - Custom GPTs mit ‚Knowledge‘
  - Retrieval-Augmented Generation (RAG): Textaufbereitung mit Embeddings, Semantic Search, Kontextanreicherung des Prompts
- Training von LLMs mit eigenen Daten (Fine-tuning)
  - Instruction tuning
  - Knowledge distillation

Relevante Arbeiten (in Auszügen):

- Chae, Youngjin, und Thomas Davidson. „Large Language Models for Text Classification: From Zero-Shot Learning to Fine-Tuning“. Preprint. SocArXiv, 24. August 2023. <https://doi.org/10.31235/osf.io/sthwk>.
- Møller, Anders Giovanni, Jacob Aarup Dalsgaard, Arianna Pera, und Luca Maria Aiello. „Is a Prompt and a Few Samples All You Need? Using GPT-4 for Data Augmentation in Low-Resource Classification Tasks“. arXiv, 26. April 2023. <http://arxiv.org/abs/2304.13861>.
- Sun, Xiaofei, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, und Guoyin Wang. „Text Classification via Large Language Models“, o. J.
- Zhou, Wenxuan, Sheng Zhang, Yu Gu, Muhao Chen, und Hoifung Poon. „UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition“, 2023. <https://doi.org/10.48550/ARXIV.2308.03279>.

- Teilbereiche:
  - Korrektur des OCR-Ergebnisses
  - Erkennung von Dokumentgrenzen
  - Klassifizierung der erkannten Dokumente (Transkript, Regest oder Fragment)

## Experiment: OCR-Korrektur, Dokumenterkennung, Klassifizierung

- Teilbereiche:
  - Korrektur des OCR-Ergebnisses
  - Erkennung von Dokumentgrenzen
  - Klassifizierung der erkannten Dokumente (Transkript, Regest oder Fragment)
- Untersuchungsausschnitt:
  - Ground Truth Transkription von 53 Einzelseiten
    - 16 aus Band 1
    - 37 aus Band 2
    - Umfang: 2033 Zeilen und 19207 Worte
  - Basis für Auswertung der bisherigen Qualität
  - Grundlage für Trainingsdatensatz zur OCR-Korrektur

# Aufbau Trainingsdatensatz

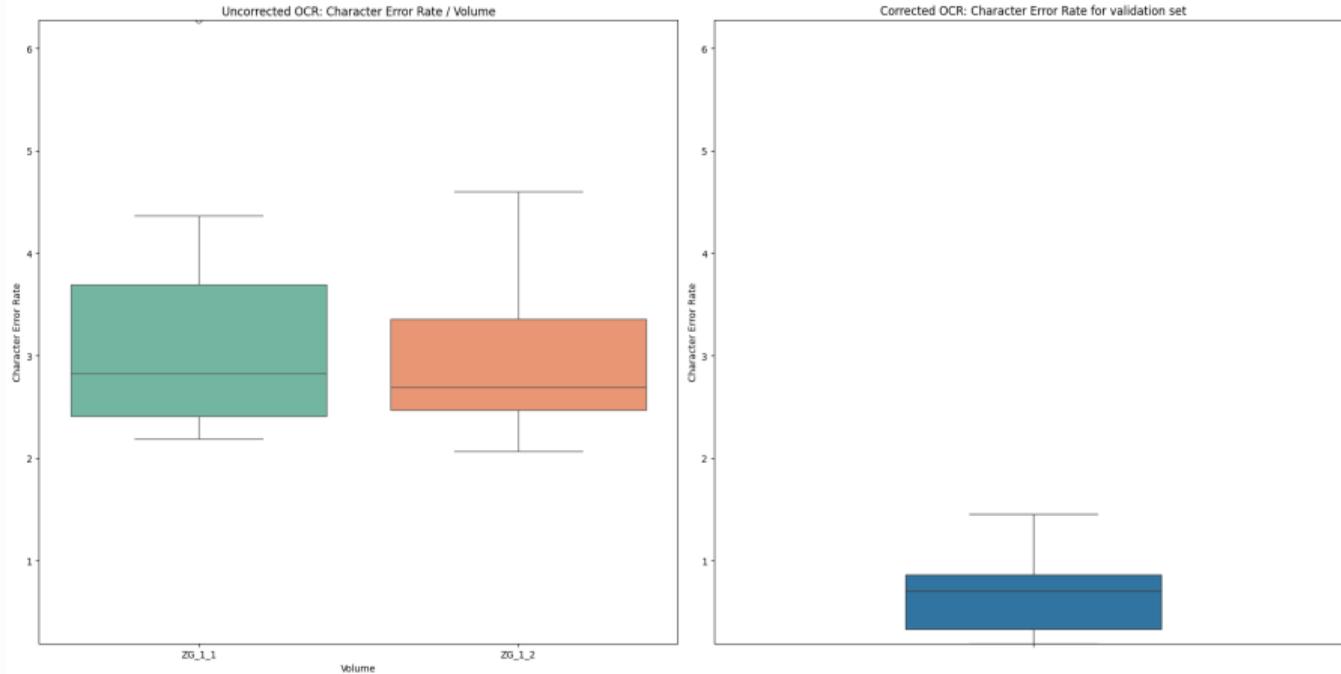
```
{
  "role": "system",
  "content": "You are an helpful research assistant named Walther, who helps researchers to correct
↵ texts, classify them and extract information from them. You have an extremely good knowledge in
↵ scholarly editing."
},
{
  "role": "user",
  "content": "Take a deep breath and look at the following text created with OCR. Correct the text,
↵ line by line, without modernizing it. Keep the linebreaks. The text starts and ends with ```
↵ as markers. Here is the text: \n\n ```\n346 \nII. Stadt und Amt\nfürsprech der justiz wider
↵ den armen sündler klaget und begehrt \ndie achte urtheil. Ob es besser oder wegerer seye, daß
↵ der arme ...```"
},
```

Aufbau des Trainingsprompts für OCR-Korrektur (Systemnachricht und Benutzereingabe)

```
{  
  "role": "assistant",  
  "content": "Certainly! I'll provide the corrected version of  
  ↪ the OCR text, because I'm your helpful assistant. Here's  
  ↪ the corrected text:\n\n  ```\n346 II. Stadt und  
  ↪ Amt\nfürsprech der justiz wider den armen sündler klaget und  
  ↪ begehrt\ndie achte urtheil. Ob es besser oder wegerer seye,  
  ↪ daß der arme ..```\n}"
```

Aufbau des Trainingsprompts für OCR-Korrektur (Antwort)

# OCR-Qualität nach Korrektur durch trainiertes GPT-3.5 Modell



# Auswertung OCR, Dokumenterkennung, Textklassifikation

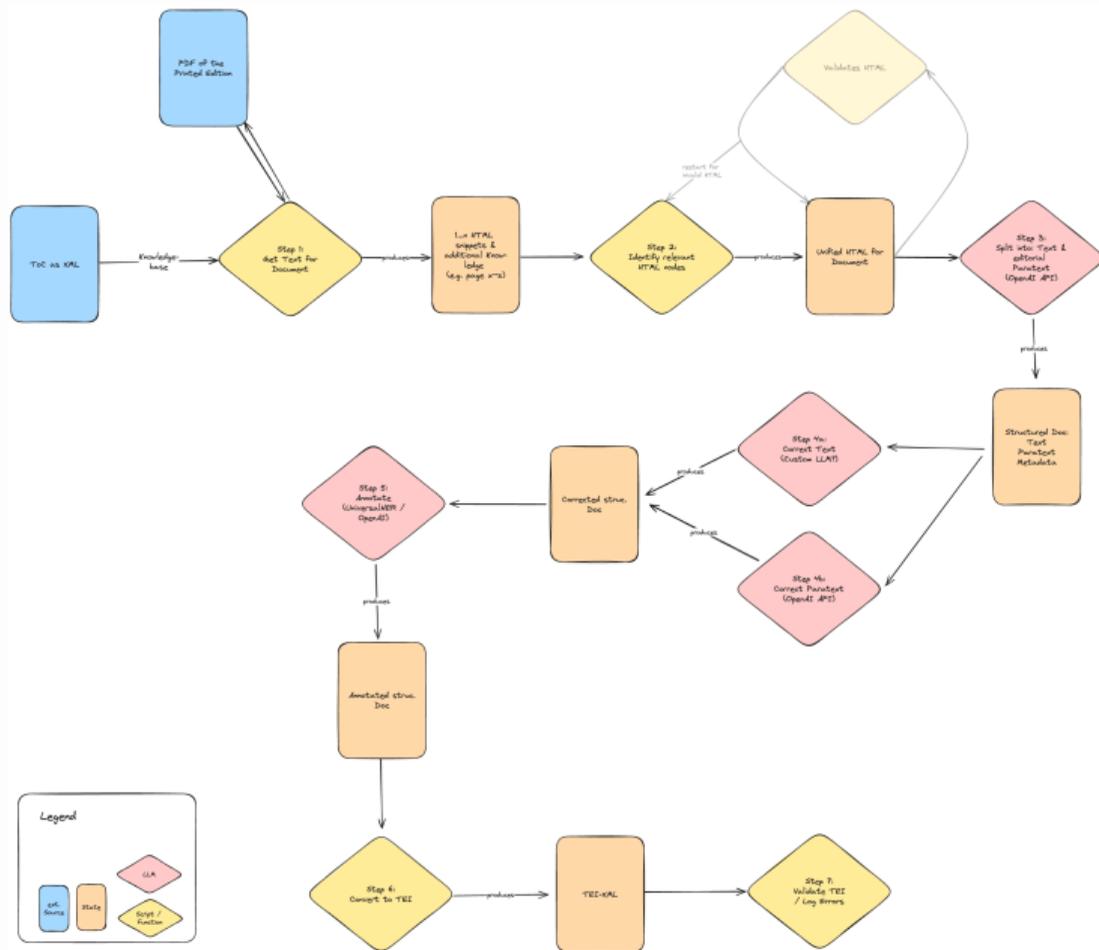
	CER	Kosinusähnlichkeit	F1	Genauigkeit
OCR	0.7	0.99		
Dokumenterkennung (Anzahl)			1.0	1.0
Dokumenterkennung (Position)			0.14	0.14
Klassifikation			0.28	0.28

Ergebnisse der Auswertung der Dokumenterkennung und Klassifikation, siehe: [Validation-Notebook](#) im Github-Repo

## Probleme:

- Prompts für Dokumenterkennung und Klassifikation nicht optimiert
  - A section represents the transcribed text of a document. This section usually starts with a date line and a title. Analyse the text and divide it into the different sections. Mark the start of a section with [START]. **Headers can be ignored.**
- Under- oder Overfitting des Modells
- teilweise unstrukturierte Textdaten erschweren Auswertung
- Arbeit auf Plaintext
- Arbeit gegen das Inhalts- / Denkmodell der Edition: Umwandlung von Druckseiten statt Dokumenten

- atomarerer Aufbau
- zusätzlicher Input; XML-ToC enthalten: Titel, Seitenangaben, Datum, Stücknummer
- Extraktion von HTML aus dem PDF statt Plaintext
- Kombination lokaler Funktionsaufrufe sowie Anfragen an ein LLM
- sofern möglich: Verwendung lokaler Modelle statt OpenAI



1. Extraktion der relevanten aus dem PDF und Umwandlung zu HTML mit PyMuPDF
2. Aufbereitung mithilfe der Informationen aus dem XML-ToC & XPath

→ Reihe von relevanten HTML-Ausschnitten mit Angaben zur Typografie im Buch (Schriftart, Schriftgröße, Schriftschnitt, Position auf der Seite)

Beispiel: GitHub-Repo

- Verwendung von GPT-3.5-turbo über die OpenAI-API
- Few-Shot-Prompting
- JSON-Ausgabe mit festem Schema
- 7 repräsentative Beispiele für erwartete Klassifikationen
- anschließende Validierung des Ergebnisses

You are an expert in Text Classification with proven knowledge in Textual Scholarship. Your task is to accept Text as HTML `p`-tag as input and extract the text nodes of this element and classify them. The paragraph is extracted of an printed book scanned with OCR. Each paragraph is equivalent to a line in the printed book. The book is a scholarly edition and part of the Swiss Law Sources. It mainly contains texts in old and middle German.

...

Assume these definitions are written by an expert and follow them closely.

1. ...

2. **TEXT**: Marks the main text / transcript. Is written in old or middle german and is typeset in a regular font. Respect the following constraint: A paragraph can never contain a **TITLE** and **TEXT**. If there is a text node in italic in the same paragraph as text nodes, which are **TEXT**, it is most likely a **SUMMARY**.

Try to classify the full text. Include only the text nodes in the output. Remove any HTML tags. Enclose your answer in three backticks. The language should be set to json.

Your response must follow the format below:

...

Here are some examples of the input and output:...

Here is the paragraph to classify – it belongs to the document with the number 23.: ...

```
{
  "text": "<p style='top:176.1pt;left:99.9pt;line-height:10.0pt'><i><span style='font-family:Times New
↳ Roman,serif;font-size:10.0pt'>Von Seiten der Bürgerschaft wird der Anzug gemacht, wie das bei
↳ so</span></i><span style='font-family:Times New Roman,serif;font-size:10.0pt'> </span><span
↳ style='font-family:Times New Roman,serif;font-size:10.0pt'>Kurtz auffgesetzte reforme</span></p>",
  "spans": [
    {
      "text": "Von Seiten der Bürgerschaft wird der Anzug gemacht, wie das bei so",
      "label": "SUMMARY",
      "reason": "Summary created by the editors in italics; can't be a title, because it appears in the
↳ same line as the title"
    },
    {
      "text": "Kurtz auffgesetzte reforme",
      "label": "TRANSCRIPT",
      "reason": "Transcript text - old / middle german - in a regular font"
    }
  ]
}
```

Beispiel Textklassifikation für Few-Shot-Prompting

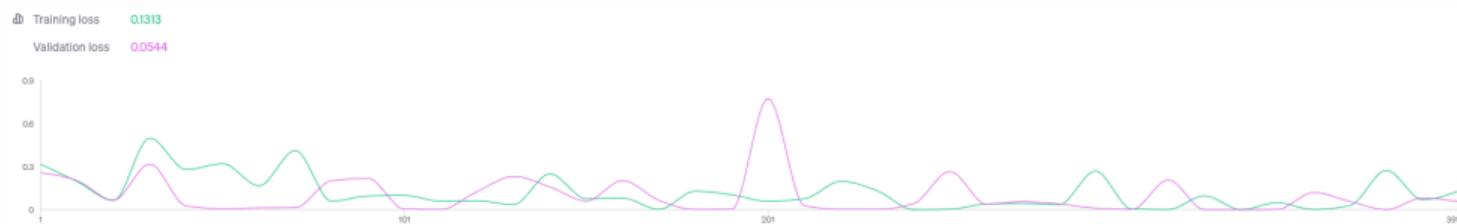
Funktioniert das?

- Titel werden z.T. als Regest erkannt (lässt sich manuell abfangen)
- für einige wenige Beispiel Accuracy > 90%
- ...

Beispiel: GitHub-Repo

Versuch der Verwendung eines offenen Modells für diese Aufgabe:  
Fine-tuning Mistral & Anwendungsprobleme

- zeilen- statt seitenweise Korrektur wie zuvor
- auf Zeilen ausgerichteter Prompt
- Fine-Tuning (GPT 3.5-turbo) anhand des Ground Truth
  - 493 Zeilen Trainingsmaterial und 124 Zeilen zur Validierung (80 / 20 Split)



- CER: 3.14 → 1.4

Beispiel: GitHub-Repo

	CER	WER	Kosinusähnlichkeit
Vorher	3.14	0.019	0.98
Nachher	1.40	0.014	0.98

## NER

- Few-Shot-Prompting
- Verwendung von `spacy-llm`
- LLM: GPT-4

## TEI-Konvertierung

- lokaler Funktionsaufruf
- Ausgabe von TEI-XML mithilfe eines Templates
- Einbindung von NER-Ergebnissen durch Auslesen der strukturierten Ausgaben vorheriger Schritte

# Ausführung der Pipeline / Demo

The screenshot displays the SSRQ Retro Lab web interface. At the top left, there is a green checkmark icon followed by the text "SSRQ Retro Lab". Below this, there is a dropdown menu labeled "Select an article to process". Underneath the dropdown, there are two side-by-side panels. The left panel is titled "Named Entities found" and contains a bar chart icon. The right panel is titled "TEI XML" and contains a code editor icon. Below these panels, there is a text input field labeled "Corrected OCR Text" with the placeholder text "Corrected OCR Text". A large grey button labeled "Start LLM chain" is positioned below the text input. At the bottom of the interface, there is a "Logs" section with a text area. At the very bottom of the page, there is a small footer that reads "Use via API" with a red arrow icon and "MIT Creative Commons" with a CC icon.

<https://github.com/SSRQ-SDS-FDS/ssrq-retro-lab?tab=readme-ov-file#demo>

## **Ausblick**

---

- Vollständige Automatisierung noch nicht möglich (Human in the Loop notwendig?)
- Wie lässt die Leistungsfähigkeit von LLMs in der Pipeline zuverlässig evaluieren?
- Hürden und Herausforderungen
  - Nachhaltigkeit
    - Reproduzierbarkeit
    - Zugang zu Modellen
  - Ausführungszeit (und -kosten)
  - Wartbarkeit
  - Umgang mit Sprachvielfalt in den Quellen