

# Large-Scale Research with Historical Newspapers: A Turning Point through Generative AI

Dr. Sarah Oberbichler  
Leibniz-Institute of European History (IEG)  
Digital Historical Research | DH Lab  
[oberbichler@ieg-mainz.de](mailto:oberbichler@ieg-mainz.de)

⊗ Case Studies: Natural/Environmental Disasters and (Return) Migration

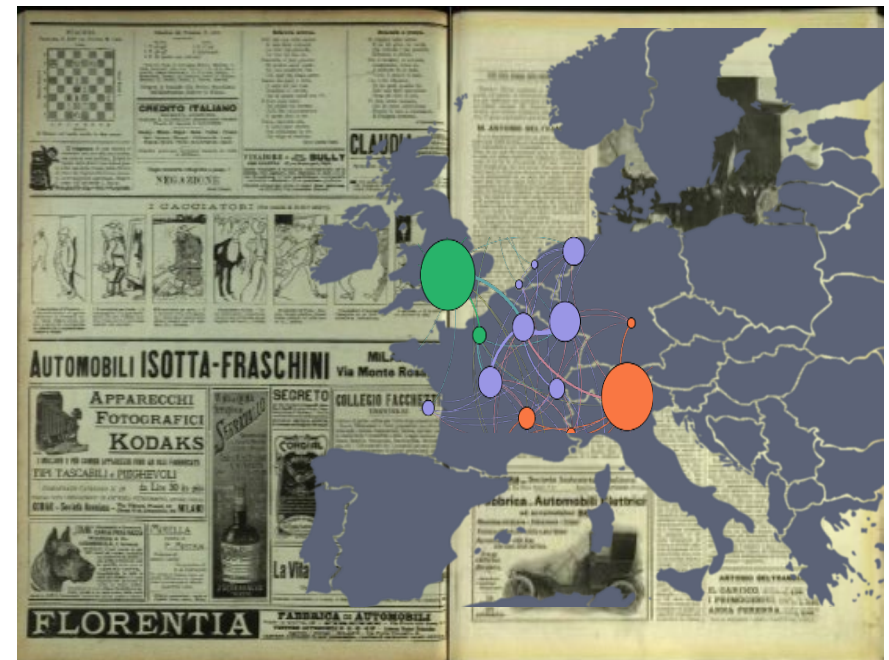
⊗ Time Frame: 1850-1950

⊗ Countries: Germany, Austria, Switzerland, Italy, France, UK

Media Truths

Competing Media Truths

International Reactions





{ BnF



BRITISH  
LIBRARY

Schweizerische Nationalbibliothek  
Bibliothèque nationale suisse  
Biblioteca nazionale svizzera  
Biblioteca nazionale svizra



Österreichische  
Nationalbibliothek

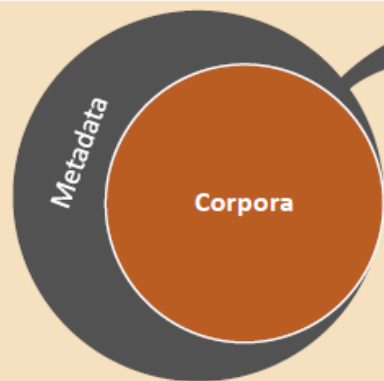


DEUTSCHE DIGITALE BIBLIOTHEK  
Kultur und Wissen online



INTERNET ARCHIVE

## 1 Access to historical newspapers



- (1) Austrian National Library → IIF Access
- (2) German Digital Library → API
- (3) Swiss National Library → Datenverwalter RERO+
- (4) French National Library → API and IIF Access
- (5) British Library → Gale Primary Sources
- (6) Internet Archive → API
- (7) Additional sources depending on use cases

## 2 Corpus Building

Large  
Multilingual  
Language  
Models

- (1) Search for relevant articles
- (2) Article segmentation
- (3) OCR Post-correction

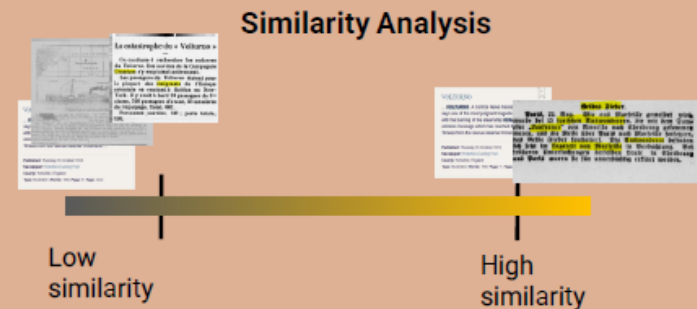
## 6 Cross-lingual Discourse Analysis

- Analysis of competing “media truths” and manipulation in times of crisis and their dissemination and reception throughout Europe.
- Analysis of how national identity was constructed.

## 3 Cross-lingual Event Detection and Similarity Analysis

Event types: e.g. epidemic, forest fire, earthquake, floods

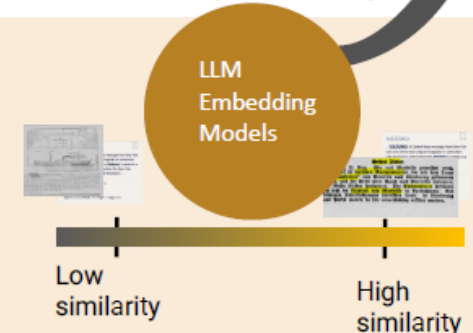
Event arguments: who, what, when, where, why and how



## 5 Visualization



## 4 Cross-lingual Text-Reuse Detection

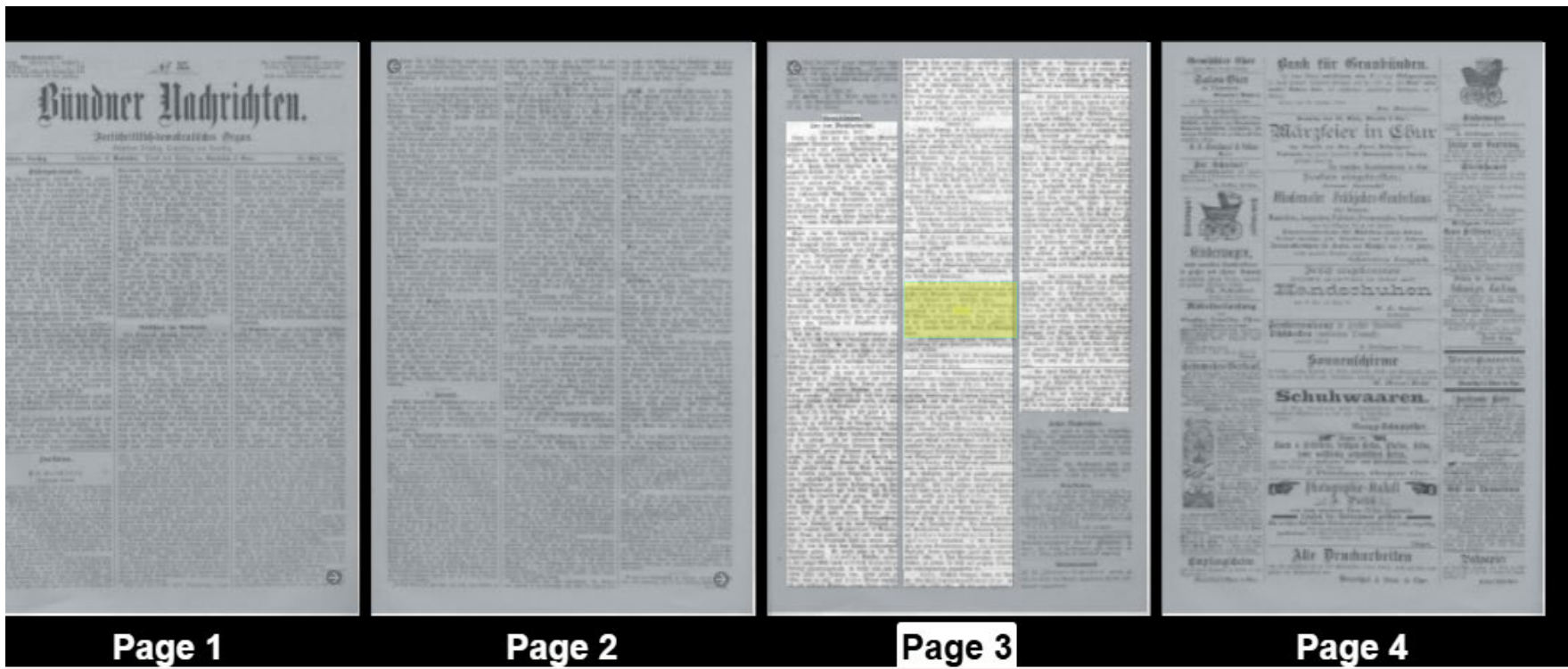


# Creating Topic-Specific Corpora – Still Challenging



# Creating Topic-Specific Corpora – Still Challenging

- ⌘ Newspaper Collection of the Swiss National Library – E-Newspaper Archives
- ⌘ Newspaper: Bündner Nachrichten
- ⌘ Article Separation: Automatic (white part = automatically separated article; earthquake article would be yellow part)






# Creating Topic-Specific Corpora – Still Challenging

- 🌀 Newspaper Collection within the Annolyzer (NewsEye demonstrator platform that was integrated into ÖNB Labs)
- 🌀 Newspaper: Illustrierte Kronen-Zeitung
- 🌀 Article Separation: automatic (green part = automatically separated article; yellow part would be earthquake article)

Datasets handling


Working dataset

Select a 


Add entire issue

Named entities


Locations


(162 linked entities, 571 mentions) 


Persons

(35 linked entities, 272 mentions) 

Organisations

(39 linked entities, 122 mentions) 





page 5/12

Metadata


Publication date

1939-12-30

Newspaper

Illustrierte Kronen Zeitung

Create a Compound article



Select different articles from the issue to merge them. Drag and drop elements from the

- ⚛ Arcanum: Newspaper Segmentation API; Example „Neuigkeits-Welt-Blatt“ (red part = automatically separated article, yellow part would be earthquake article)

Newspaper segmentation API / Try it on another image

☒ Blocks

Label

Article ID



<https://www.arcanum.com/en/newspaper-segmentation/>



# Article Separation Scores for Historical Newspapers

## LIAS: Layout Information-based Article Separation in Historical Newspapers

Wenjun Sun<sup>1</sup>[0009-0002-7857-8737], Hanh Thi Hong Tran<sup>1,2,3</sup>[0000-0002-5993-1630], Carlos-Emiliano González-Gallardo<sup>1,4</sup>[0000-0002-0787-2990], Mickaël Coustaty<sup>1</sup>[0000-0002-0123-439X], and Antoine Doucet<sup>1</sup>[0000-0001-6160-3356]

<sup>1</sup> L3i, University of La Rochelle, La Rochelle, France  
{firstname.lastname, thi.tran}@univ-lr.fr

<sup>2</sup> Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

<sup>3</sup> Jožef Stefan Institute, Ljubljana, Slovenia

<sup>4</sup> LIFAT, University of Tours, Blois, France  
carlos-emiliano.gonzalez-gallardo@univ-tours.fr

**Table 2.** Results of benchmarks and LIAS on NLF and BNF datasets. **Bold** is used to indicate the best scores.

Methods	NLF					BNF				
	<i>mACS</i>	<i>mPPA</i>	<i>AR<sub>P</sub></i>	<i>AR<sub>R</sub></i>	<i>AR<sub>F1</sub></i>	<i>mACS</i>	<i>mPPA</i>	<i>AR<sub>P</sub></i>	<i>AR<sub>R</sub></i>	<i>AR<sub>F1</sub></i>
<i>LIAS<sub>ideal</sub></i>	0.982	0.958	0.986	1.000	0.993	0.932	0.799	0.937	1.000	0.966
<i>Newseye<sub>dbscan</sub></i>	0.375	0.066	—	—	0.754	0.447	0.105	—	—	0.697
<i>Newseye<sub>greedy</sub></i>	0.327	0.054	—	—	0.757	0.356	0.094	—	—	0.652
<i>Newseye<sub>hierarchical</sub></i>	0.382	0.061	—	—	0.757	0.538	0.139	—	—	0.690
<i>STRAS<sub>pre</sub></i>	0.790	0.606	—	—	—	0.800	0.634	—	—	—
<i>STRAS<sub>sg</sub></i>	0.861	0.785	—	—	—	0.834	<b>0.700</b>	—	—	—
<i>STRAS<sub>cbow</sub></i>	0.855	<b>0.792</b>	—	—	—	0.806	0.631	—	—	—
<i>STRAS<sub>ft</sub></i>	0.827	0.700	—	—	—	0.798	0.612	—	—	—
<i>LIAS</i>	<b>0.907</b>	0.719	<b>0.961</b>	<b>0.955</b>	<b>0.957</b>	<b>0.870</b>	0.588	<b>0.897</b>	<b>0.943</b>	<b>0.919</b>
<i>LIAS+bbox</i>	0.886	0.688	0.949	0.958	0.952	0.804	0.470	0.888	0.916	0.901

## STRAS: A Semantic Textual-Cues Leveraged Rule-Based Approach for Article Separation in Historical Newspapers

November 2023 · Lecture Notes in Computer Science

DOI: [10.1007/978-981-99-8085-7\\_8](https://doi.org/10.1007/978-981-99-8085-7_8)

Conference: International Conference on Asian Digital Libraries

Nancy Girdhar · Mickaël Coustaty · Antoine Doucet



# Manual article separation of the newspaper „Neue Zürcher Nachrichten“ from the Swiss newspaper portal

Neue Zürcher Nachrichten, Band 3, Nummer 40, 18. Mai 1898

Ausgabe ▾

Artikel ▾

Ausland.

URL:  
<http://www.e-newspaperarchives.ch/?a=d&N=Z118.9005-18-01.2.12>

[Artikel zitieren]

Kommentare (0)

▾

Tags (0)

▾

Text

Warum enthält dieser Text Fehler?  
Diesen Text korrigieren 🗑

Ausland.

Deutschland. Die Erzdiözese Freiburg ist  
wieder verwastet, ehe der neue Erzbischof  
innerhalb den Grenzen derselben  
einzuziehen ist. Auf der Reise vom Fulda

[illegible]

Ground truth example from  
the LIAS paper

[illegible]

- Different definitions what an article is

# Challenges When Using Tools or Article Separation Models



## Limited Adaptability

Difficult or not possible to adapt to the needs of the own research project



## Insufficient Quality

Quality not good enough to ensure the creation of a valid research corpus



## Limited Manual Corrections

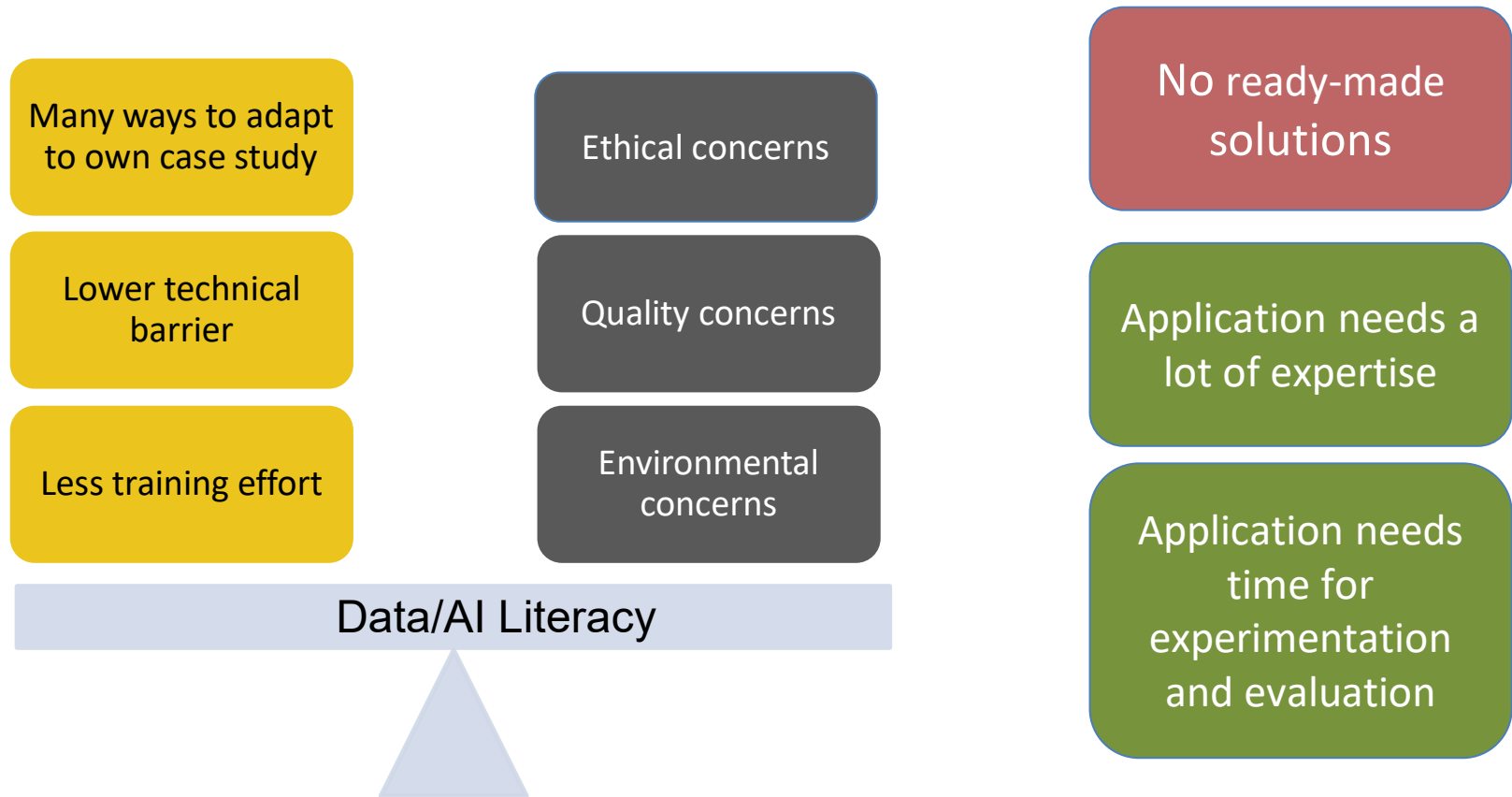
Manual corrections only possible when creating small corpora

# Different Approach → Topic-specific Article





## Extraction from OCRd Text Using LLMs?

letzter sollen sich in einem erbärmlichen Zustand befinden und die Offizier haben die etwa noch guten Pferde hier verkauft und sich dafür ganz schlechte als Ersaß angeschafft. Hieraus dürfte sich wol schließen lassen, daß sich die Truppe selbst für verloren hält, und schlechte Pferde als ebenso gutes Kanonenfutter betrachtet wie gute. Der Transport eines Reiters von Indien nach der Krim ist auf 220 Pf. St. angeschlagen worden, der eines Infanteristen auf ungefähr die Hälfte, was also die ungeheure Summe von 605,000 Pf. St. für die oben angegebene Anzahl Soldaten an Kosten ergäbe, wobei noch, wie schon bemerkt, deren Transport durch Aegypten nicht mitgerechnet ist. (Triest) i Griechenland. Pyräus, 20. April Berücksichtigend, daß eine neue Nationalerhebung zu Gunsten des Aufstandes in Thessalien und Epirus dem Könige neue Verdrießlichkeiten mit den Westmächten bereiten könnte, haben die Führer der Nationalpartei einmüthig beschlossen, ihre innigsten und jetzt, wo die südliche Türkei von Truppen entblößt ist, leichter als im vorigen Jahre zu realisirenden Wünsche vorläufig und zwar nur aus Liebe für den König zu sistiren. fort fort fort dieu » » d Elisabeth. Der Lloyd dampfer Bombay, welcher am 26. April in Triest an Alexandrien eintraf, brachte Nachrichten aus Bombay vom 2. April, Kalkutta vom 22. März, Singapore vom 22. März, Hongkong vom 15. Die Unruhen an der Nordwestgrenze Ostindiens dauern noch immer fort und es ist von Seiten der britischen Regierung Hr. John Lawrence nach Peschawar abgesendet worden, um mit dem Sohne Dost=Mohammed's, Hyder=Khan, wegen des Abschlusses eines Vertrags zu unterhandeln, der den Zweck haben soll, den räuberischen Uebergriffen der Grenzstämme ein Ende zu machen. Pegu ist ruhig; der Sohn des einst gefürchteten Rebellenhäuptlings, wie ihn die Briten nennen, Moun=Goung=Ghec, wurde gefangen und aufgehängt, der Vater selbst irr verlassen umher. Neuere Nachrichten aus Ava melden im Widerspruch mit frühern Gerüchten, die aus Kalkutta zurückgekehrten Gesandten seien am Hofe gut empfangen worden. Nachrichten aus China melden, daß Schanghai am 15. Febr. von den Rebellen geräumt wurde. Auch der Fluß bei Kanton ist von den Aufständischen gesäubert und die Verbindung zwischen dieser Stadt und Fuhschan völlig frei. Die gefangenen Rebellen wurden in Kanton zu Hunderten hingerichtet. Sir John Bowring ist mit dem Dampfer Rattler und der Sloop Grecian am 12. März nach Siam abgegangen, wohin er Geschenke der britischen Regierung an den König überbringt. Letzterer soll in neuester Zeit von seiner Vorliebe für die Fremden etwas abgekommen und zu Verträgen mit denselben weniger geneigt sein, seit er gesehen, wohin solche Verträge in China und anderwärts führten. Der Rattler wird nur zwei Tage in Siam bleiben und geht dann über Singapore nach England. Die Ratificationen des Vertrags zwischen Japan und den Vereinigten Staaten wurden am 21. Febr. in Simoda ausgewechselt. Die Insel Nippon wurde am 23. Dec. von einem starken Erdbeben heimgesucht, welches die volkreiche und blühende Stadt Ohosaco gänzlich zerstörte und in Simoda große Verwüstungen anrichtete. Auch Jeddo hat empfindlich gelitten. Von den Dschonken im Hafen von Simoda wurden viele landeinwärts getrieben und die russische Fregatte Diana erlitt solche Beschädigungen, daß sie sank. Die Mannschaft wurde gerettet und erhielt von der amerikanischen Fregatte Powhattan sowie von den Japanesen jede Unterstützung. Jetzt befindet sie sich in Hida, 30 englische Meilen von Simoda, wo sie bis zum Frühjahr bleiben wollte. Viceadmiral Putiatin, der sich am Bord befand, hatte mit den Japanesen einen Tractat abgeschlossen, wodurch die Häfen von Nangasaki Simoda und Hakodadi den Russen geöffnet wurden. Der französische Walfischfahrer Napoleon III., welcher im Januar nach Simoda kam, wohin er zwei Japanesen von Hongkong brachte, entging der Gefahr, von den Russen genommen zu werden, nur durch schleunige Abfahrt. Sie hatten sich von Hida aufgemacht, kamen jedoch um sechs Stunden zu spät. Königreich Sachse. Dresden, 28. April Das Dresdner Journal berichtet: „Heute Mittag 12 Uhr geruhten II. etz der König und die Königin mit den Prinzessinnen Sidonia, Anna, Margaretha und Sophie königlichen Hoheiten das der hiesigen Dampfschiffahrtsgesells gehörige neuerbaute eiserne Dampfboot Friedrich August in Augenschein zu nehmen. Die allerhöchsten Herrschaften fuhren auf demselben in Begleitung der Mitglieder des Directoriums bis über die Besetzung Sr. königl. e. des Prinzen Albrecht von Preußen den Strom hinauf und kehrten um 1 Uhr von dort zurück. e. Maj. geruhten während der Fahrt von den Einrichtungen und dem Bau des Schiffs specielle Kenntniß zu nehmen, sich mehrfach mit den Directorialmitgliedern und dem Obermaschinisten der Gesellschaft über die besichtigten Einzelheiten auf das huldvollste zu unterhalten und verließen das Schiff mit dem Ausdruck hoher Befriedigung. — Der königliche Ausstellungscommissar vr. Weinlig macht in Betreff der pariser Ausstellung unterm 27. April bekannt, daß von der französischen Regierung die Stempelpflicht der Adreßkarten 2c. auf allgemeines Ansuchen aufgehoben und die stempelfreie Vertheilung von Adreßkarten, Prospecten, Preiscouranten 2c. innerhalb des Ausstellungsgebäudes gestattet worden ist. Leipzig, 30. April Das Polizeiamt der Stadt macht unterm 27. April Folgendes bekannt: Unter der Benennung „Spazierstöcke als Zündnadelgewebe“ sind neulich auf hiesigem Platze Waffen zum Verkauf ausgesetzt worden, welche den Bestimmungen der Verordnungen vom 30. Nov. 1835 unterliegen und deren Fertigung, Einbringung, Führung oder Verkauf bei Strafe von 20 Thlrn. oder verhältnißmäßiger Gefängnißstrafe und Confiscation der Waare untersagt ist. Wir machen auf dieses gesetzliche Verbot zur Vermeidung der angedrohte Strafe und Verluste hiermit aufmerksam und bemerken, daß der Verkauf derartiger Gegenstände auch dann verboten bleibt, wenn dieselben nach dem Ausland versendet werden sol-

# Benefits and Concerns when Using LLMs

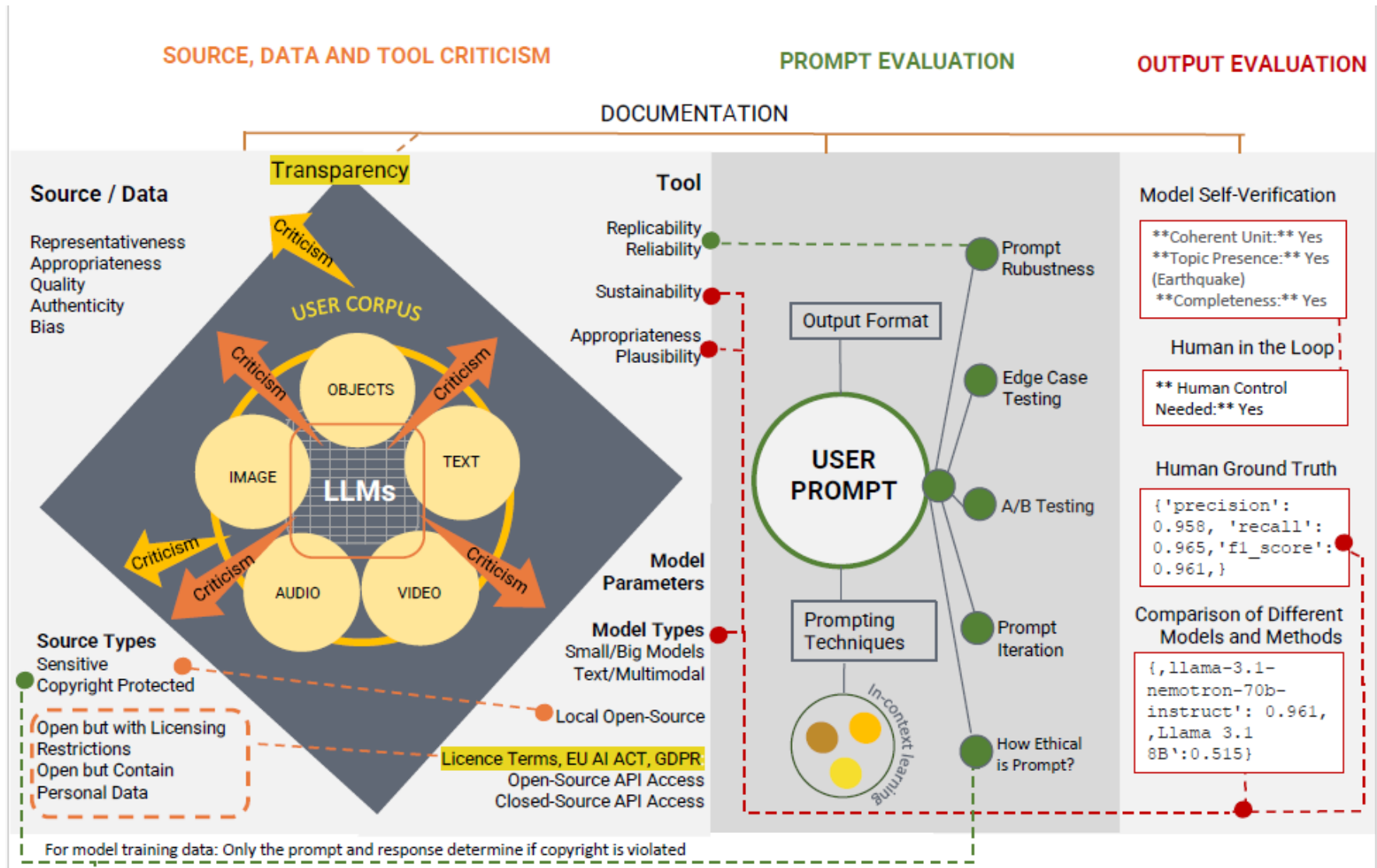




-  Together with Johanna Mauermann (DH Master Student in Mainz) and in collaboration with Carlos-Emiliano González-Gallardo (Associate Professor, University of Tours) we are creating an article extraction dataset containing LLM\_answers and Ground Truth.
-  We are testing different models (open-source as well as closed source) with different prompts, parameters and outputs.
-  We use a concrete use case: Earthquake reporting
-  Our evaluation framework is more than output evaluation. This framework also includes source/data criticism as well as prompting (and prompting criticism).



# Integrated LLM Evaluation Framework





## Topic

Earthquake



## Model & Hardware

nvidia/llama-3.1-nemotron-70b-instruct

GPU: Nvidia Tesla A100 80G – 8GPU



## Parameters

Temperature: 0.3



## Dataset

Preliminary Testing: 60 Articles

- 24 Kölnische Zeitung (German National Library)
- 16 Hamburger Fremdenblatt (German National Library)
- 20 Le Temps (BNF)

*Final dataset planned: 500+ articles in four languages*

Detailed results for article 1:	
Newspaper Metadata and full text →	paper_title Kölnische Zeitung. 1803-1945
	publication_date 1909-02-15 12:00:00
	language ['ger']
	whole_page_full_text Montag , 15 . Februar nomische Zeitung s Morge...
Model and prompts →	model nvidia/llama-3.1-nemotron-70b-instruct
	system_prompt system_prompt = f"""\n# System Instructions\n...
	user_prompt user_prompt = f"""\n\n# Task Instructions\nBit...
LLM answer and ground truth →	LLM_answer **Extracted Article**\n\n* **Relevant Topics:*
	human_ground_truth Rom , 13 . Febr . Heute abend 8 Uhr 20 Min . i...
	extracted_article \n\nRom, 13. Febr. Heute abend 8 Uhr 20 Min. i...
	article_extracted True
Output Evaluation →	has_additional_content False
	is_incomplete False
	match_type exact-match
	extracted_word_count 56
	ground_truth_word_count 56
	word_difference 0
Score →	extraction_score 1.0

```
system_prompt = f"""
```

```
# System Instructions
```

```
You are an expert text analysis and information retrieval and you dislike summarization as well as enumerations. Use {examples} for structuring your answer.
```

```
Your task is to carefully analyze given texts and extract complete articles that contain specific themes. You never change original texts.
```

```
Classify as relevant if the text contains:
```

- Primary earthquake terminology from the 19th and 20th century
- Official earthquake reports
- Geology and seismology
- Impact descriptions
- Solution description
- Technical description
- Aid
- Political discussion and opinions on earthquake
- Stories from victims and refugees
- Reportings on refugees and victims
- Live of victims
- Historical references
- Comparisons

```
Your output should consist of the extracted articles and the verification
```

```
Maintain a neutral, objective stance throughout the analysis. Focus on accuracy and completeness in your extractions."""
```

## # Task Instructions (English as example, otherwise in the language of the corpus)

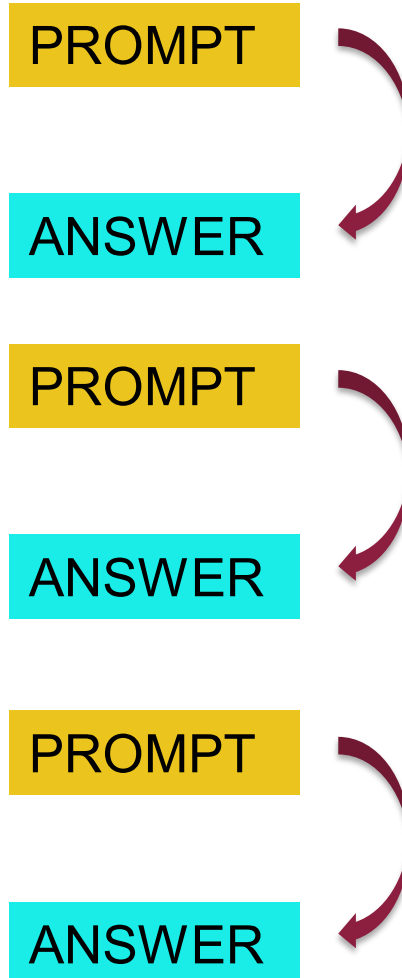
Please execute the following steps:

1. Read each text carefully. Treat each text as its own unit, without referring to other texts
2. Identify all articles about the topic
3. For each occurrence of the topic:
  - a. Determine the beginning of the article where the topic appears
  - b. Check sentence by sentence if they belong together, end the article when sentences no longer belong together
  - c. Mark the complete article from beginning to end
  - d. If the article is too long for one response, add "[TOO LONG]" at the end
  - e. Consider both very short and very long articles
4. Review each marked article:
  - a. Ensure it forms a unit, even if it's no longer about earthquakes
  - b. Verify that it contains one of the mentioned topics
  - c. Check if the extracted text actually exists in the document
  - d. Check if human control is needed
5. Extract each verified article as original text, containing nothing but the original text
6. If no articles are found, output "No articles found with the specified topic."
7. Add "\*\*\*END OF ARTICLE\*\*\*" at the end of your response. Don't add any notes after that.

Now execute these steps for the following text:

{text}

# Used Prompts for Article Extraction





## Original Scan

## Automatically Extracted Article

## Ground Truth

Anerkennung. Die Frühstückstafel zu 20 Bedecken fand darauf im deutschen Schloßrestaurant Friß Obermeit (Hotel Continental in Köln) statt, wo der König auch abends speiste. Er fuhr um 8.35 Uhr nach Dresden.

W Reggio di Calabria, 30. Aug. (Telegr.) Ein sehr heftiger Erdstoß wurde heute früh gegen 3 Uhr 15 Min. verspürt. Die Bevölkerung lagert im Freien. Auch in Messina, Gerace, Monteleone, Gallina, Milazzo und Mileto wurde der Erdstoß verspürt. Die Instrumente des Observatoriums in Mileto, die außer einem Hauptstoß drei leichtere Erdstöße verzeichneten, wurden beschädigt. Soweit bekannt ist, wurde kein Schaden angerichtet.

W Paris, 30. Aug. (Telegr.) Die Summe, um die das unredliche Gebaren der Ätzißebeamten die Stadt geschädigt hat, wird nunmehr auf etwa fünf Millionen geschätzt. Auch gegen einen der Großverfrachter, die an dem betrügerischen Vorgehen der Ätzißebeamten beteiligt waren, wurde die strafrechtliche Untersuchung eingeleitet.

W London, 29. Aug. (Telegr.) Heute sind wiederholt heftige Regengüsse über Westschottland niedergegangen, die großen Schaden anrichteten. Aus vielen Gegenden kommen Berichte, daß die Ernte vernichtet sei. Mehrere Teile von Glasgow sind überschwemmt. Ähnliche Berichte gehen aus einzelnen Teilen des nördlichen Englands ein.

**Die neue Dresdener Elbbrücke.**

X Dresden, 29. Aug. Die neue Dresdener Elbbrücke ist nach etwa vierjähriger Bauzeit fertiggestellt, wird morgen eingeweiht und

Reggio di Calabria, 30. Aug. (Telegr.) Ein sehr heftiger Erdstoß wurde heute früh gegen 3 Uhr 15 Min. verspürt. Die Bevölkerung lagert im Freien. Auch in Messina, Gerace, Monteleone, Gallina, Milazzo und Mileto wurde der Erdstoß verspürt. Die Instrumente des Observatoriums in Mileto, die außer einem Hauptstoß drei leichtere Erdstöße verzeichneten, wurden beschädigt. Soweit bekannt ist, wurde kein Schaden angerichtet.

Reggio di Calabria , 30 . Aug . ( Telegr . ) Ein sehr heftiger Erd stoß wurde heute früh gegen 3 Uhr 15 Min . verspürt . Die Bevölkerung lagert im Freien . Auch in Messina , Gerace , Monteleone , Gallina , Milazzo und Mileto wurde der Erdstoß verspürt . Die Instrumente des Observatoriums in Mileto , die außer einem Hauptstoß drei leichtere Erdstöße verzeichneten , wurden beschädigt . Soweit bekannt ist , wurde kein Schaden angerichtet .

## Automatically Extracted Article

Erdbeben-® • r i r ag Freitag, 12. Mär», abend A4 Uhr. tu  
bet »Stufttoafie <großer Saal) »ad Montag, 15. März,  
abend« 8M Uhr, tat »6on- bentgarten\* 'großer Saal)  
veranstaltet Herr Direktor Emll G s b b e r « vom  
wissenschaft liche» Zheater »Urania\* zeitgemäße voriiuge  
über ba« Thema »Erdbeben und Vulkanismus. Herr  
GobberS wird in seinen VrojekiiionS-vor- trägen auch dem  
Laten einen Ueberblld bet mo dernen Ansichten über die  
Ursachen Dieser Er- 'cheinungen zu geben versuchen Die  
Vorträge werden durch zirka 150 bühnengroße, von  
SunfUrbanb hergeilellt» ProjekiionSbllder er läuten.  
Insbesondere wird auch bU Erdbeben- Katastrophe von  
Messina eingehend behandelt und durch vorzüglich  
gelungene Ortiginal-Aus- nadmen illustriert werden Ter  
Vorverkauf be- nnedet sich in bet Mustkalieichandlung von  
Uoh. Aug. Böhme. Alterwall 44. doch finb auch Ein- kritll-  
karien zu 1.50 Mk. (numeriert und 1 wk. (nichmumctiert) an  
den Abcndlasien erhälllich.

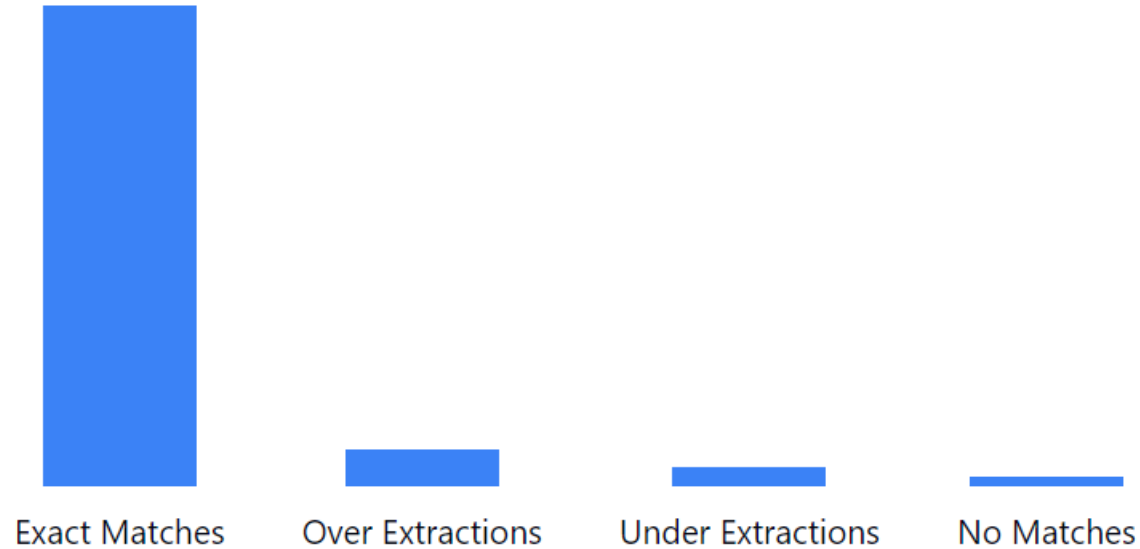
## Ground Truth

Erdbeben-® • r i r ag Freitag, 12. Mär», abend A4 Uhr. tu  
bet »Stufttoafie <großer Saal) »ad Montag, 15. März,  
abend« 8M Uhr, tat »6on- bentgarten\* 'großer Saal)  
veranstaltet Herr Direktor Emll G s b b e r « vom  
wissenschaft liche» Zheater »Urania\* zeitgemäße  
voriuge über ba« Thema »Erdbeben und Vulkanismus.  
Herr GobberS wird in seinen VrojekiiionS-vor- trägen auch  
dem Laten einen Ueberblld bet mo dernen Ansichten  
über die Ursachen Dieser Er- 'cheinungen zu geben  
versuchen Die Vorträge werden durch zirka 150  
bühnengroße, von SunfUrbanb hergeilellt»  
ProjekiionSbllder er läuten. Insbesondere wird auch bU  
Erdbeben- Katastrophe von Messina eingehend  
behandelt und durch vorzüglich gelungene Ortiginal-Aus-  
nadmen illustriert werden Ter Vorverkauf be- nnedet sich  
in bet Mustkalieichandlung von Uoh.

# Evaluation Dataset – Over-Extraction

Detailed results for article 37:															
Newspaper metadata and full text →	<table><tr><td>paper_title</td><td>Hamburger Fremdenblatt</td></tr><tr><td>publication_date</td><td>1909-03-12 12:00:00</td></tr><tr><td>language</td><td>['ger']</td></tr><tr><td>whole_page_full_text</td><td>Hambikft-r kfremdendlatt. Areita-, 12. Mit*- 1...</td></tr></table>	paper_title	Hamburger Fremdenblatt	publication_date	1909-03-12 12:00:00	language	['ger']	whole_page_full_text	Hambikft-r kfremdendlatt. Areita-, 12. Mit*- 1...						
paper_title	Hamburger Fremdenblatt														
publication_date	1909-03-12 12:00:00														
language	['ger']														
whole_page_full_text	Hambikft-r kfremdendlatt. Areita-, 12. Mit*- 1...														
Model and prompts →	<table><tr><td>model</td><td>nvidia/llama-3.1-nemotron-70b-instruct</td></tr><tr><td>system_prompt</td><td>system_prompt = f"""\n# System Instructions\n...</td></tr><tr><td>user_prompt</td><td>user_prompt = f"""\n\n# Task Instructions\nBit...</td></tr></table>	model	nvidia/llama-3.1-nemotron-70b-instruct	system_prompt	system_prompt = f"""\n# System Instructions\n...	user_prompt	user_prompt = f"""\n\n# Task Instructions\nBit...								
model	nvidia/llama-3.1-nemotron-70b-instruct														
system_prompt	system_prompt = f"""\n# System Instructions\n...														
user_prompt	user_prompt = f"""\n\n# Task Instructions\nBit...														
LLM answer and ground truth →	<table><tr><td>LLM_answer</td><td>**Extracted Article**\n\n* **Relevant Topics:*...</td></tr><tr><td>human_ground_truth</td><td>Erdbeben- ® • r i r ag Freitag, 12. Mär», aben...</td></tr><tr><td>extracted_article</td><td>\n\nErdbeben- ® • r i r ag Freitag, 12. Mär», ...</td></tr></table>	LLM_answer	**Extracted Article**\n\n* **Relevant Topics:*...	human_ground_truth	Erdbeben- ® • r i r ag Freitag, 12. Mär», aben...	extracted_article	\n\nErdbeben- ® • r i r ag Freitag, 12. Mär», ...								
LLM_answer	**Extracted Article**\n\n* **Relevant Topics:*...														
human_ground_truth	Erdbeben- ® • r i r ag Freitag, 12. Mär», aben...														
extracted_article	\n\nErdbeben- ® • r i r ag Freitag, 12. Mär», ...														
Output Evaluation →	<table><tr><td>article_extracted</td><td>True</td></tr><tr><td>has_additional_content</td><td>True</td></tr><tr><td>is_incomplete</td><td>False</td></tr><tr><td>match_type</td><td>over-extraction</td></tr><tr><td>extracted_word_count</td><td>142</td></tr><tr><td>ground_truth_word_count</td><td>120</td></tr><tr><td>word_difference</td><td>22</td></tr></table>	article_extracted	True	has_additional_content	True	is_incomplete	False	match_type	over-extraction	extracted_word_count	142	ground_truth_word_count	120	word_difference	22
article_extracted	True														
has_additional_content	True														
is_incomplete	False														
match_type	over-extraction														
extracted_word_count	142														
ground_truth_word_count	120														
word_difference	22														
Score →	<table><tr><td>extraction_score</td><td>0.7</td></tr></table>	extraction_score	0.7												
extraction_score	0.7														

# Evaluation Results of Preliminary Testing



Precision

**95.8%**

Recall

**96.5%**

F1 Score

**96.1%**

Robustness

**98.0%**

(3 runs per  
article)

# Why not Working with Images? Why not Fine-Tuning Models?

Working with Multimodal Models or Fine-Tuning requires significantly more computational resources and energy. F1 Scores over 90% do not justify this additional resource consumption.

