



# Community of Practice KIPerWeb

Austausch zur Nutzung und Entwicklung KI-gestützter Webanwendungen



**KIPerWEB**



Forschungsinstitut  
Betriebliche Bildung

- **Update**
  - News & Leaderboard-Update
- **Input**
  - „KI und der Chat mit den eigenen Daten“
- **Diskussion**

# News & Update (06.11.2024)



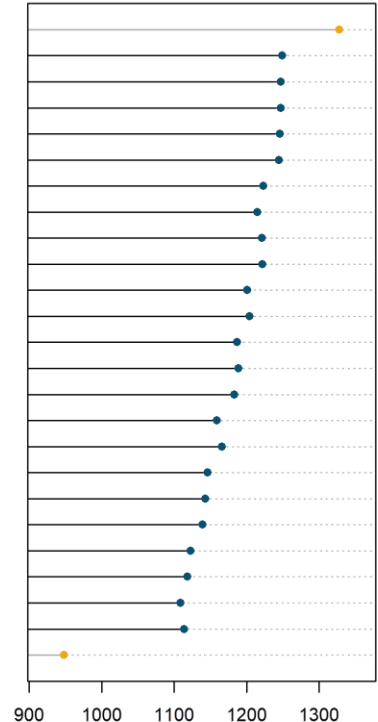
- ChatGPT-4o-latest (2024-09-03) führt das Leaderboard an mit Arena Score 1327, Schlusslicht ist aktuell Gemma-1.1-2b-it mit 948
  - Beide rechts zur besseren Einordnung eingetragen
- Rechts ausgewiesen sind die Arena Scores in der Kategorie „German“ für alle Modelle mit (a) nicht-proprietärer Lizenz 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)
- Neues abseits des Leaderboard?
  - SD 3.5, Red-panda, Recraft v3
  - Mini-Omni2, LongVU
  - MobileLLM, SmolLm2, OLMo-1b
  - SauerkrautLM-v2-14b
  - Oasis (AI Minecraft: <https://oasis.decart.ai/overview>)



## Arena Score German

based on lmarena.ai on 06.11.2024

ChatGPT-4o-latest (2024-09-03) [Proprietary]  
Qwen-Max-0919 [Qwen]  
Meta-Llama-3.1-405b-Instruct-bf16 [Llama 3.1]  
Mistral-Large-2407 [Mistral Research]  
Meta-Llama-3.1-405b-Instruct-fp8 [Llama 3.1]  
Athene-70b [CC-BY-NC-4.0]  
Deepseek-v2.5 [DeepSeek]  
Gemma-2-9b-it-SimPO [MIT]  
Qwen2.5-72b-Instruct [Qwen]  
Meta-Llama-3.1-70b-Instruct [Llama 3.1]  
Deepseek-v2-API-0628 [DeepSeek]  
Gemma-2-27b-it [Gemma]  
Nemotron-4-340B-Instruct [NVIDIA Open Model]  
Command R+ (02-2024) [CC-BY-NC-4.0]  
Gemma-2-9b-it [Gemma]  
DeepSeek-Coder-V2-Instruct [DeepSeek]  
Llama-3-70b-Instruct [Llama 3]  
Qwen2-72B-Instruct [Qianwen]  
Meta-Llama-3.1-8b-Instruct [Llama3.1]  
Mixtral-8x22b-Instruct-v0.1 [Apache 2.0]  
Command R (02-2024) [CC-BY-NC-4.0]  
Qwen1.4-110B-Chat [Qianwen]  
Gemma-2-2b-it [Gemma]  
Mixtral-8x7b-Instruct-v0.1 [Apache 2.0]  
Gemma-1.1-2b-it [Gemma]



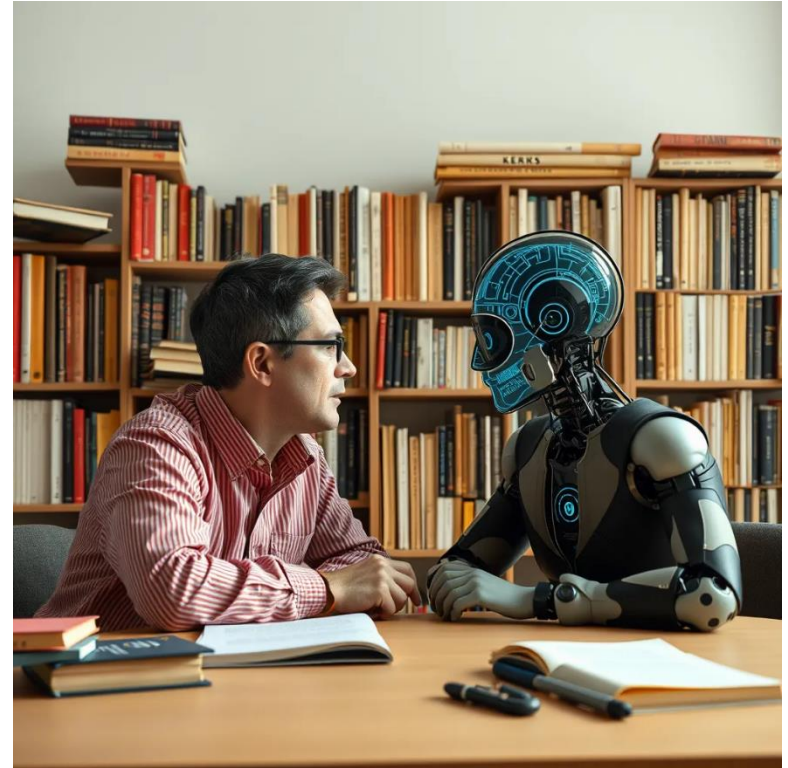
Datengrundlage: <https://chat.lmsys.org/?leaderboard>

(c) 2024 Andreas Fischer

# Fokusthema: KI und der Chat mit den eigenen Daten

- Prompt „Ein Mensch chattet mit einer Künstlichen Intelligenz. Im Hintergrund sind Stapel von Büchern“ interpretiert von FLUX.1 [schnell]

Links Variante 2:  
„Eine Frau und eine Künstliche Intelligenz lesen gemeinsam ein Buch.“



# Generative Question Answering



- „Chat with your document“ via LMStudio (>0.3), AnythingLLM, HuggingChat, etc.

The image displays three overlapping screenshots of AI chat interfaces. On the left is LMStudio, showing a chat window with a user message 'Hallo' and an AI response in German. In the center is AnythingLLM, showing a workspace with various tools like 'Transkription' and 'Data Science'. On the right is HuggingChat, showing a chat window with a user message 'Ich mache ne game. Würde empfehlen?' and an AI response in German. The HuggingChat interface also shows a 'Current Model' dropdown set to 'meta-llama/Meta-Llama-3.1-70B-Instruct' and a list of 'Available tools' including 'Image Generation', 'Image Editor', 'Fetch URL', 'Document Parser', 'Calculator', and 'Web Search'. A hand cursor is pointing at the 'Document Parser' tool.

# Alternative: Extractive Question Answering



**Hugging Face** Search models, datasets, users... Models Datasets Spaces Posts Docs Pricing Log In Sign Up

Tasks

## Question Answering

Question Answering models can retrieve the answer to a question from a given text, which is useful for searching for an answer in a document. Some question answering models can generate answers without context!

**Inputs**  
**Question**  
Which name is also used to describe the Amazon rainforest in English?  
**Context**  
The Amazon rainforest, also known in English as Amazonia or the Amazon Jungle

**Question Answering Model**

**Output**  
**Answer**  
Amazonia

Deploy on Inference Endpoints

**Compatible libraries**

Adapters AllenNLP Transformers Transformers.js

**Question Answering demo**

using [deepset/roberta-base-squad2](#)

Question Answering Examples

Your sentence here... Compute

Context

Please input some context...

View Code Maximize

**Models for Question Answering** Browse Models (12,423)

# Exemplarische RAG-Prompts



Hier exemplarisch RAG-Prompts von AnythingLLM & LMStudio 0.3.5:

- **AnythingLLM (System Prompt RAG Scenario):** 'Given the following conversation, relevant context, and a follow up question, reply with an answer to the current question the user is asking. Return only your response to the question given the above information following the users instructions as needed.\n\nContext: \n[CONTEXT 0]:\n<document\_metadata>\nsourceDocument: *{title1}*\npublished: *{date1}*\n</document\_metadata>\n\n*{text1}*\n[END CONTEXT 0]\n\n'
- **LMStudio (Single-Turn RAG scenario):** 'The following citations were found in the files provided by the user:\n\nCitation 1: "*{text1}*"\n\nCitation 2: "*{text2}*"\n\nCitation 3: "*{text3}*"\n\nUse the citations above to respond to the user query, only if they are relevant. Otherwise, respond to the best of your ability without them.\n\nUser Query:\n\n*{query}*'
- **LMStudio (Single-Turn Enriched Context scenario):** 'This is a Enriched Context Generation scenario.\n\nThe following content was found in the files provided by the user.\n\n\*\* *{title}* full content \*\*\n\n*{text}* \*\* end of [object Object] \*\*\n\nBased on the content above, please provide a response to the user query.\n\nUser query: *{query}*'

# Formatierung des RAG-Kontextes?

## Varianten?

- Prefixes für Kontext-Dokumente ("Context #1"),
- Rahmung (z.B. "BEGINCONTEXT" & "ENDCONTEXT", „<context></context>“)
- Newline-Trennung (+ Entfernung von Newlines aus Kontext-Dokumenten)
- Erfahrung(s)wissen:
  - Gute Erfahrungen habe ich mit einem Prefix und Anführungszeichen:  
*Kontext 1: "Text"*
  - Bei Rahmung mit XML-Tags wie *<Kontext 1>Text</Kontext 1>* scheint Mixtral häufig ins Englische zu wechseln

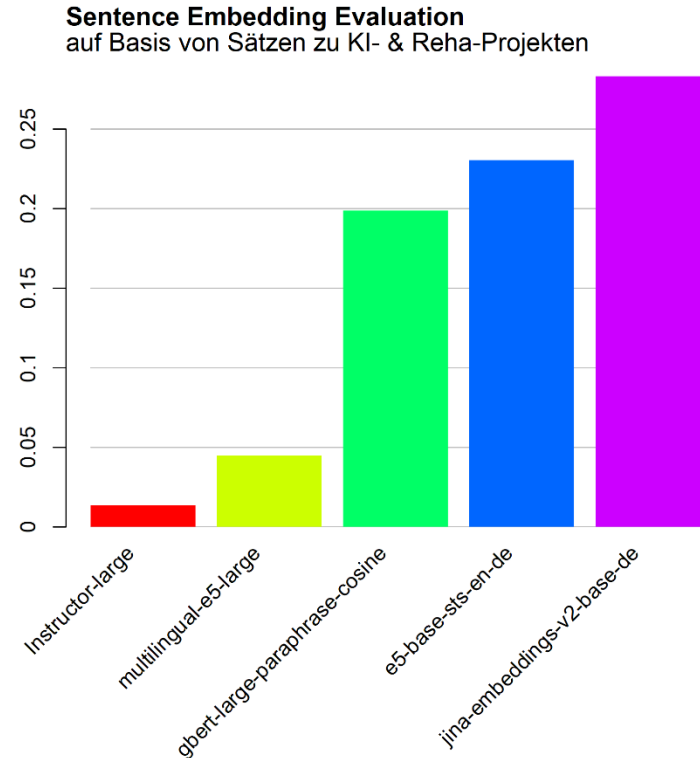


# Formatierung von Query & History für RAG

- Umformulierung des Dialogs durch LLM (pot. ressourcenintensiv)
- Routing durch LLM (dito)
- Nur letzter Query (oft schlecht bei Nachfragen)
- Formatierter Dialog-Prompt (Formatierungs-Overhead)
- Schlicht formatierter Dialog-Prompt (z.B. „USER: Test.\nASSISTANT:Test“)
- Letzte User-Queries only (vermeidet Dialog-Struktur und spart Zeichen)
- Erfahrung(halb)wissen:
  - Gute Erfahrungen habe ich damit gemacht, die erste Anfrage als Text zu embedden und bei multi-turn-Dialogen den Dialog (oder Auszüge) schlicht als Fließtext zu formatieren
  - Für Embedding-Modelle sind fremde prompt-templates mutmaßlich „Code“

# Embedding-Modelle für deutsche Texte?

- Auf der Suche nach neuen Embedding-Modellen habe ich recherchiert und verglichen
- Die Grafik rechts zeigt für ausgewählte Themen (KI und Reha) eine Bewertung, die umso größer ausfällt, je *eindeutiger* die Embeddings von Queries zu KI und Reha jeweils thematisch passenden Sätzen zugeordnet werden (bei 4 Queries und 4 Sätzen je Thema)
- Welche Modelle sind noch empfehlenswert?



# Beispiel: Frag-dein-PDF – RAG mit User-Content



- **Frag-dein-PDF:** Online-Demo auf Basis von *Jina-Embeddings-v2-base-de* & *Mixtral-8x7B-Instruct-v0.1* (on-prem eher *gemma-2-9b-it-SimPO*)
- System Prompt analog zu AnythingLLM, aber auf Deutsch
- Chunking mit Overlap und auf Basis von Absatz- & Satzgrenzen (regex)
  - Viel besser und robuster als einfache harte Zeichengrenzen

