



Community of Practice KIPerWeb

Austausch zur Nutzung und Entwicklung KI-gestützter Webanwendungen



KIPerWEB



Forschungsinstitut
Betriebliche Bildung

- **Update**
 - News & Leaderboard-Update
- **Input**
 - „ Best Practice zur Zusammenfassung und Analyse von Texten mit generativer KI“
- **Diskussion**

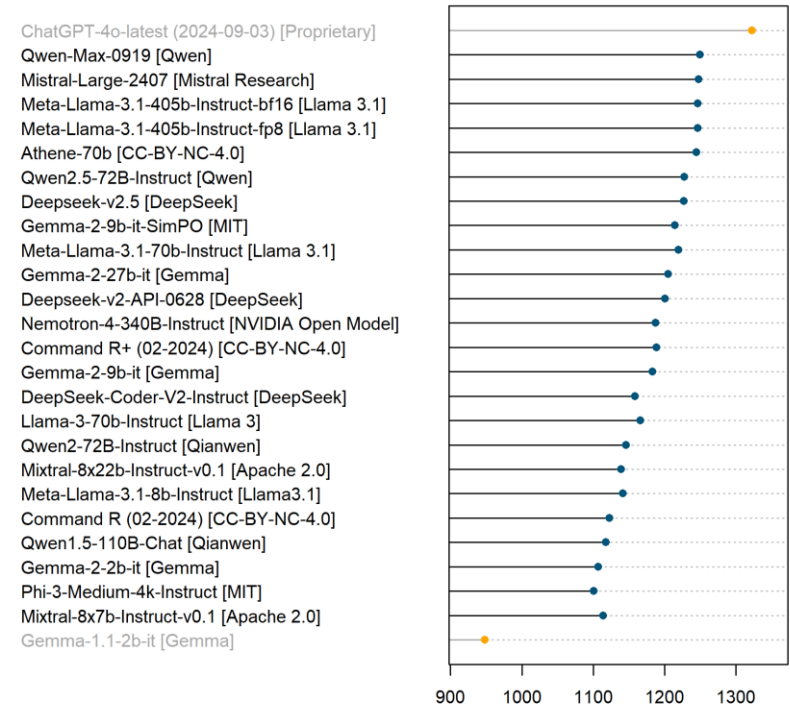
News & Update (15.11.2024)



- ChatGPT-4o-latest (2024-09-03) führt das Leaderboard an mit Arena Score 1323, Schlusslicht ist aktuell Gemma-1.1-2b-it mit 948
 - Beide rechts zur besseren Einordnung eingetragen
- Rechts ausgewiesen sind die Arena Scores in der Kategorie „German“ für alle Modelle mit (a) nicht-proprietärer Lizenz und (b) einem Ranking oberhalb von Mixtral-8x7b-Instruct-v0.1 mit 1114
- Top-Kandidat für den Hausgebrauch on-premises ist m.E. Gemma-2-9b-it-SimPO mit 1215 (unter MIT Lizenz)



Arena Score German
based on Imarena.ai on 15. Nov 2024



Fokusthema: Zusammenfassung und Analyse von Texten



- Prompt an Llama-3.1-70B-Instruct:
„Erstelle eine knappe Beschreibung für ein Cover-Bild zur Veranstaltung "Best Practice zur Zusammenfassung und Analyse von Texten mit generativer KI", die ich als Prompt für eine bildgenerierende KI verwenden kann.“
- Visualisierung von FLUX.1 [schnell] zum Prompt:
„Ein stilisierter Hintergrund mit einer Kombination aus Binär-Code-Elementen und Notizen, die sich zu einem Puzzle zusammensetzen. Im Vordergrund ein großes, modernes Stifte-Set, das auf einem digitalen Tablet liegt. Die Stifte sind um einen zentralen Punkt angeordnet, der eine KI-ähnliche Struktur darstellt. Die Farben sind eine Kombination aus Blau und Grün, die für Technologie und Kreativität stehen.“



- IDEs/IDE-Extensions für „Ghost Text Completion“ (z.B. Github Copilot in VSCode)
- NLP-Methoden z.B. für Clustering und Klassifikation (cf. Bittermann & Fischer, 2024)
 - LLM-Embeddings für semantische Suche & Verschlagwortung (z.B. WISY@KI, 2023)
- LLM-Generierung - ggf. mit RAG für komplexere Texte oder Kategoriensysteme
 - Multi-Turn „Cyborg-Modus“ (cf. Lindner, 2024)
 - Single-Turn „Outsourcing“ - ggf. auch iterativ für multiple Texte/Textteile
 - Topic Labeling (z.B. Rieger et al., 2024)
 - KI-gestützte Qualitative Inhaltsanalyse (Fischer et al., in Vorbereitung)
 - Zusammenfassung/Verdichtung von Texten/Daten
 - Strukturierung/Kategorienzuordnung zu Texten/Daten
 - Explikation/Anreicherung von Texten/Daten



Abkürzungsverzeichnis:

IDE = Integrated Development Environment; NLP = Natural Language Processing;
LLM = Large Language Model; RAG = Retrieval Augmented Generation

Quellen:

Bittermann & Fischer (2024): <https://econtent.hogrefe.com/doi/10.1027/2151-2604/a000568>

Rieger et al. (2024): <https://cran.rstudio.com/web/packages/topiclabels/index.html>

Lindner (2024): https://www.linkedin.com/posts/martin-lindner_ki-kmu-newwork-activity-7224736566567469056-d3vN

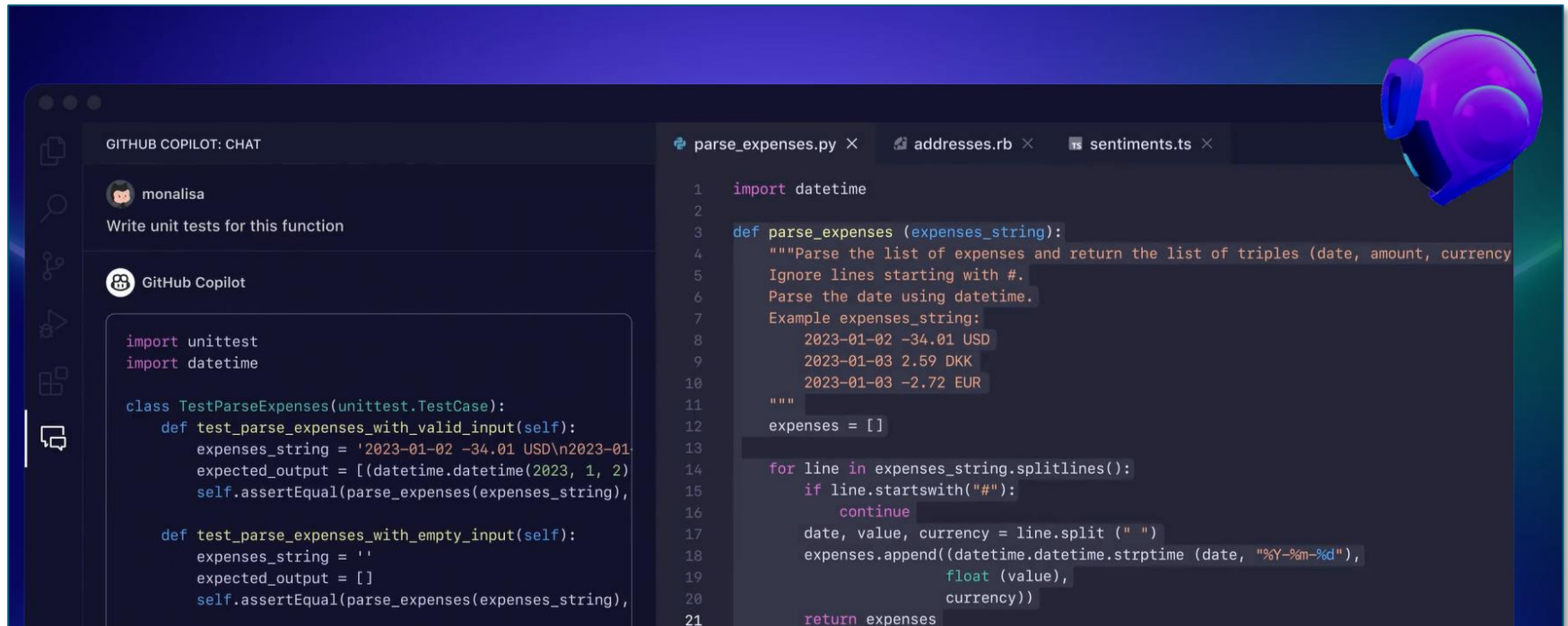
WISY@KI (2023): <https://documenter.getpostman.com/view/5192576/2s935pqP3v#20a20689-c97f-4b3a-8639-d8238eb77152>

Best Practice zu Textanalyse mit moderner KI



- Open-weights Modelle on premises empfehlen sich
 - insb. für replizierbare datenschutzkonforme Lösungen
- Leistungsstarke schlanke Modelle wie Gemma2-9b-it-SimPO hilfreich
 - insb. für effiziente Auswertungen großer Textmengen on-prem 👉 [CoP-Session #14](#)
- Für Verifikation/Lektorat sollte man Zeit und Expertise einplanen (vgl. Potkalitsky, 2024)
 - ggf. stichprobenhaft und evtl. schon vor Festlegung der finalen Analysestrategie
- Multiple Durchläufe bei nicht-deterministischen Analysen oder Modell-/Prompt-Variationen
 - ggf. nach iterativer Prompt-Optimierung iSv. RITE 👉 [CoP-Session #8](#)
- Datenaufbereitung und Prompting-Strategien essenziell für hochwertige Ergebnisse
 - z.B. ROMANE 👉 [CoP-Session #3](#) – generell kann für die Vorhersage jedes Wort zählen
- ...

- KI als Copilot für traditionelle Data Science Workflows (Standalone vs. Extension), z.B.



Verschlagwortung über Text Embeddings

- Embeddings z.B. über *jinaai/jina-embeddings-v2-base-de*
Ähnlichkeiten von
Texten/Abschnitten zu
Schlagworten (cosine-sim.)
- Cutoff z.B. auf Basis von
Sprüngen in der Rangreihe
der Ähnlichkeiten

```
def similarityCutoff(x=[0.2,0.7,0.8,0.9],verbose=True):  
    x0=sorted(x,reverse=True)  
    x1=x0[0:-1]  
    x2=x0[1:]  
    diff = [x1[i]-x2[i] for i in range(len(x0)-1)]  
    if(verbose==True): print(  
        "Max diff at pos "+str(x0.index(x0[diff.index(max(diff))])))+  
        " of sorted list "+str(x0)+  
        " --> cutoff value: "+str(x0[diff.index(max(diff))]))  
    return(x0[diff.index(max(diff))])  
  
similarityCutoff()
```


Verschlagwortung via multiple best-matches



Chunking und Verschlagwortung eines Texts zu Teilqualifikation (TQ):

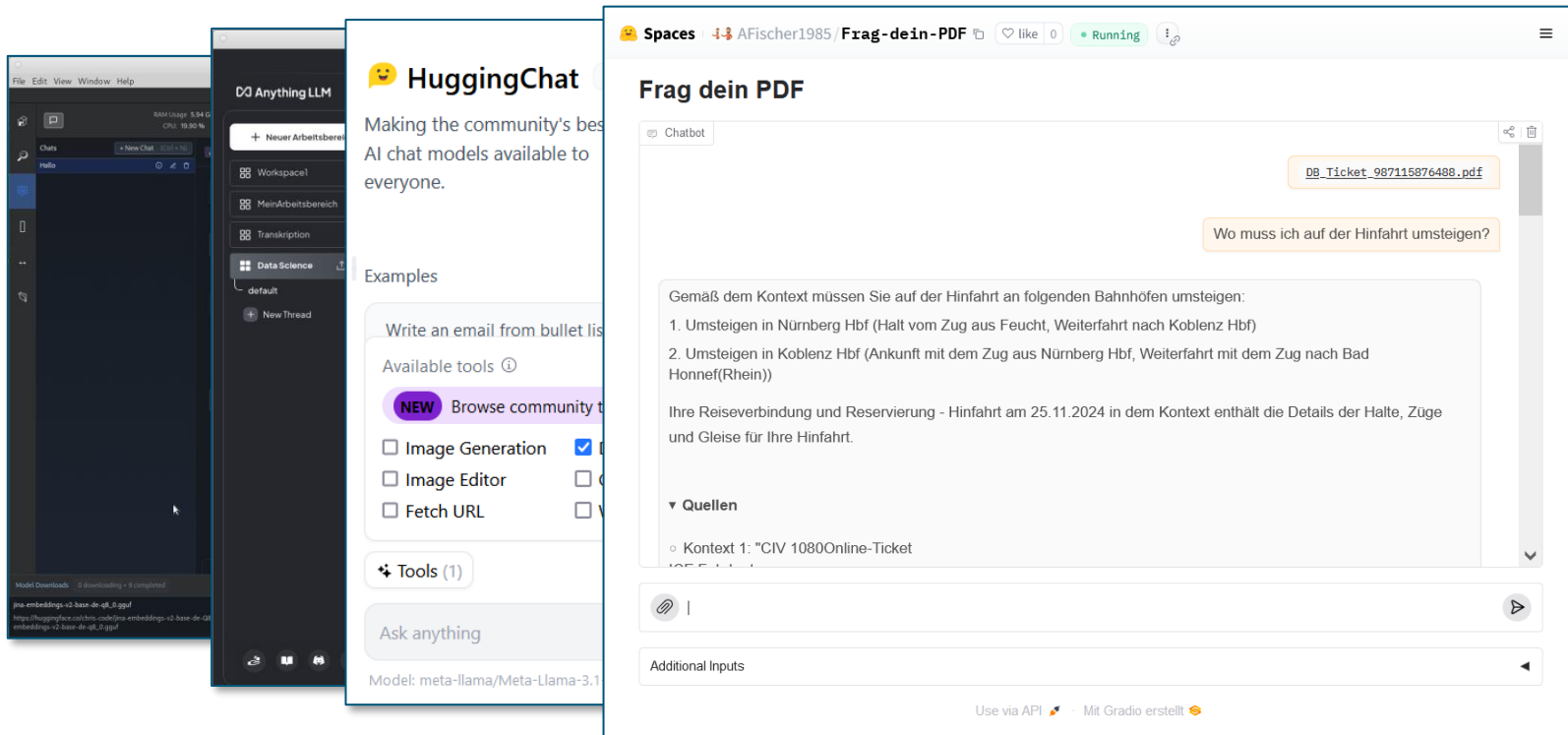
Text (TQ-Beschreibung)	Best match (ESCO-skills/-occupations)	Ähnlichkeit
Fachinformatiker:in, Daten und Prozessanalyse - Teilqualifizierung 5 - Datenbank; Datenbankgrundlagen; Datenbankmanagement mit My SQL; Objektorientiertes PHP; Weiterführende Programmiertechniken mit Java Script	PHP: Techniken und Grundsätze der Softwareentwicklung wie Analyse, Algorithmen, Programmierung, Testen und Kompilieren von Programmierparadigmen in PHP.	0,96889228
Fachinformatiker:in, Daten und Prozessanalyse - Teilqualifizierung 5 - Datenbank	Akademische und vergleichbare Fachkräfte für Datenbanken und Netzwerke	0,96902251
Datenbankgrundlagen	Datenbankentwickler und -administratoren	0,97706938
Datenbankmanagement mit My SQL	Design und Verwaltung von Datenbanken und Netzwerken	0,97528487
Objektorientiertes PHP	PHP: Techniken und Grundsätze der Softwareentwicklung wie Analyse, Algorithmen, Programmierung, Testen und Kompilieren von Programmierparadigmen in PHP.	0,95962518
Weiterführende Programmiertechniken mit Java Script	JavaScript: Techniken und Grundsätze der Softwareentwicklung wie Analyse, Algorithmen, Programmierung, Testen und Kompilieren von Programmierparadigmen in JavaScript.	0,97539222

Topic Labeling via Topiclabels

- Prompt-Beispiel: *"You are an expert in labeling outputs from topic models.\n\nCould you please help me labeling the following topic based on its top terms? "The top terms of the topic are listed below:\n\n{terms}\n\nOutput your response in JSON format with a single field called 'label', specifying the best label for the topic in no more than {max_length_label} words."*
- Default-Modell: *mistralai/Mixtral-8x7B-Instruct-v0.1*

Text Analysis via Generative Question Answering

„Chat with your document“ via LMStudio, AnythingLLM, HuggingChat, „Frag-dein-PDF“, etc.



The image displays three overlapping screenshots of AI chat interfaces. The background screenshot is LMStudio, showing a file explorer and a chat window. The middle screenshot is AnythingLLM, showing a chat window with a PDF document. The foreground screenshot is HuggingChat, showing a chat session with a PDF document titled "Frag dein PDF".

HuggingChat Interface:

- Header:** Spaces | AFischer1985 / Frag-dein-PDF | 0 likes | Running
- Chatbot:** Chatbot
- Input:** DB_Ticket_987115876488.pdf
- Question:** Wo muss ich auf der Hinfahrt umsteigen?
- Answer:**

Gemäß dem Kontext müssen Sie auf der Hinfahrt an folgenden Bahnhöfen umsteigen:

 1. Umsteigen in Nürnberg Hbf (Halt vom Zug aus Feucht, Weiterfahrt nach Koblenz Hbf)
 2. Umsteigen in Koblenz Hbf (Ankunft mit dem Zug aus Nürnberg Hbf, Weiterfahrt mit dem Zug nach Bad Honnef(Rhein))

Ihre Reiseverbindung und Reservierung - Hinfahrt am 25.11.2024 in dem Kontext enthält die Details der Halte, Züge und Gleise für Ihre Hinfahrt.
- Quellen:**
 - Kontext 1: *CIV 1080Online-Ticket
- Footer:** Use via API | Mit Gradio erstellt