Community of Practice KI*PerWeb*

Austausch zur Nutzung und Entwicklung KI-gestützter Webanwendungen

KIPERWEB

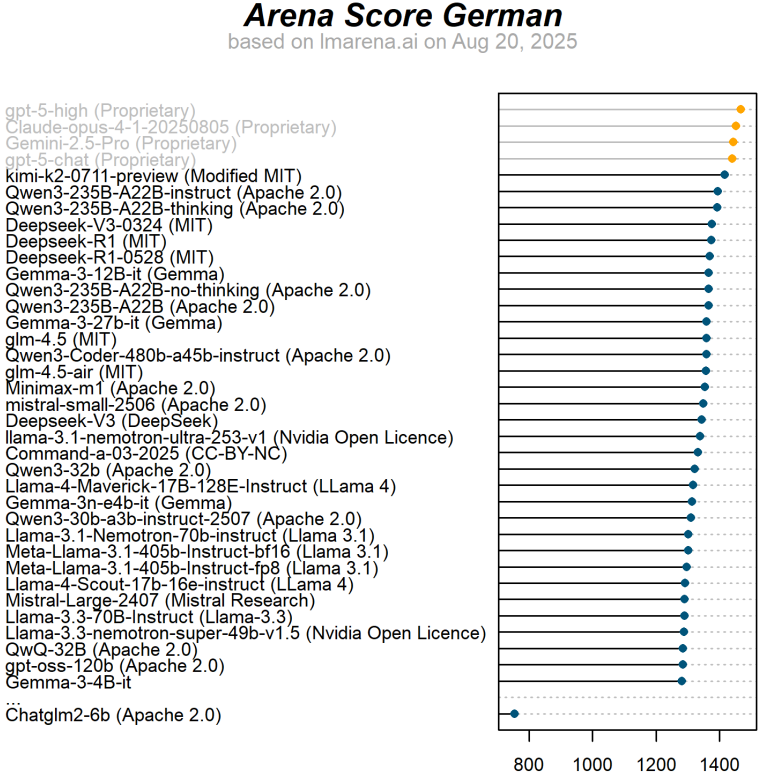
f-bb Forschungsinstitut Betriebliche Bildung

Agenda

- Update
  - News & Leaderboard-Update
- Input
  - "Multimodal Retrieval-Augmented Generation on-premises?"
    (Gastbeitrag: Henry Herkula)
- Diskussion

Leaderboard-Update (20.08.2025)

- Gpt-5-high (nicht gpt-5-chat) löst Gemini-2.5-Pro als ehemaligen Spitzenreiter ab
- Arena-Scores von *nicht*-proprietären Modelle sind rechts ausgewiesen sofern sie mindestens das Niveau von Gemma-3-4B-it erreichen:
  - OpenAIs gpt-oss-120b enttäuscht weit hinten (gpt-oss-20b ist mit 1262 gar nicht mehr im Bild weil hinter Gemma3-4B-it mit 1282)
- Schlusslicht auf dem Leaderboard bleibt Chatglm2-6b



Datengrund

(c) 2025 Andreas Fischer

Fokusthema: Multimodal Retrieval-Augmented Generation

- Prompt:
  „Multimodal Retrieval-Augmented Generation on-premises“

  (rechts visualisiert von FLUX.1-schnell, seed 1211218464)

...

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

Gastbeitrag

- Henry Herkula zu "Multimodal Retrieval-Augmented Generation on-premises?"

Brandenburgische Technische Universität Cottbus - Senftenberg

Benutzerfreundliches Retrieval-Augmented-Generation (RAG)

Welche lokalen Lösungen sind zuverlässig?

Henry Herkula
Künstliche Intelligenz, LLM, RAG, Lokale
Community of Praxis KIPerWeb
2025-08-22, Cottbus

Zentrum für wissenschaftliche Weiterbildung

Brandenburgische Technische Universität Cottbus - Senftenberg

Inhaltsverzeichnis

1. Was ist lokale KI?

- Eine **lokale Künstliche Intelligenz** ist ein Programm auf dem eigenen Computer, das nicht mit dem Internet verbunden sein muss, um verschiedene KI-Fähigkeiten nutzen zu können.
- **Retrieval-Augmented-Generation (RAG)** ist die Bezeichnung für ein Verfahren, mit dem ein KI-Modell auf Informationen außerhalb der eigenen Trainingsdaten zugreifen kann und darauf aufbauend eine Antwort generiert. Es ist ein mögliches Element in einem KI-Prozessablauf.

FLUX.1 [schnell], “Retrieval-Augmented-Generation”

1. Was ist lokale KI?

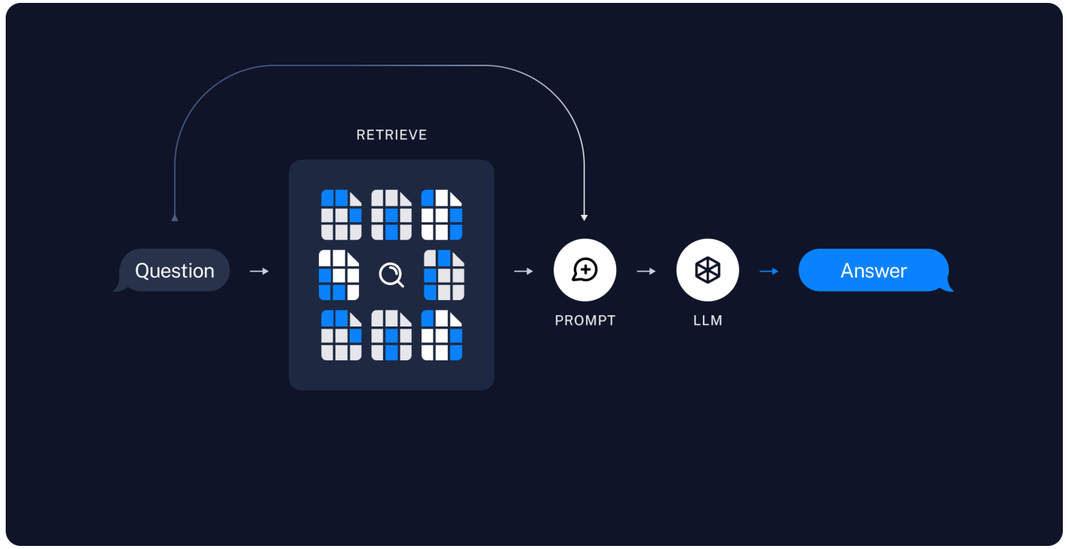
- KI-Prozessabläufe bestehen aus Ketten von Anweisungen, die hintereinander ausgeführt werden. Dabei können sie sehr unterschiedlich programmiert sein.
- Wenn zum Beispiel ein Link in das Chatfenster eines Chatbots eingetragen wird, dann kann eine kleinere KI oder ein Algorithmus zunächst entscheiden, ob ein Skript oder ein Agent eine Webseite auslesen soll oder ob dieser Link vielleicht schon vor fünf Minuten schon einmal ausgelesen wurde und die Indexierung übersprungen werden kann.
- Wenn die Webseite noch nicht ausgelesen wurde, dann werden zum Beispiel die Textdaten der Webseite anhand einer Auslesestrategie (Chunking) in Vektorrepräsentationen (Zahlen) umgewandelt und gemeinsam mit dem jeweiligen Textausschnitt in einer Datenbank abgelegt.
- Mit dem Auslesen der Daten aus der Vektordatenbank startet anschließend der RAG-Prozess. Die Ausschnitte der Webseite können nun mit der Vektorrepräsentation der

1. Was ist lokale KI?

eigenen Eingabe verglichen werden, damit die passendsten Ergebnisse ausgewählt werden.

- Die wahrscheinlichsten Passagen werden nach der höchsten Übereinstimmung sortiert und auf Basis eines Cutoffs (Top-K [Anzahl] oder Top-P [Wahrscheinlichkeit]) anschließend für die Augmentierung mit der eigenen Textangabe allerdings unsichtbar kombiniert.
- Der aus der eigenen Eingabe und dem angehängten Text entstandene Prompt wird nun vom KI-Modell genutzt, um eine Antwort mit den gezielt ausgewählten Informationen zu generieren.

1. Was ist lokale KI?



Beispiel: Der LangChain-RAG-Prozess mit einer Frage als Nutzereingabe und einer Antwort als Ausgabe, https://python.langchain.com/docs/tutorials/rag/

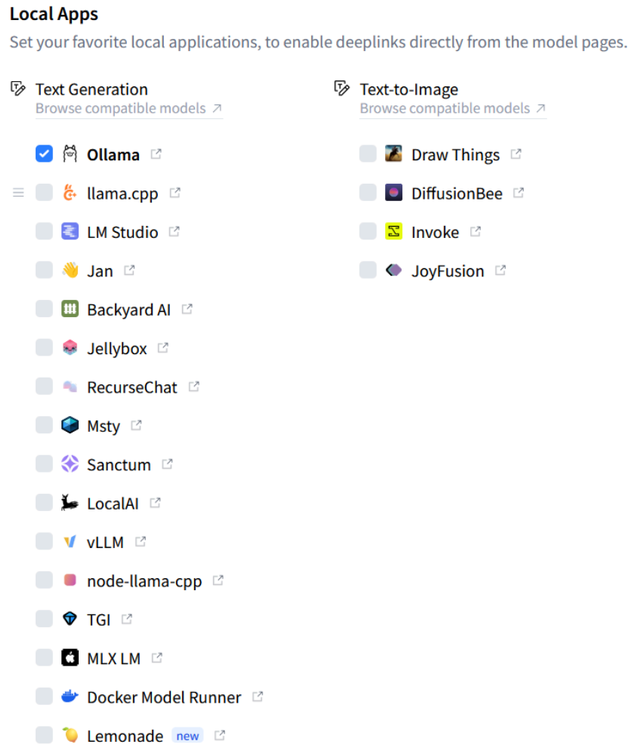
- Der große Vorteil eines RAGs besteht darin, dass die Antworten aktueller, faktenbasierter und nachvollziehbarer werden können, ohne dass das Modell neu trainiert werden muss.
- **Bedeutung**: Ein lokales RAG macht diese Funktionalität auch ohne Abhängigkeit von großen Unternehmen für jeden sicher und vertraulich zugänglich.

2. Agentic RAG

- RAG-Prozess, der unabhängige Programme (Tools) nutzt, um den Auslese-Teil des Programms zu verbessern.
- **Beispiele**:
  - Erstellung von individuellen SQL-Abfragen, um mit Informationen aus Datenbank-Tabellen arbeiten zu können
  - Auswahl an geeigneten Programmen, die Chunks je nach Anfrage neu einlesen
- Agentic RAG ist nur mit individuellen Templates zu erhalten. Keine meiner getesteten Benutzeroberflächen setzt bisher so tief an.

3. Untersuchte Anwendungen

- Auswahl an Anwendungen, die KI auf dem eigenen Computer laufen lassen, ohne dass eine Internetverbindung nötig ist.
- Hugging Face ist eine der größten Seite für die Veröffentlichung von KI-Modellen im Internet.
- Wenn eine lokale Anwendung es hierher schafft, hat sie eine bestimmte Reife erreicht.
- Getestet wurde die Möglichkeit, Dateien als Eingabe für beliebige Modelle zu verwenden.



Liste lokaler Anwendungen, die von Hugging Face unterstützt werden: https://huggingface.co/settings/local-apps, 2025-08-22

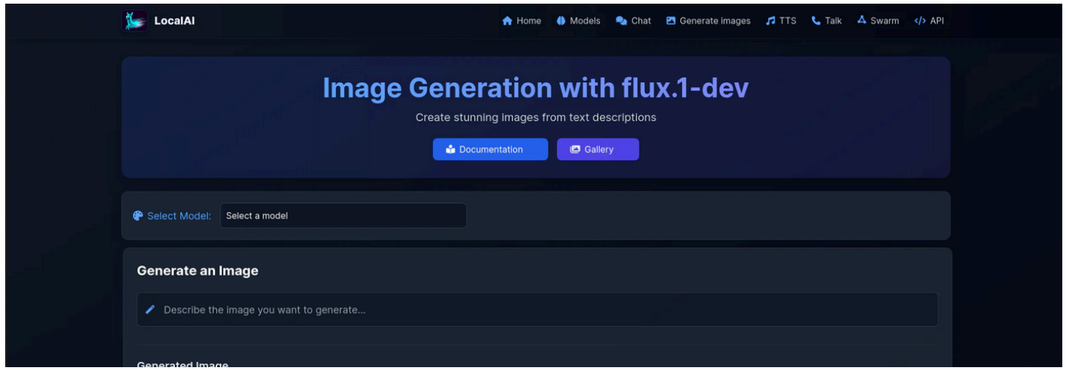
4. Einfachste Anwendung: Ollama

- MIT-Lizenz, Open Source, kostenlos und kommerziell nutzbar
- Einfache Benutzeroberfläche mit Drag-&-Drop-Funktionalität für einen Austausch mit Dateien und Bildern
- Anbindung an eigene Skripte möglich, zum Beispiel über Flowise oder n8n
- **Fazit**: Für den Austausch mit einer einzelnen Datei schnell und super. Für Datenbanken ist eine Anbindung über Skripte erforderlich und daher nicht mehr benutzerfreundlich.

Benutzeroberfläche der Ollama-Anwendung, https://ollama.com/

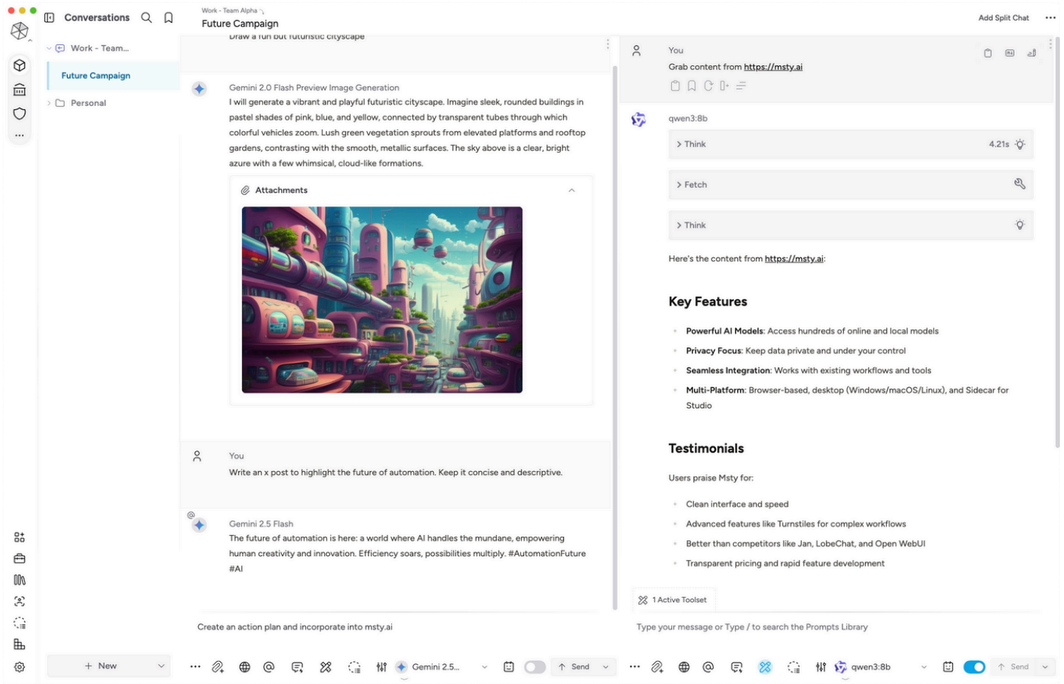
5. Beste Flexibilität: Local AI

- MIT-Lizenz, Open Source, kostenlos und kommerziell nutzbar
- Kompatible REST-API und WebUI für einfache Bedienung
- RAG lässt sich in der WebUI über verschiedene Datenbanken oder Dateien anbinden.
- **Fazit**: Robuste Lösung, aber zu komplex in der Einrichtung. *In diesem Fall kann man sich dann auch einfach einen eigenen Software-Stack bauen.*



WebUI von Local AI, https://localai.io/

6. Bester Kompromiss: Msty



Benutzeroberfläche von Msty, https://msty.ai/

- Proprietär, mit kostenfreier Privatversion, Premium/Unternehmensfeatures kostenpflichtig
- Intuitive, GUI-basierte Anwendung
- Funktionen wie Multiverse-Chats, Knowledge Stacks, Flowchat, Echtzeit-Websuche
- **Fazit**: Für den privaten Gebrauch hat Msty die größte Individualisierbarkeit mithilfe der Knowledge-Stacks-Funktionalität.

7. Beobachten

- **GPT4All** (https://www.nomic.ai/gpt4all) bietet eine Open-Source-Variante und in diesem Bereich die meisten Individualisierungsmöglichkeiten, wird aber zurzeit schon seit Januar 2025 nicht mehr weiterentwickelt; vielleicht ändert sich das in diesem Jahr noch
- **LM-Studio** (https://lmstudio.ai/) hat sich wesentlich weiterentwickelt und ist nach der Einrichtung mit ihrem neuen Plugin-System wahrscheinlich die flexibelste Lösung. Allerdings ist es weiterhin proprietär (obwohl es seit Juli auch kommerziell genutzt werden darf) und die RAG-Funktionen sind bisher nicht so zugänglich wie bei Msty.

Kontakt

Henry Herkula
BTU Cottbus-Senftenberg
Projekt KOMBiH

<henry.herkula@b-tu.de>
T: +49 (0)355 69 3728

Erich-Weinert-Straße 1
03046 Cottbus