



Benchmarking großer Sprachmodelle im Rahmen eines Explainability Frameworks

Sarah Oberbichler/Johanna Mauermann

AGKI Webinar 04.07.2024

Warum Benchmarking in den Geisteswissenschaften?



**Transparenz und
methodische Rigorosität**



Qualitätssicherung

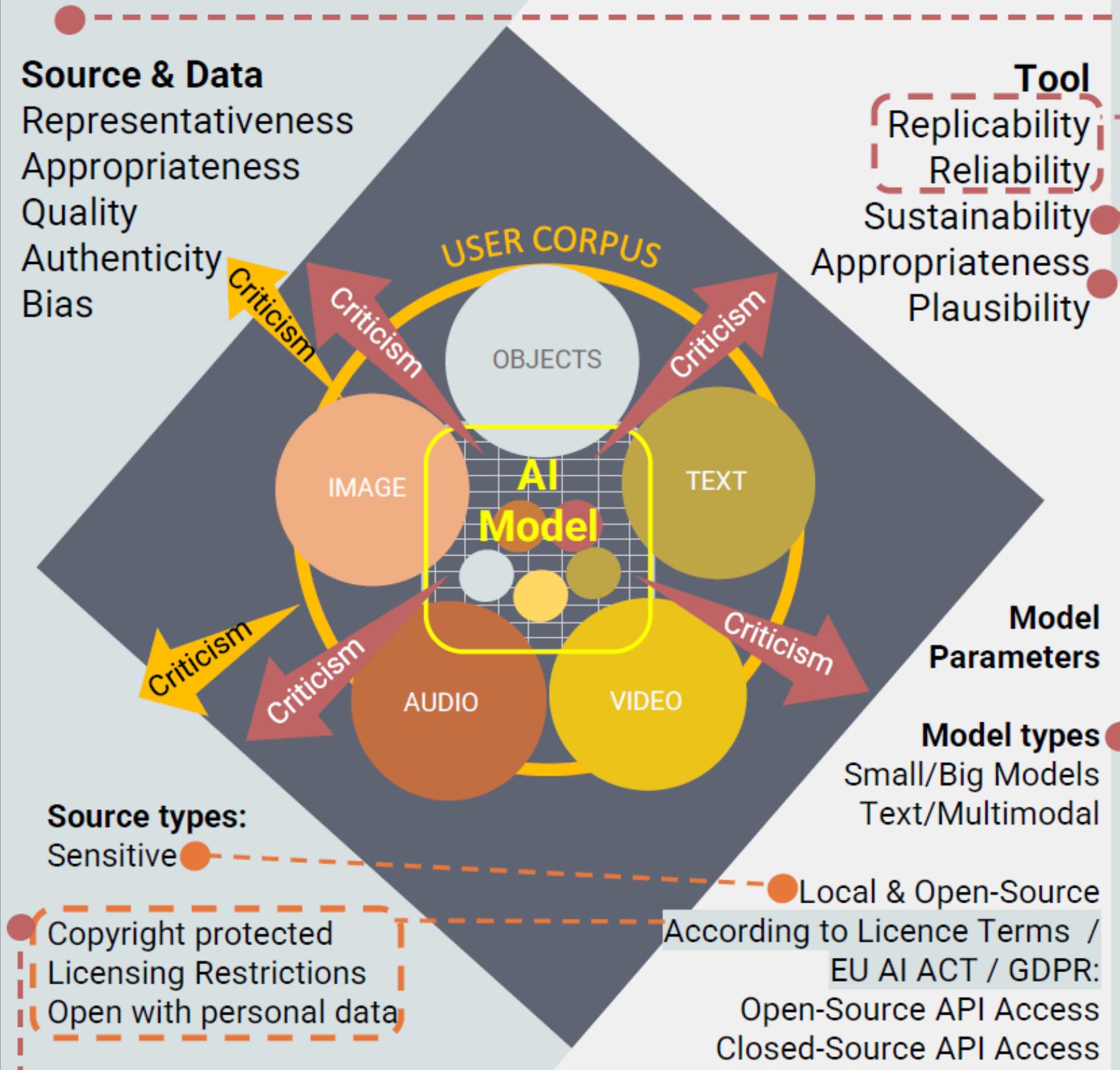


**Integraler Teil
historischer Quellen-
und Toolkritik**

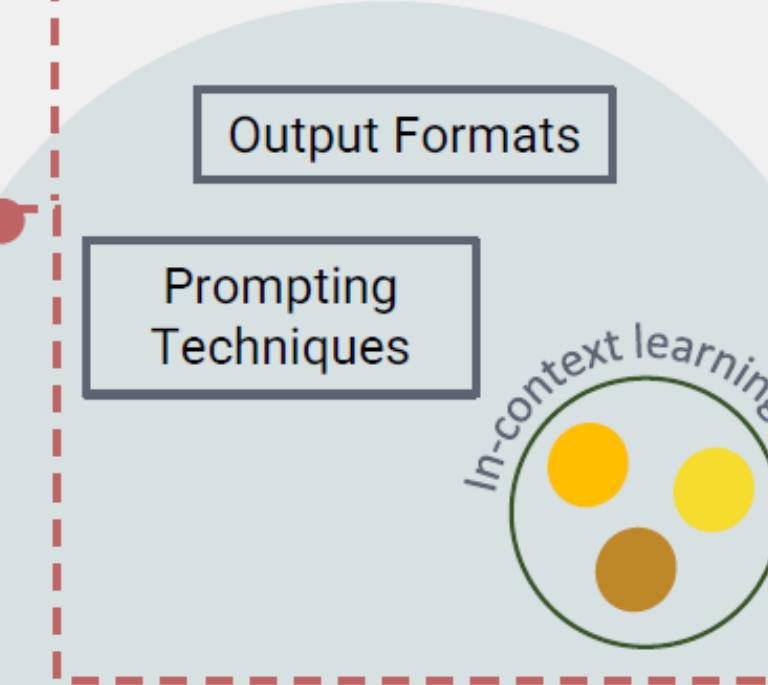
SOURCE, DATA AND TOOL CRITICISM

PROMPT AND OUTPUT EVALUATION

DOCUMENTATION & TRANSPARENCY



USER PROMPT



Model Self-Verification

**** Coherent Unit:** Yes
** Topic Presence:** Yes
(Earthquake)
** Completeness:** Yes**

Human in the Loop

**** Human Control Needed:** Yes**

Human Ground Truth

```
{'precision':  
0.958, 'recall':  
0.965, 'f1_score':  
0.961,}
```

Comparison of Models and Methods

```
{'llama-3.1-nemotron-70b-instruct': 0.961,  
'Llama 3.1 8B': 0.515}
```



Vorbereitungen fürs Benchmarking...

- Auswahl der Modelle, die verglichen werden sollen
- Wahl der Metriken (e.g., Precision, Recall, F1)
- Wahl der Evaluierungsmethode
- Erstellen von Ground Truth oder Wahl von bereits vorhandenen Evaluationsdatensets



Aus der Praxis: Benchmarking für Artikelextraktion aus historischen Zeitungen

Extraktion von Artikeln aus Zeitungstranskripten



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

November 2023 · Lecture Notes in Computer Science

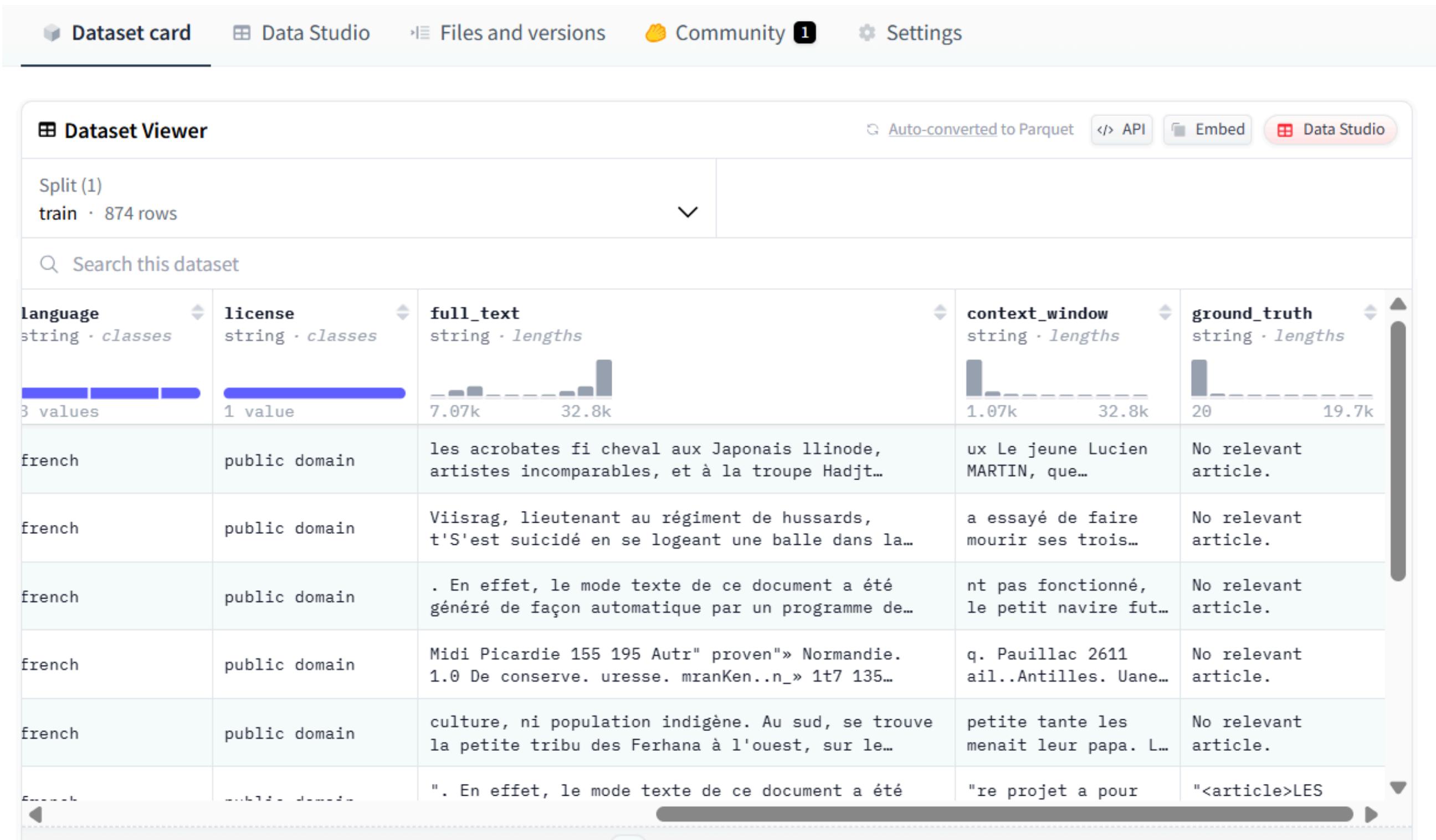
DOI: [10.1007/978-981-99-8085-7_8](https://doi.org/10.1007/978-981-99-8085-7_8)

Conference: International Conference on Asian Digital Libraries

• Nancy Girdhar · • Mickaël Coustaty · • Antoine Doucet

https://link.springer.com/chapter/10.1007/978-981-99-8085-7_8

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 wohl 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 einmütig beschlossen, ihre innigsten und jetzt, wo die südliche Türkei von Truppen entblößt ist, leichter als im vorigen Jahre zu realistrenden Wünsche vorläufig und zwar nur aus Liebe für den König zu sistiren. fort fort fort fort dieu » » d Elisabeth. Der Lloyddampfer 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 Peschawer 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, Moung=Goung=Ghec, wurde gefangen und aufgehängt, der Vater selbst irr verlassen umher Neuere Nachrichten aus Ava melden im Widerspruch mit fröhlichen 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 Ratifikationen 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 Mitag 12 Uhr geruhten II. etz der König und die Königin mit den Prinzessinnen Sidonia, Anna, Margaretha und Sophie königlichen Hoheiten das der hiesigen Dampfschiffahrtsgesellschaft gehörige eiserne Dampfboot Friedrich August in Augenschein zu nehmen. Die allerhöchsten Herrschaften fuhren auf demselben in Begleitung der Mitglieder des Directoriums bis über die Besitzung 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 besichtigen 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, Preiscuranten 2c. innerhalb des Ausstellungsgebäudes gestattet worden ist. Leipzig, 30. April Das Polizeiamt der Stadt macht unterm 27. April Folgendes bekannt: Unter der Benennung „Spazierstücke als Zündnadelgewebe“ sind neulich auf hiesigem Platze Waffen zum Verkauf ausgeboten worden, welche den Bestimmungen der Verordnungen vom 30. Nov. 1835 unterliegen und deren Fertigung, Einbringung, Führung oder Verkauf bei Strafe von 20 Thlrn. oder verbültümäßiger 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-



Dataset card Data Studio Files and versions Community 1 Settings

Dataset Viewer

Split (1)
train · 874 rows

Search this dataset

language · classes
string · classes
3 values

license · classes
string · classes
1 value

full_text · lengths
string · lengths
7.07k 32.8k

context_window · lengths
string · lengths
1.07k 32.8k

ground_truth · lengths
string · lengths
20 19.7k

language	license	full_text	context_window	ground_truth
string · classes	string · classes	string · lengths	string · lengths	string · lengths
3 values	1 value	7.07k 32.8k	1.07k 32.8k	20 19.7k
french	public domain	les acrobates fi cheval aux Japonais llinode, artistes incomparables, et à la troupe Hadjt...	ux Le jeune Lucien MARTIN, que...	No relevant article.
french	public domain	Viisrag, lieutenant au régiment de hussards, t'S'est suicidé en se logeant une balle dans la...	a essayé de faire mourir ses trois...	No relevant article.
french	public domain	. En effet, le mode texte de ce document a été généré de façon automatique par un programme de...	nt pas fonctionné, le petit navire fut...	No relevant article.
french	public domain	Midi Picardie 155 195 Autr" proven"» Normandie. 1.0 De conserve. uresse. mranKen..n_» 1t7 135...	q. Pauillac 2611 ail..Antilles. Uane...	No relevant article.
french	public domain	culture, ni population indigène. Au sud, se trouve la petite tribu des Ferhana à l'ouest, sur le...	petite tante les menait leur papa. L...	No relevant article.
		". En effet, le mode texte de ce document a été	"re projet a pour	"<article>LES

Erstellen von Ground Truth: Case Study Messina Erdbeben 1908

Wahl der LLM-Modelle für den Vergleich

01 Open-Weight Modelle

- **Llama 3 70B (Meta)**
- **Nemotron 3.1 70B (Nvidia)**
- **Qwen 2.5 72B (Alibaba)**

02 Proprietäre Modelle

- **GPT-4o (OpenAI)**
- **Claude 3.5 Sonnet (Anthropic)**
- **DeepSeek V3**

Original Scan

Anerkennung. Die Frühstückstafel zu 20 Gedekken fand darauf im deutschen Schloßrestaurant Friz 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 Mileko 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 Alzisebeamten die Stadt geschädigt hat, wird nunmehr auf etwa fünf Millionen geschätzt. Auch gegen einen der Großverfrachter, die an dem betrügerischen Vorgehen der Alzisebeamten 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.

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

Automatically Extracted Article

<article>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 Mileko 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.</article>

Ground Truth

<article>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 Mileko 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 .</article>

Automatically Extracted Article

<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 voriuge über ba« Thema »Erdbeden 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 SunfUerbanb hergeilellt» ProjekiiionSbllder er läuten. Insbesondere wird auch bU Erdbeben- Katastrophe von Messina eingehend behandelt und durch vorzüglich gelungene Ortginal-Aus- nadmen illustriert werden Ter Vorverkauf be- nndet sich in bet Mustkalieichandlung von Uoh. Aug. Böhme. Alterwall 44. doch finb auch Ein- kritli- karien zu 1.50 Mk. (numeriert und 1 wk. (nichmumctiert) an den Abcndlasiens erhälllich. </article>

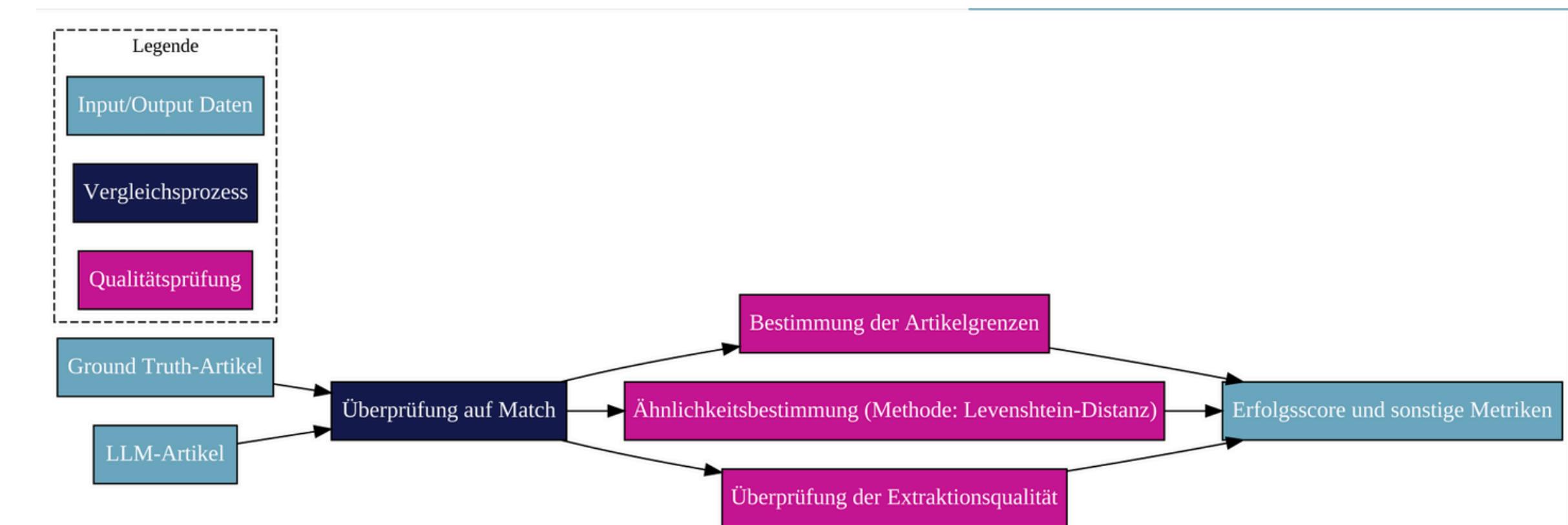
Ground Truth

<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 voriuge über ba« Thema »Erdbeden 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 SunfUerbanb hergeilellt» ProjekiiionSbllder er läuten. Insbesondere wird auch bU Erdbeben- Katastrophe von Messina eingehend behandelt und durch vorzüglich gelungene Ortginal-Aus- nadmen illustriert werden Ter Vorverkauf be- nndet sich in bet Mustkalieichandlung von Uoh.</article>

Kriterien für erfolgreiche Extraktion

- Artikel vollständig erfasst: Keine Kürzungen
- Nur relevanter Artikel erfasst: Keine Hinzufügungen
- Keine Veränderungen gegenüber Original
- Separate Erfassung mehrerer Artikel
- Wenn kein relevanter Artikel, entsprechend kennzeichnen

Workflow, um Erfolg zu quantifizieren



Metriken für die Evaluierung

Berechnung: 0-1 Skala
basierend auf Alignment-Kategorie

- False Positives/Negatives: **SSCr** = 0
- True Negatives: **SSCr** = 1
- Aligned Articles: **SSCr** = $\text{textsim}(g,a)$

Success Score (SSCr) -
Hauptmetrik

$$\text{textsim}(g,a) = \max(0, 1 - (\max(0, d^{\text{lev}} - \tau) + \alpha \cdot d^{\text{omit}} + \beta \cdot d^{\text{add}}) - \delta)$$

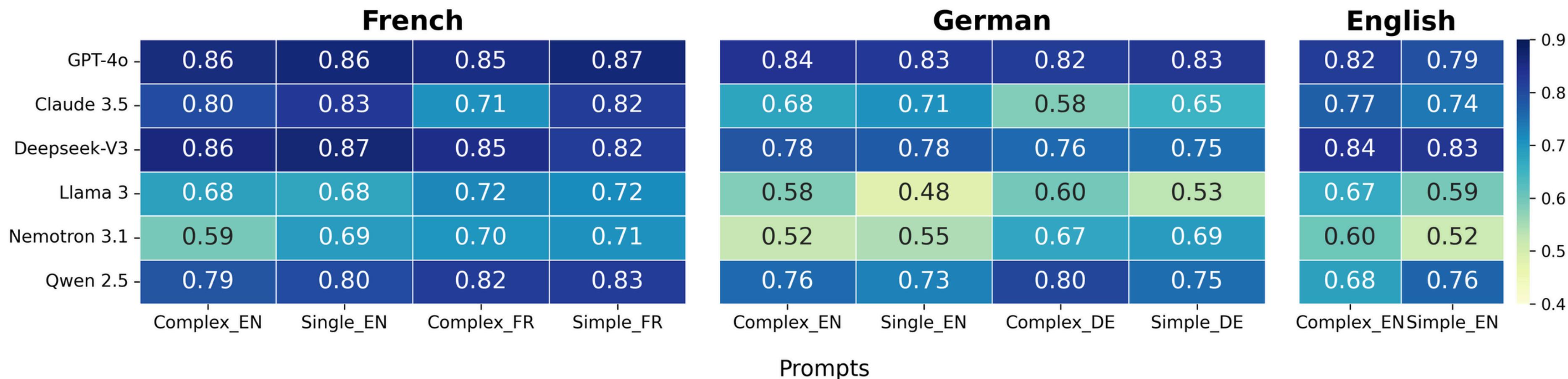
Formel für Textsimilarität

→ vom perfekten Score 1 ausgehend, gibt es proportionale Abzüge bei einer zu hohen Levenshtein-Distanz, sowie falsch gesetzten Artikelgrenzen

- d^{lev} = Levenshtein-Distanz (normalisiert)
- τ = 0.1 (OCR-Toleranz)
- d^{omit} = ausgelassene Zeichen (normalisiert)
- d^{add} = hinzugefügte Zeichen (normalisiert)
- α = 1.5 (Penalty für ausgelassene Zeichen)
- β = 1.0 (Penalty für hinzugefügte Zeichen)
- δ = 0.2 (bei Formatierungsfehler, ansonsten = 0)

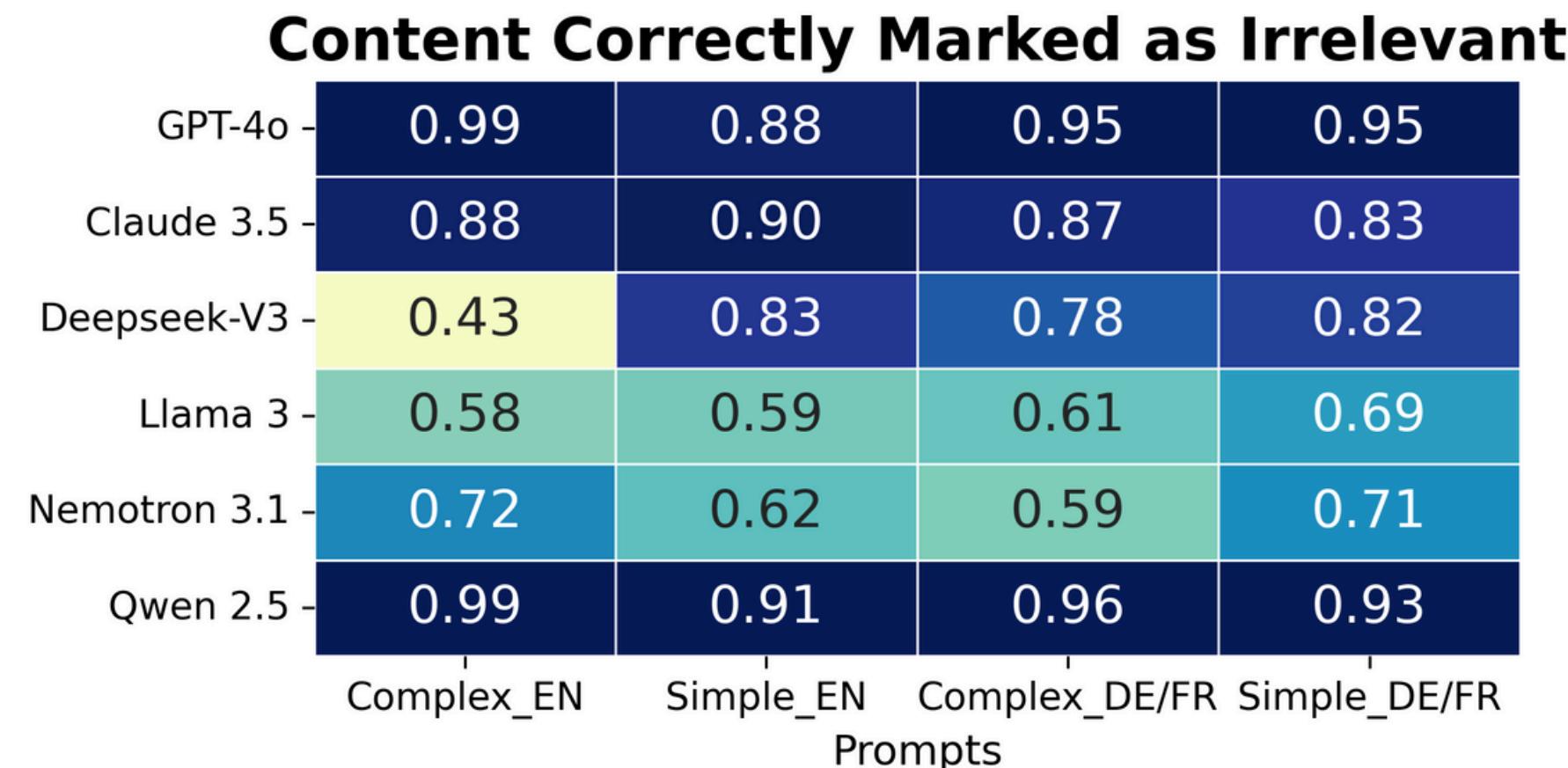
Parameter

Gesamte durchschnittliche Genauigkeit (Prozentsatz korrekter Vorhersagen) für die französischen, deutschen und englischen Datensätze

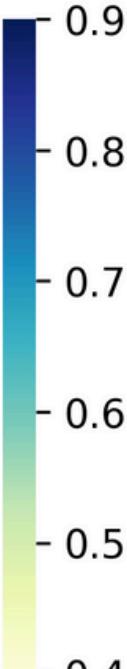
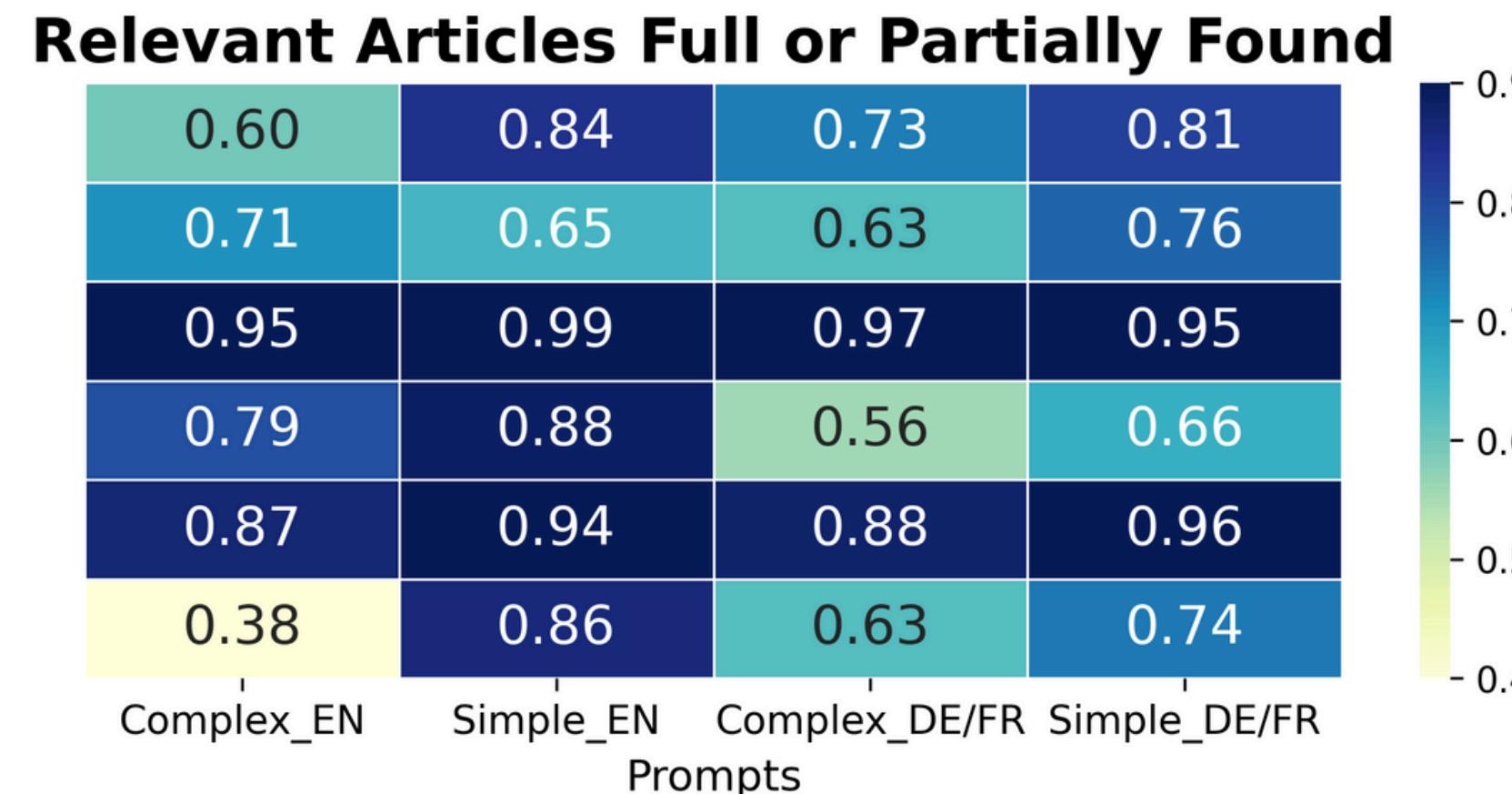


Anteil gefundener relevanter vs. irrelevanter Artikel bei Modelltemperatur von 0.2, aggregiert über alle Sprachen

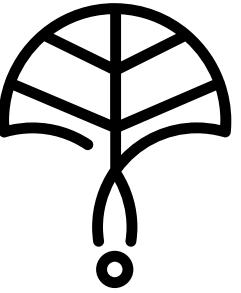
Precision



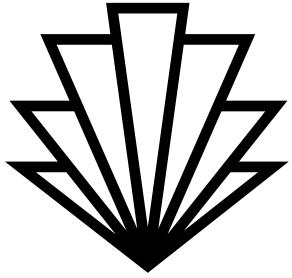
Recall



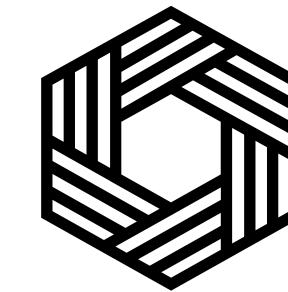
Explainability als Benchmarking-Verbesserung



Die meisten Benchmarks konzentrieren sich ausschließlich auf die Ausgabegenaugigkeit (was das Modell produziert)



Binäre Erfolg/Misserfolg-Metriken verfehlten ein nuanciertes Verständnis des Modellverhaltens

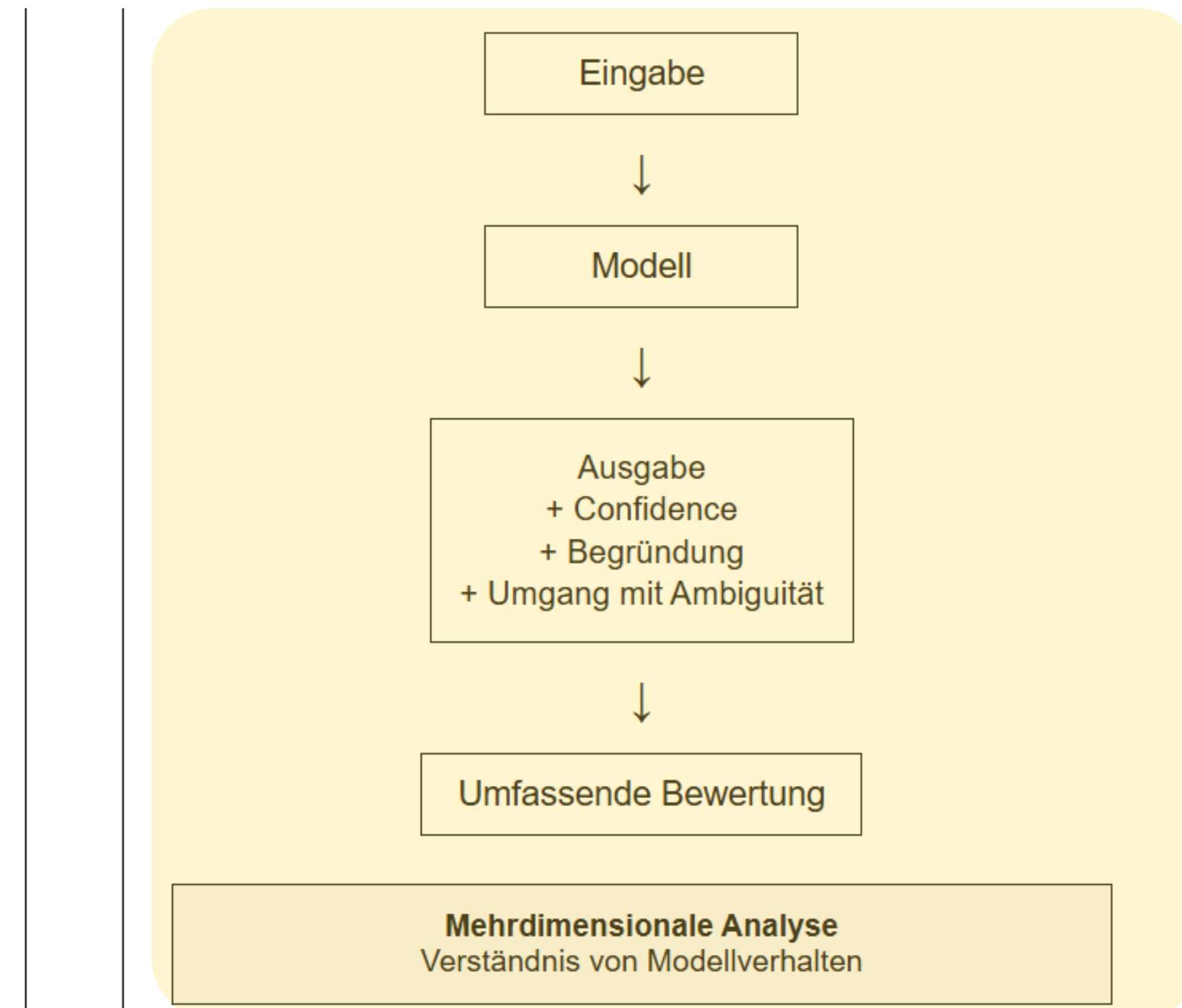


Sie enthüllen nicht, warum Modelle in bestimmten Kontexten versagen oder erfolgreich sind

Übliches Benchmarking



Mehrdimensionales Benchmarking Framework



<article> TERRIBLE WEATHER IN CALABRIA. Reggio, February 12. A violent storm accompanied by showers of hail and heavy rain swept the ruined town last night, while this morning an earthquake shock was felt. <article>

<article> Rome, February 12. An official decree announces that all consignments of goods from abroad destined for sufferers from the earthquake, including provisions, transportable houses, and materials for building, will be allowed access duty free until June 30th. <article>

<verification> Is the unit coherent? Yes. Is the subject present? Yes. Is the article complete? Yes. Have all articles been found? Yes. <verification>

<human_verification_needed> False <human_verification_needed>

<confidence> 0.9 <confidence>

<confidence_explanation> The relevant articles are clearly identifiable, complete, and directly related to the Messina earthquake and its consequences.

<confidence_explanation>

**Beispiel einer
LLM Antwort
mit
Explainability
Elementen**

BIAS-BEWUSSTES BENCHMARKING



Beispiel: Kontextueller Synthesebias

Article from the Dresden Daily:

EARTHQUAKE IN ITALY. Monte leone, December 28.

Extensive damage was done here and at Mileto, Cantazaro, San Giorgio, and other Calabrian districts by a violent earthquake which occurred at 5.50 o'clock this morning. [...]



Claude 3.5 Answer

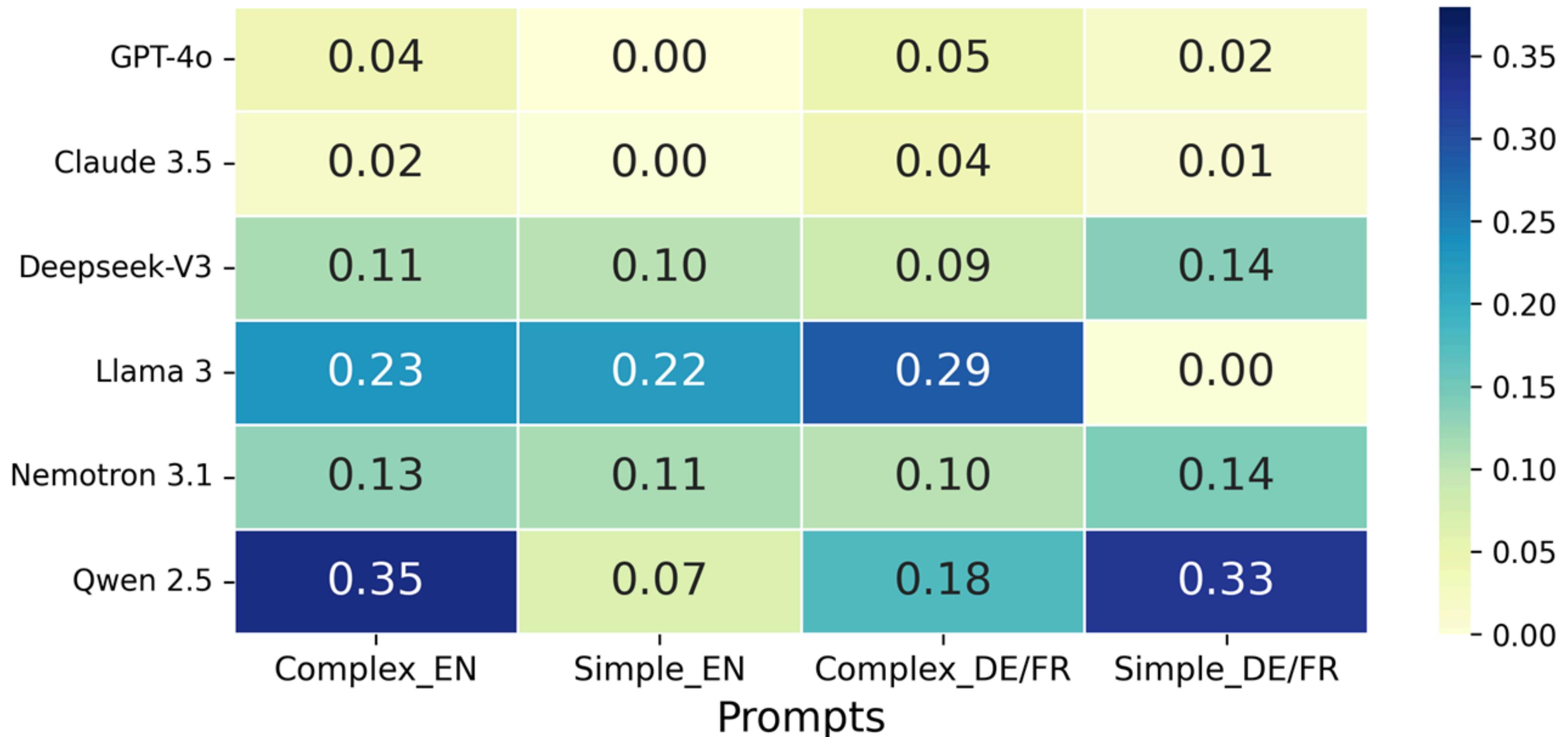
While this article does discuss an earthquake in Italy in December 1908, it appears to be reporting on a different seismic event that occurred in **Calabria** [...]. The Messina earthquake affected **Sicily and Calabria** and was much more devastating, killing over 80,000 people.

Selbst sehr große Sprachmodelle wie Claude schaffen es nicht, geografisches, zeitliches und kontextuelles Wissen verlässlich miteinander in Verbindung zu setzen

BIAS-BEWUSSTES BENCHMARKING



Human Verification Needed



Studying Model Design Biases in LLMs for Multilingual Historical Newspaper Extraction; The Messina Earthquake Case Study

Sarah Oberbichler¹[0000-0002-1031-2759], Johanna
Mauermann²[0009-0000-4790-8196], The Trung Tran³[0009-0005-1261-950X], and
Carlos-Emiliano González-Gallardo⁴[0000-0002-0787-2990]

¹ Leibniz Institute of European History

oberbichler@ieg-mainz.de

² University of Mainz

jmauerma@students.uni-mainz.de

³ University of Science and Technology of Hanoi

trungtt.bi12-452@st.usth.edu.vn

⁴ Université de Tours

gonzalezgallardo@univ-tours.fr

Abstract. Large language models offer new opportunities for processing historical documents, yet their application to historical documents raises questions of reliability. This study presents a systematic framework for