Community of Practice KI*PerWeb*

Austausch zur Nutzung und Entwicklung KI-gestützter Webanwendungen

KIPERWEB

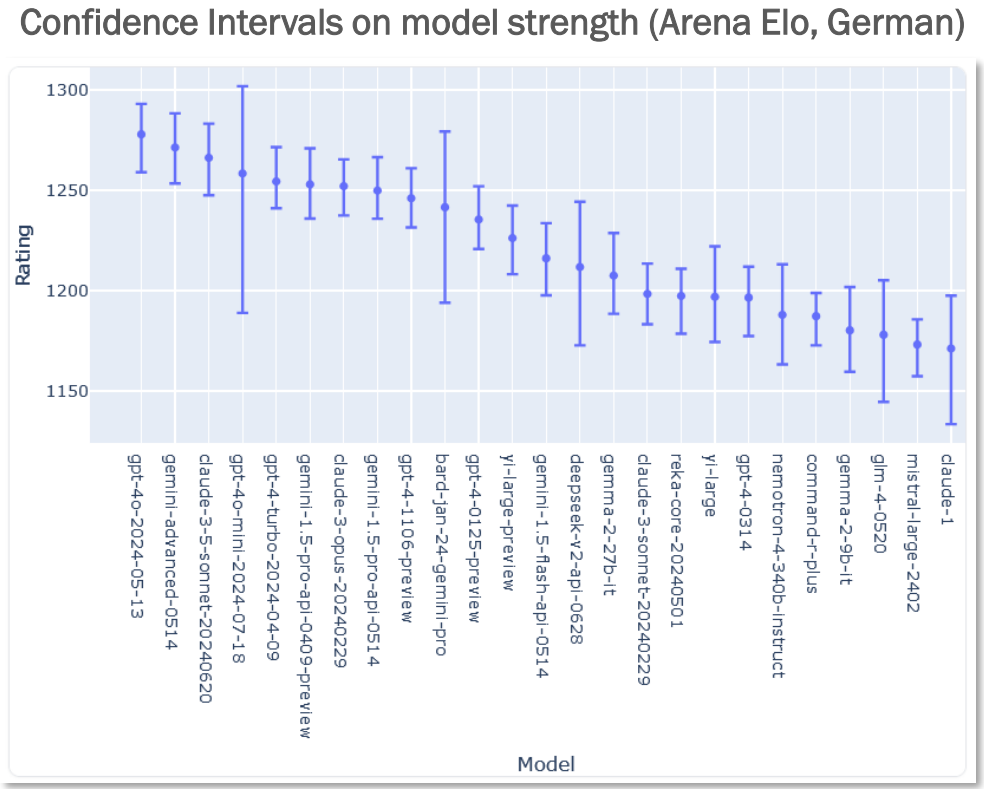
f-bb Forschungsinstitut Betriebliche Bildung

Agenda

- Update
  - News & Leaderboard-Update
- Input
  - „Flowise“ (Gastbeitrag von Henry Herkula)
- Diskussion

News & Update (24.07.2024)

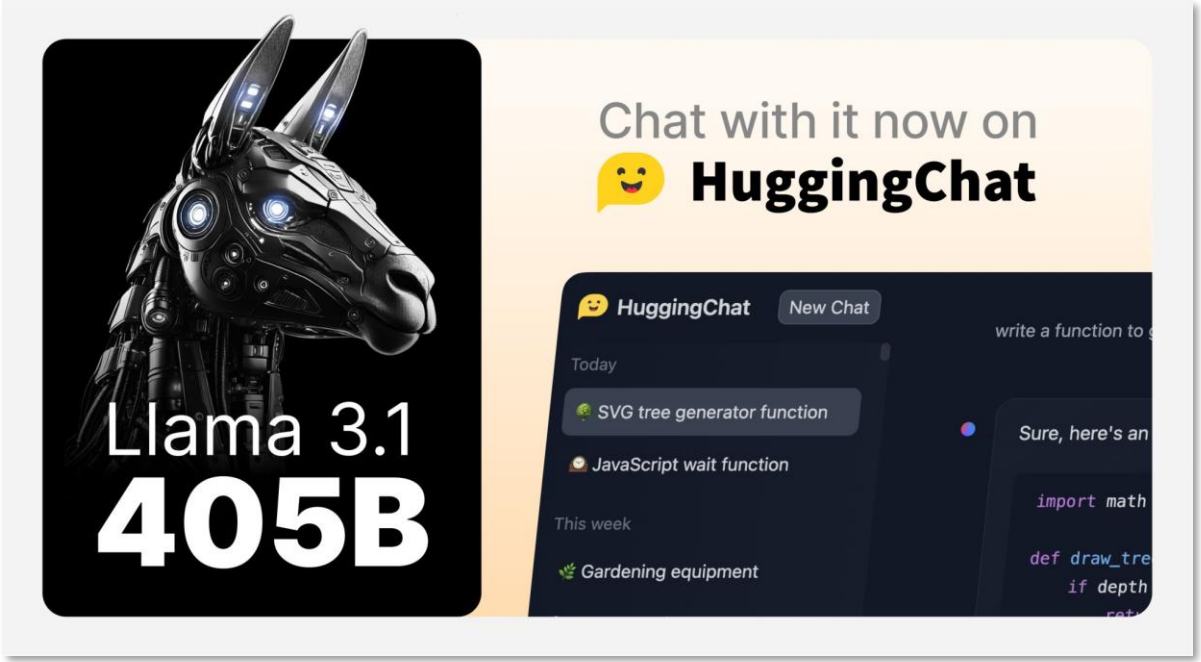
- *Gpt-4o-mini* steigt weit vorne mit ein
- *Gemini-1.5-flash-api-0514* & *deepseek v2-api-0628* vor *gemma-2-27b-it*
- *gemma-2-9b-it* auf einer Ebene mit *command-r-plus*
- *Mixtral-8x22b-Instruct-v0.1* – nicht mehr im Bild aber trotzdem hervorragend ☺ – bleibt vor *Mixtral-8x7b-Instruct-v0.1* bestes Modell unter Apache 2.0
- Neu veröffentlicht, aber noch nicht im Rennen:
  - *meta-llama/Meta-Llama-3.1-405B-Instruct*
  - *mistralai/Mistral-Nemo-Instruct-2407*
  - *nvidia/Mistral-NeMo-12B-Instruct*
  - *apple/DCLM-7B*
  - *nvidia/Minitron Family (4B & 8B)*



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

Exkurs: 🔥 Llama-3.1 🔥

- Auf den meisten Benchmarks noch vor GPT-4o und damit als erstes Open-Source LLM SOTA!
- Varianten 70B und 405B auf HuggingChat,
- GGUF-Varianten für 8B, 70B und 405B liegen ebenfalls vor: https://huggingface.co/models?search=llama-3.1%20gguf



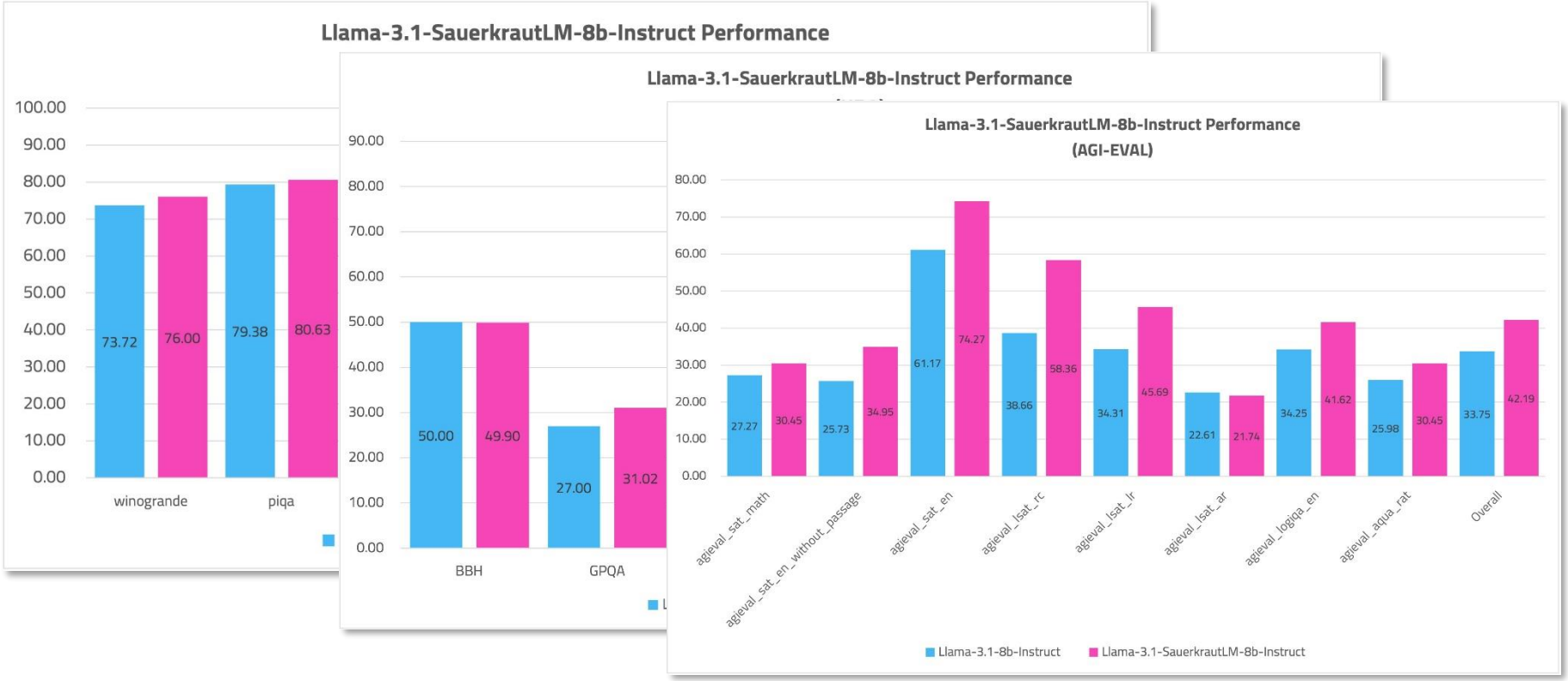
👉 https://huggingface.co/chat/models/meta-llama/Meta-Llama-3.1-405B-Instruct-FP8

Exkurs: Llama-3.1 ist neuer SOTA!

Category / Benchmark	Llama 3.1 405B	Nemotron 4 340B Instruct	GPT-4 (0125)	GPT-4 Omni	Claude 3.5 Sonnet
General MMLU (0-shot, CoT)	88.6	78.7 (non-CoT)	85.4	**88.7**	88.3
MMLU PRO (5-shot, CoT)	73.3	62.7	64.8	74.0	**77.0**
IFEval	**88.6**	85.1	84.3	85.6	88.0
Code HumanEval (0-shot)	89.0	73.2	86.6	90.2	**92.0**
MBPP EvalPlus (base) (0-shot)	88.6	72.8	83.6	87.8	**90.5**
Math GSM8K (8-shot, CoT)	**96.8**	92.3 (0-shot)	94.2	96.1	96.4 (0-shot)
MATH (0-shot, CoT)	73.8	41.1	64.5	**76.6**	71.1
Reasoning ARC Challenge (0-shot)	**96.9**	94.6	96.4	96.7	96.7
GPQA (0-shot, CoT)	51.1	-	41.4	53.6	**59.4**
Tool use BFCL	88.5	86.5	88.3	80.5	**90.2**
Nexus	**58.7**	-	50.3	56.1	45.7
Long context ZeroSCROLLS/QuALITY	**95.2**	-	**95.2**	90.5	90.5
InfiniteBench/En.MC	**83.4**	-	72.1	82.5	-
NIH/Multi-needle	98.1	-	**100.0**	**100.0**	90.8
Multilingual Multilingual MGSM (0-shot)	**91.6**	-	85.9	90.5	**91.6**

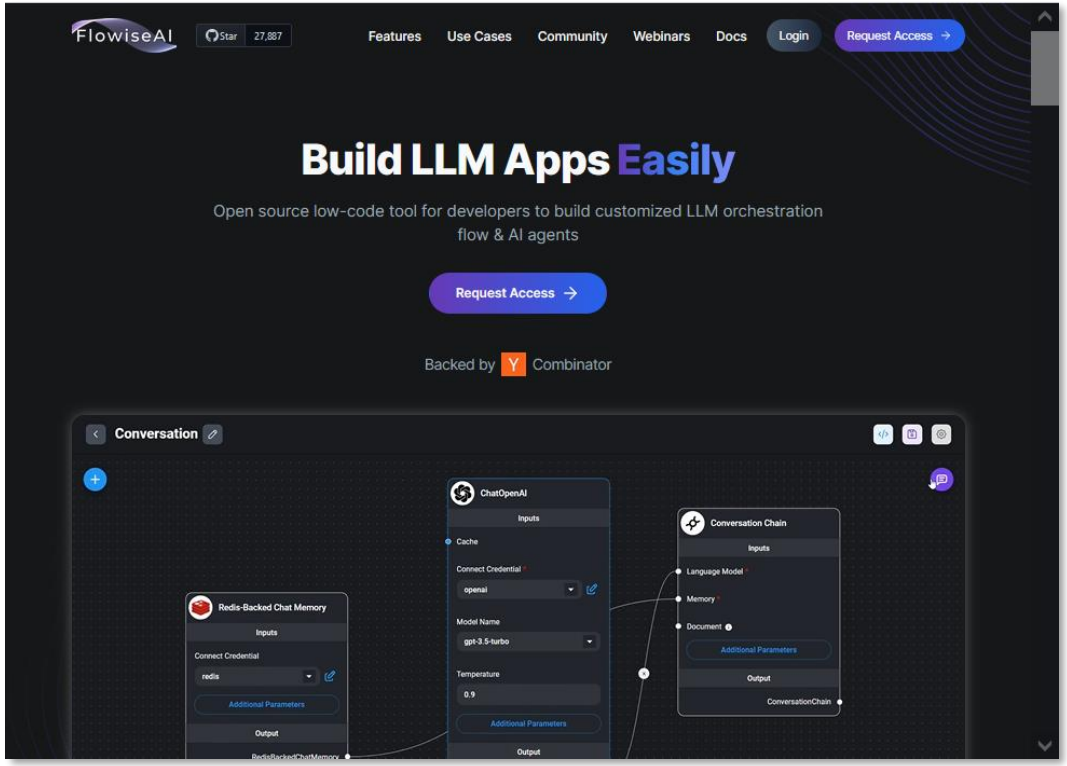
Quelle: https://ai.meta.com/blog/meta-llama-3-1/

Exkurs: Llama-3.1 gibt es jetzt auch mit Sauerkraut!



Quelle: https://x.com/AFischer1985/status/1816572063424958728

FlowiseAI



Quelle: https://flowiseai.com/;
cf. https://github.com/FlowiseAI/Flowise

Brandenburgische Technische Universität Cottbus - Senftenberg

Flowise

Open-Source-Werkzeug zur Verwaltung von KI-gesteuerten Arbeitsabläufen

Henry Herkula

Künstliche Intelligenz, Benutzeroberfläche, UX, Arbeitsabläufe, Verwaltung

KIPerWeb Community of Praxis

2024-07-26

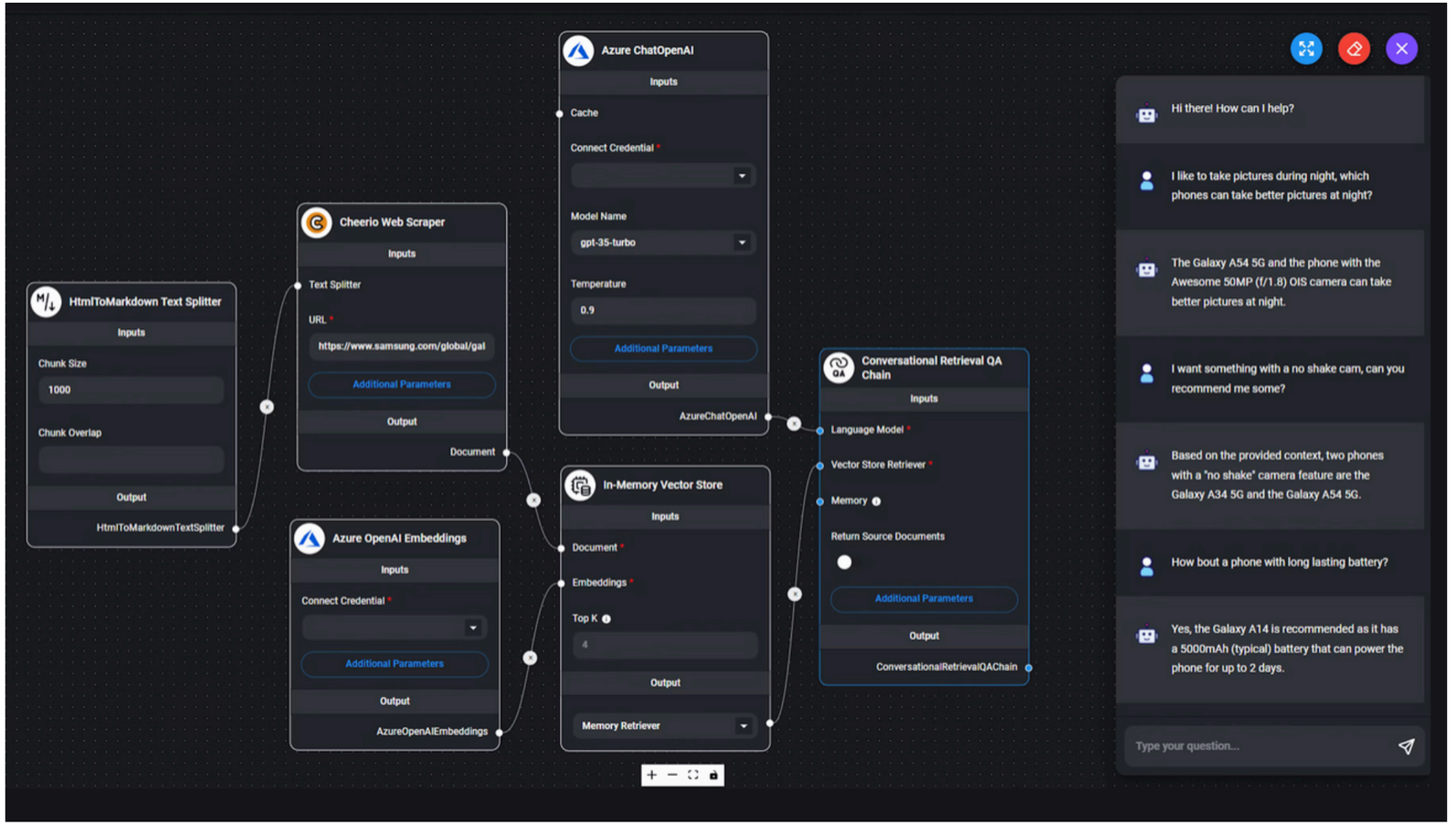
Aktualisiert: 2024-07-26

Zentrum für wissenschaftliche Weiterbildung

Brandenburgische Technische Universität Cottbus - Senftenberg

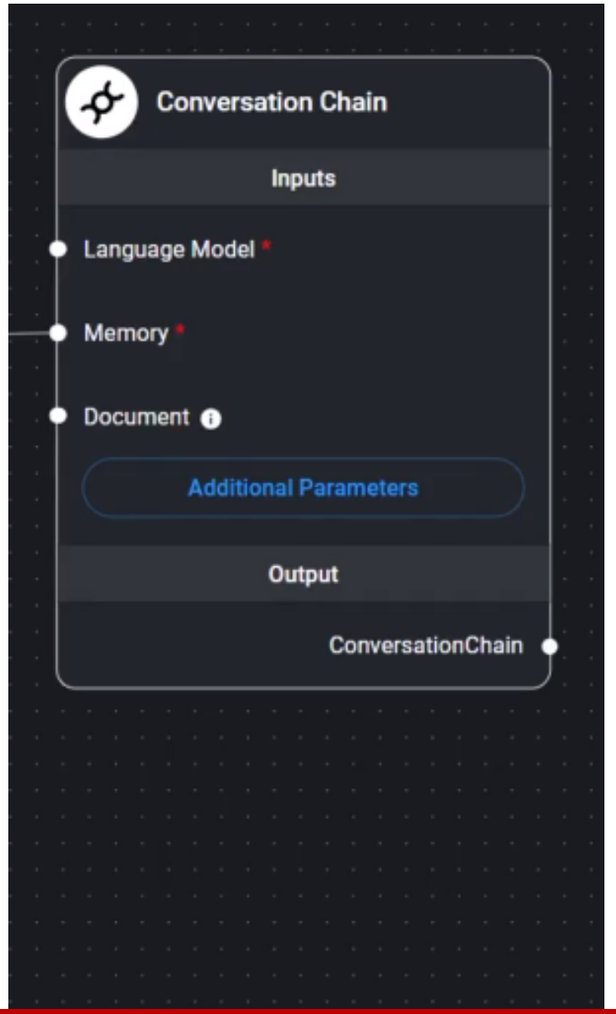
Inhaltsverzeichnis

1. Was ist Flowise?



1. Was ist Flowise?

- Flowise ist eine Benutzeroberfläche für die Zusammenstellung und Verwaltung von Arbeitsabläufen mit Künstlicher Intelligenz.
- Dafür werden Knoten, die verschiedene Aufgaben erfüllen, auf einem Arbeitsbereich verteilt und miteinander so verbunden, das sie eine Aufgabe erfüllen.



2. Was ist ein möglicher Mehrwert?

- Schneller Austausch von Lösungen für Anwendungen in der Künstlichen Intelligenz
- Überblick und Verwaltung verschiedener Arbeitsabläufe eines KI-Systems
- Integration eines eigenen Monitoring-Systems für KI-Abläufe.

3. Wie zugänglich ist die Einrichtung?

Brandenburgische Technische Universität Cottbus - Senftenberg

- Flowise benötigt als Voraussetzung einen NodeJS-Server (>= 18.15.0)

```
# install nvm (Node Version Manager)
curl -o- https://raw.githubusercontent.com/nvm-sh/nvm/v0.39.7/install.sh |
bash

# download and install Node.js
nvm install 20

# install flowise
npm install -g flowise

# start Flowise
npx flowise start
```

3. Wie zugänglich ist die Einrichtung?

- Mit diesen Befehlen konnte ich das Programm über http://localhost:3000 starten und sofort loslegen. Das Programm speichert alles von selbst.
- Flowise bietet eine umfangreiche Gemeinschaft für Vorlagen, sodass man relativ schnell eine funktionierende KI gestalten kann.

4. Was gibt es für Hindernisse?

- Flowise’ Übersichten und Benutzeroberfläche sind sehr unhandlich und es ist eine neue proprietäre Sprache von Knotenpunkten, die man lernen muss, um etwas sinnvoll zu erzeugen.
- Flowise bietet keine eigenen Anzeigen für das Durchlaufen der Prozesse, sondern man müsste dann auf Werkzeuge wie LangSmith zugreifen

5. Für wen ist Flowise gedacht?

- Flowise ist hervorragend, wenn man KI-Prozesse lernen möchte.
- Flowise lässt sich einsetzen, wenn man die Oberfläche in eigene Anwendungen einbaut, um dadurch eine Übersicht aller Prozesse zu erhalten, die man anbieten möchte.
- Flowise besitzt für mich jedoch keinen guten Mehrwert zu einer einzelnen Skript-Datei in Python.
- Alles in einer Text-Datei zur Verfügung zu stellen, gibt einem die Sicherheit, den gesamten Ablauf klar nachvollziehen zu können, während Flowise jedoch neue Komplexität einführt, die vor allem vielleicht für Monitoring-Situationen relevant wären.

Kontakt

Henry Herkula
BTU Cottbus-Senftenberg
Projekte EXPAND+ER WB³, KOMBiH
<henry.herkula@b-tu.de>
T: +49 (0)355 69 3728