Community of Practice KI*PerWeb*

Austausch zur Nutzung und Entwicklung KI-gestützter Webanwendungen

KIPERWEB

f-bb Forschungsinstitut Betriebliche Bildung

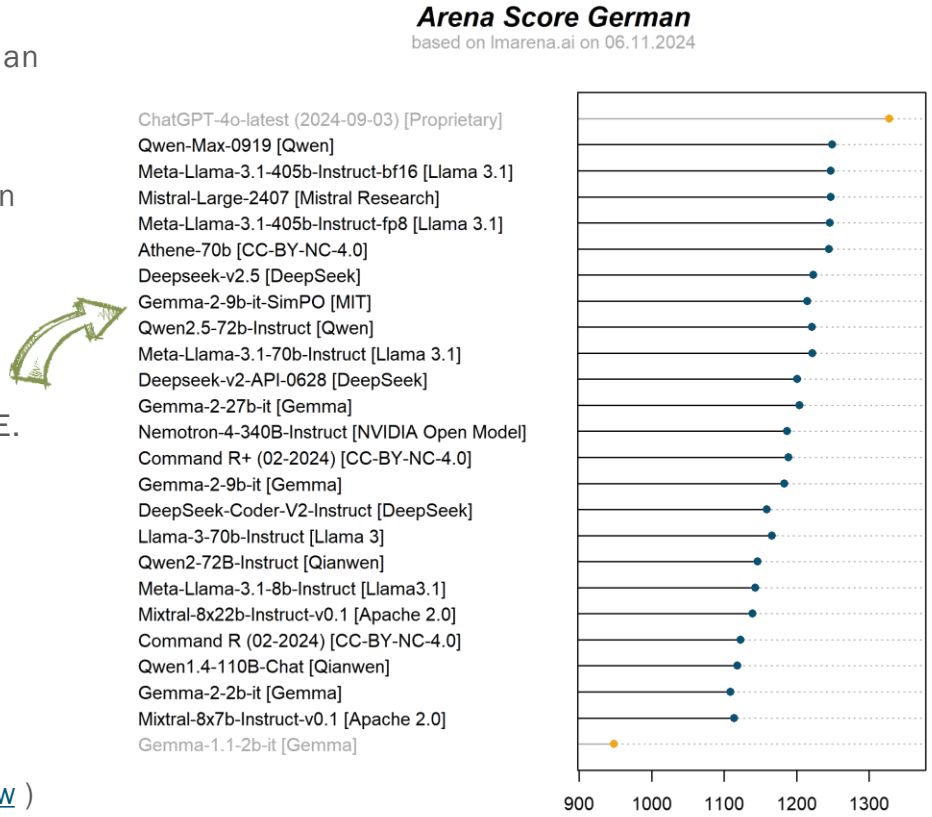
Agenda

- 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)

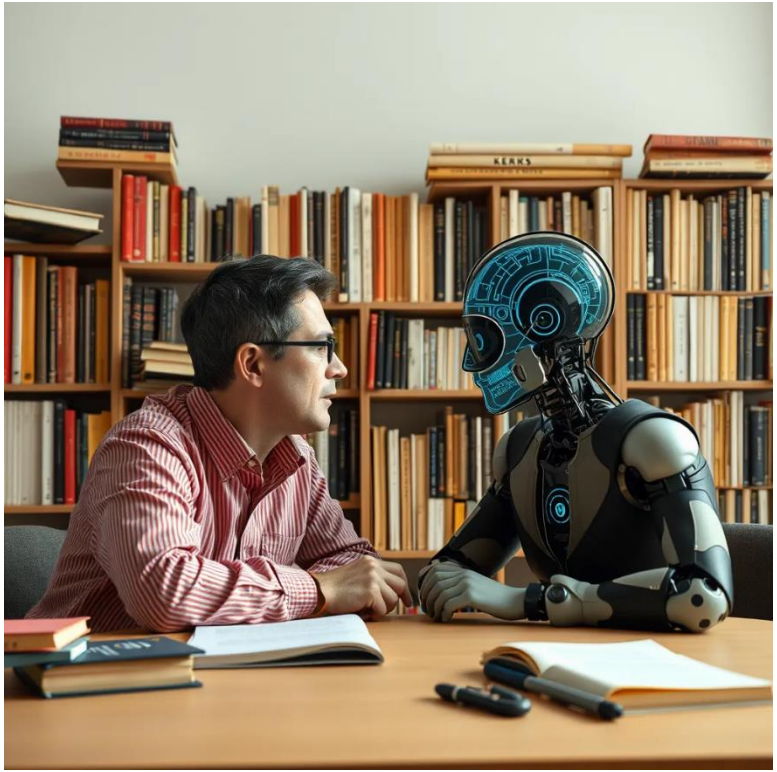
- Neues abseits des Leaderboard?
  - SD 3.5, Red-panda, Recraft v3
  - Mini-Omni2, LongVU
  - MobileLLM, SmolLm2, OLMo-1b
  - SauerkrautLM-v2-14b
  - Oasis (AI Minecraft: https://oasis.decart.ai/overview)



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

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]

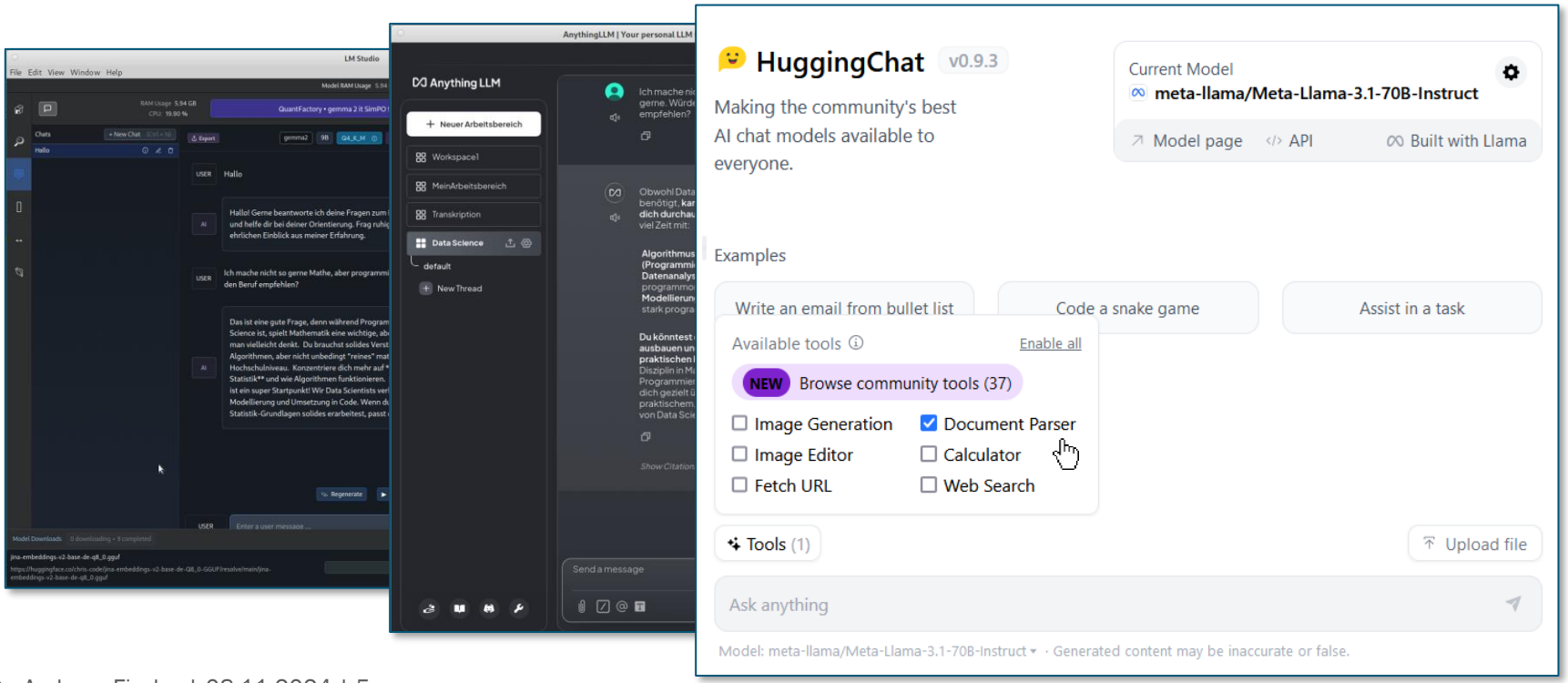


Links Variante 2: „Eine Frau und eine Künstliche Intelligenz lesen gemeinsam ein Buch.“

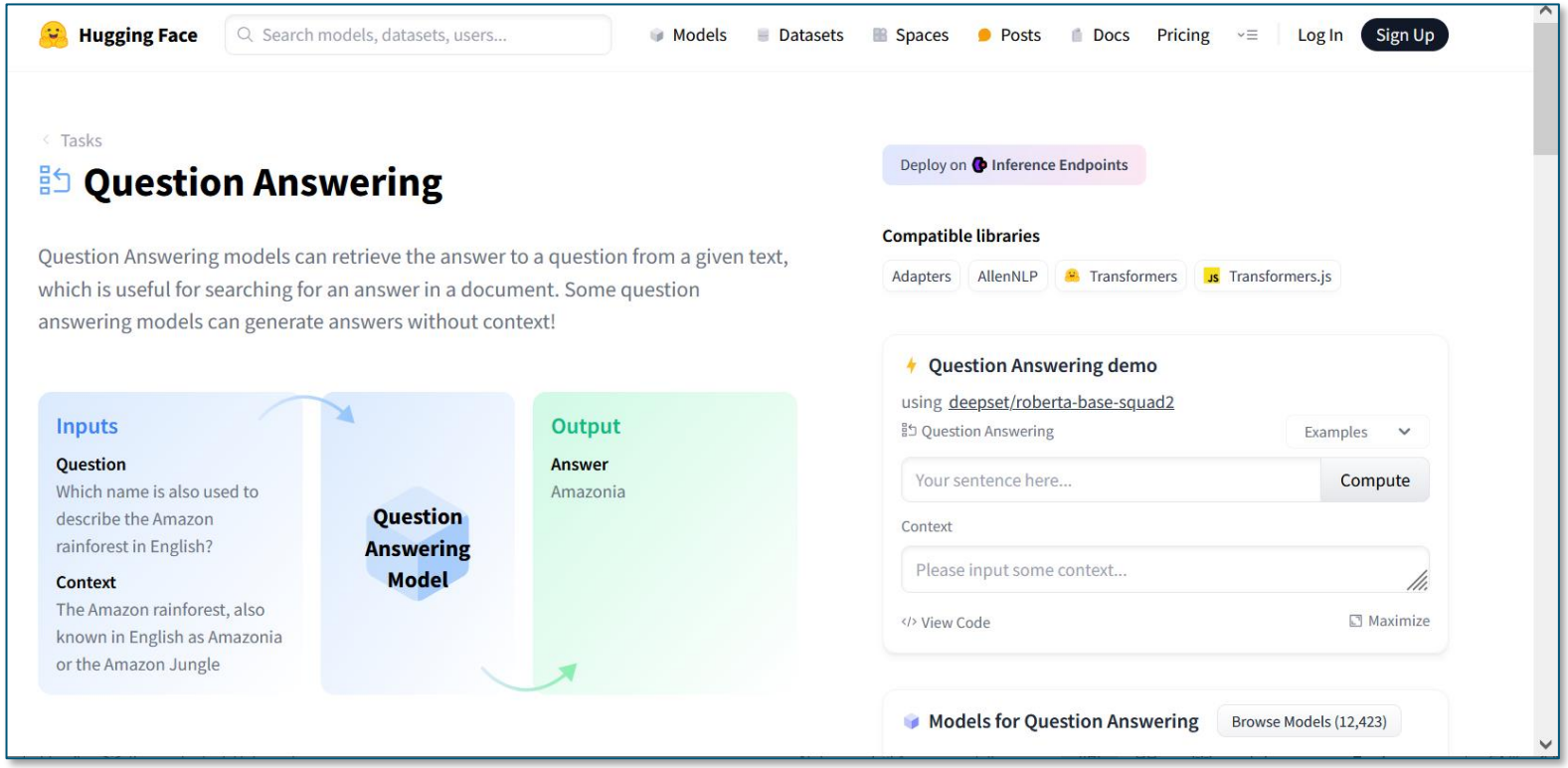
Quelle: https://huggingface.co/spaces/black-forest-labs/FLUX.1-schnell

Generative Question Answering

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



Alternative: Extractive Question Answering



Quelle: https://huggingface.co/tasks/question-answering
cf. https://www.kaggle.com/code/arunmohan003/question-answering-using-bert

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\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)

- Erfahrungs(halb)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

https://www.reddit.com/r/LocalLLaMA/comments/1b4505o/best_practive_for_formatting_context_in_rag/?&rdt=50130

Formatierung von Query & History für RAG

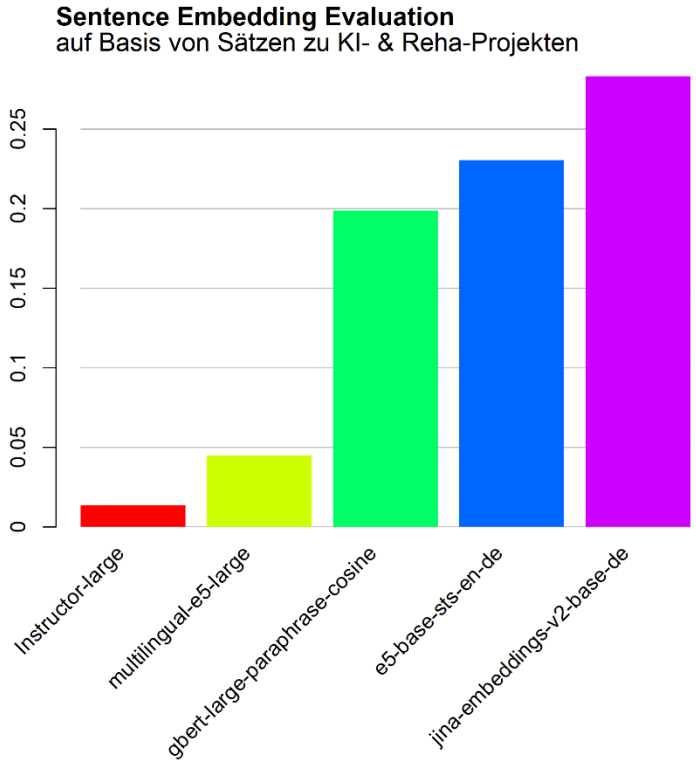
- Umformulierung des Dialogs durch LLM (pot. ressourcenintensiv)
- Routing durch LLM (dito)
- Nur letzter Query (oft schlecht bei Nachfragen)
- Formatierter Dialog-Prompt (Formatierungs-Overhead)
- Schlicht formatierter Dialog-Prompt (z.B. „USER: Test.\nASSISTANT:Test“)
- Letzte User-Queries only (vermeidet Dialog-Struktur und spart Zeichen)

- Erfahrungs(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“

https://www.reddit.com/r/LocalLLaMA/comments/1b4505o/best_practive_for_formatting_context_in_rag/?&rdt=50130

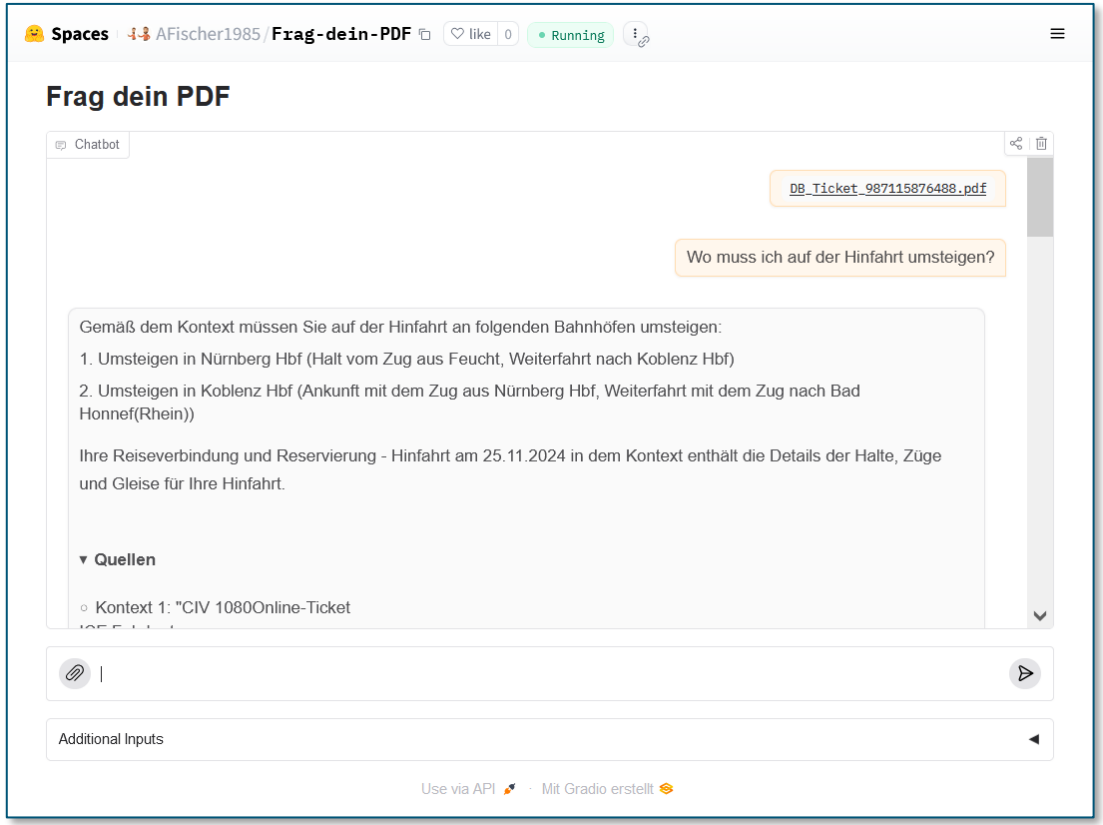
Embedding-Modelle für deutsche Texte?

- 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



Quelle: https://huggingface.co/spaces/AFischer1985/Frag-dein-PDF