



# Künstliche Intelligenz: Open Source KI in a Nutshell

Prof. Dr. Marcel Gygli

# Über mich

- Professor (Tenure Track) an der BFH
  - KI im Öffentlichen Sektor
  - Leiter Digital Sustainability Lab
  
- Innovation Fellow
  - Schweizerische Nationalbibliothek
  
- PhD in Computer Science



# Generative Künstliche Intelligenz

# Generative KI am Beispiel von Grossen Sprachmodellen



Was ist ein grosses Sprachmodell?



Ein großes Sprachmodell ist eine künstliche Intelligenz, die mithilfe von Texten und großen Datenmengen "gelernt" hat bestimmte Antworten zu geben. Daher ist es wichtig in welchem Kontext diese Antwort verstanden wird, da man mit der gewählten Eingabe auch das Problem verändern kann. Das größte bekannte Sprachmodell (Stand 2021) heißt GPT-3 (Generative Pre-training Transformer 3). Es wurde vom Unternehmen OpenAI entwickelt.

[<https://open-assistant.io>]

# Generative KI am Beispiel von Grossen Sprachmodellen



Was ist ein grosses Sprachmodell?



Ein großes Sprachmodell ist eine **künstliche Intelligenz**, die mithilfe von Texten und großen Datenmengen **"gelernt"** hat bestimmte Antworten zu geben. Daher ist es wichtig in **welchem Kontext diese Antwort verstanden** wird, da man mit der gewählten Eingabe auch das Problem verändern kann. Das größte bekannte Sprachmodell (Stand 2021) heißt GPT-3 (Generative Pre-training Transformer 3). Es wurde vom Unternehmen OpenAI entwickelt.

[<https://open-assistant.io>]

# Wie werden Grosse Sprachmodelle trainiert?

Es war einmal eine alte Geis ...

# Wie werden Grosse Sprachmodelle trainiert?

Es war einmal ...

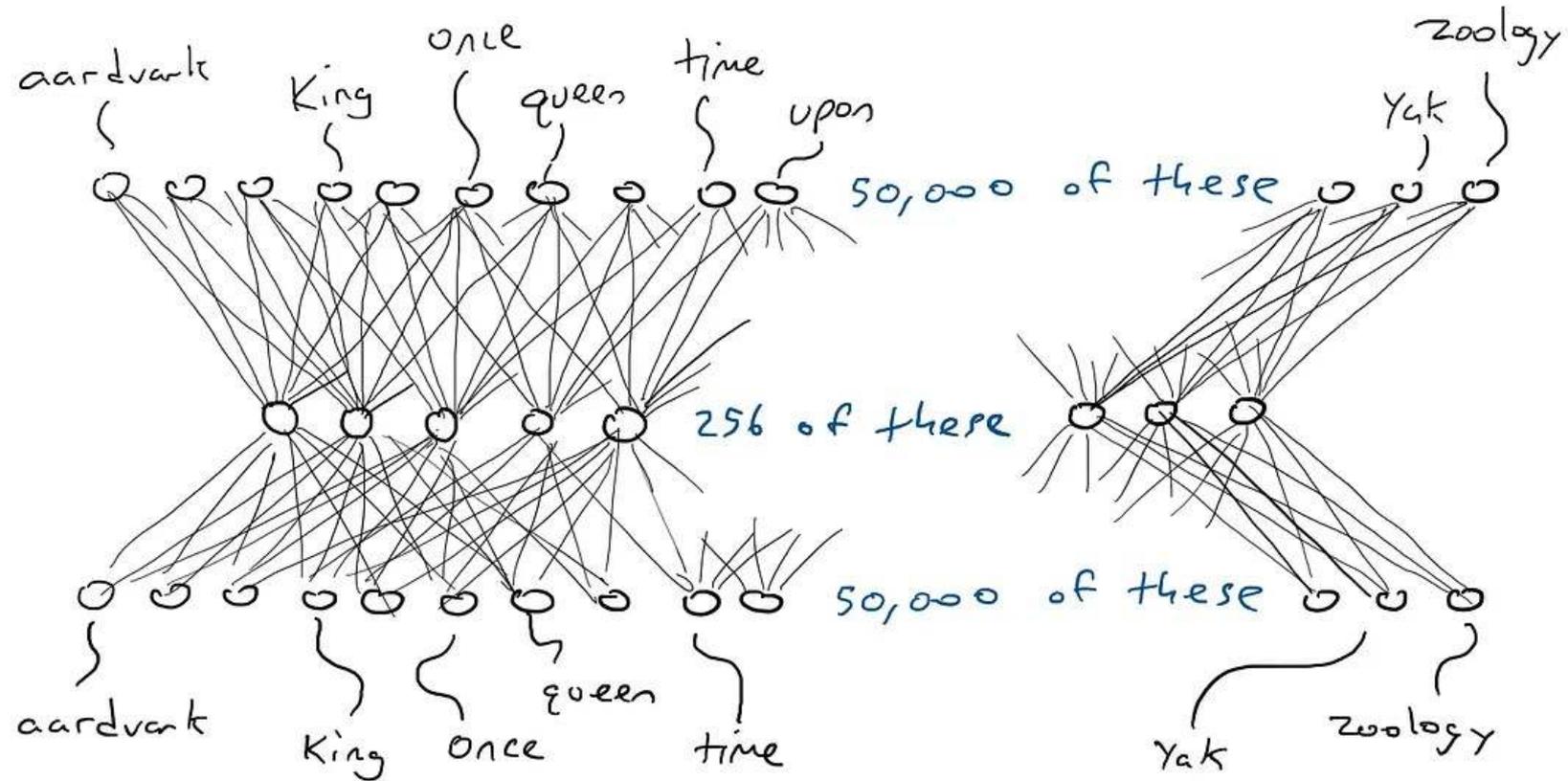
ein

eine

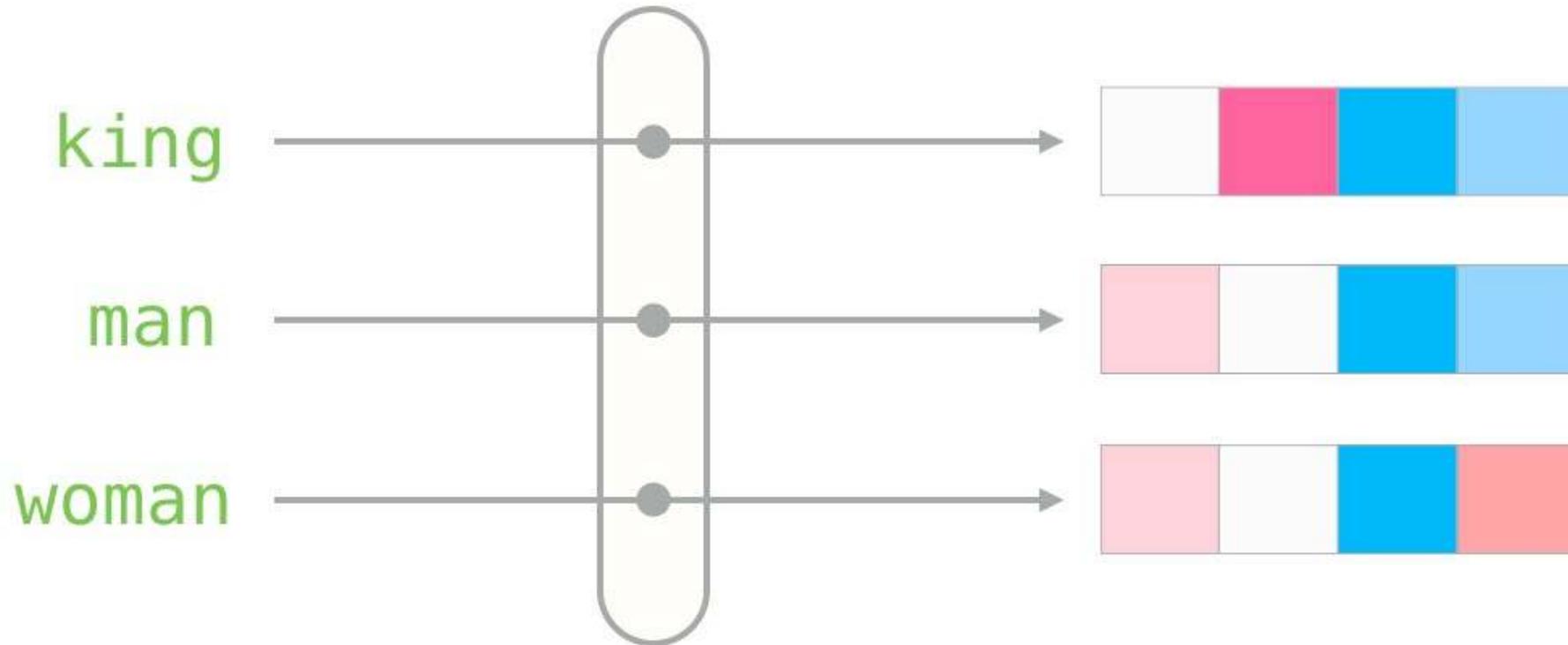
Kuchen

Apfel

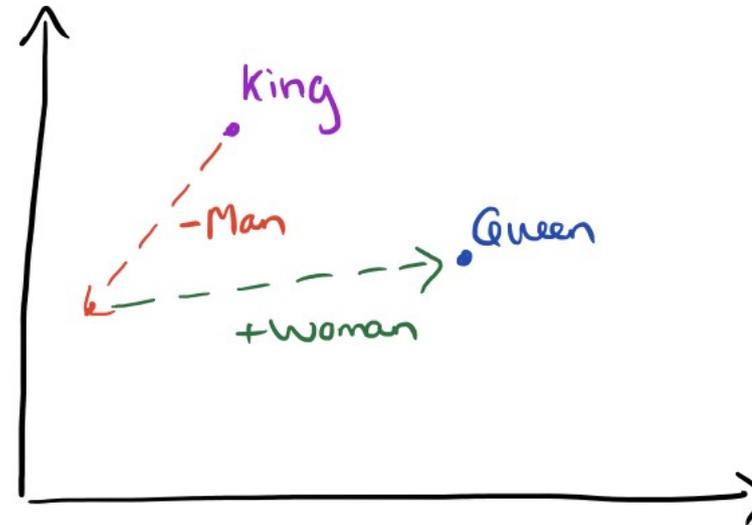
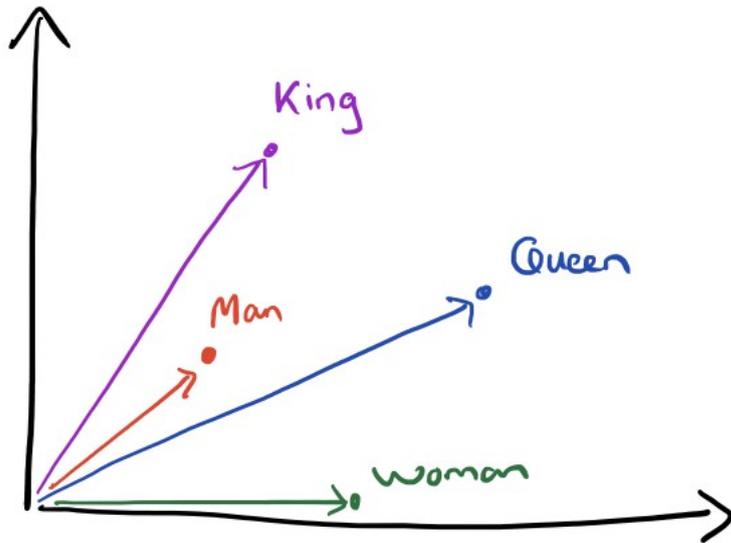
# Wir benötigen eine Zwischenrepräsentation



# Word-Embeddings als Rapresentation



# «Rechnen» mit Embeddings



# Open Source Künstliche Intelligenz

# Was sind «offene» Sprachmodelle

- Sprachmodelle von denen wir unterschiedliche Dinge kennen
  - Architektur des Modells
  - Trainingsdaten
  - Weights (Trainierte Knoten)
  
- Meisten Modelle bieten nur das letzte



# Was ist in diesen Daten eigentlich?

- Trainingsdaten bei allen bekannten Anbietern unbekannt
  - Unklare Biases in den Daten
  
- Grosser manueller Aufwand für «Sicherheit»
  - Getätigt im globalen Süden

TIME

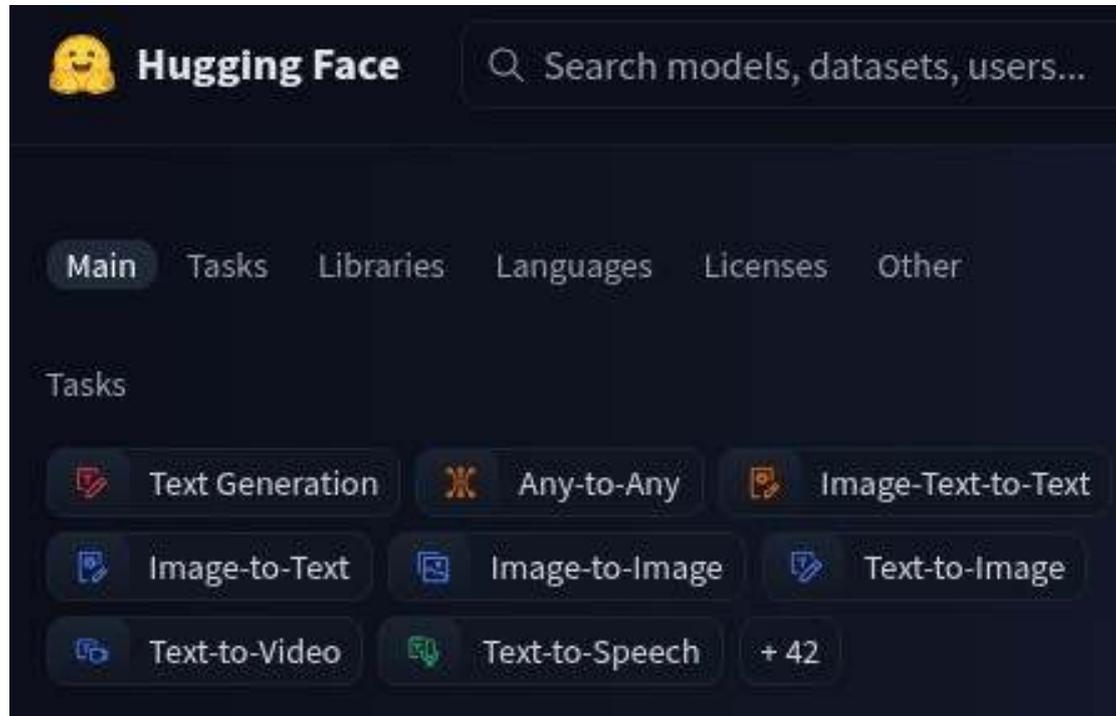
SUB

BUSINESS • TECHNOLOGY

**Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic**

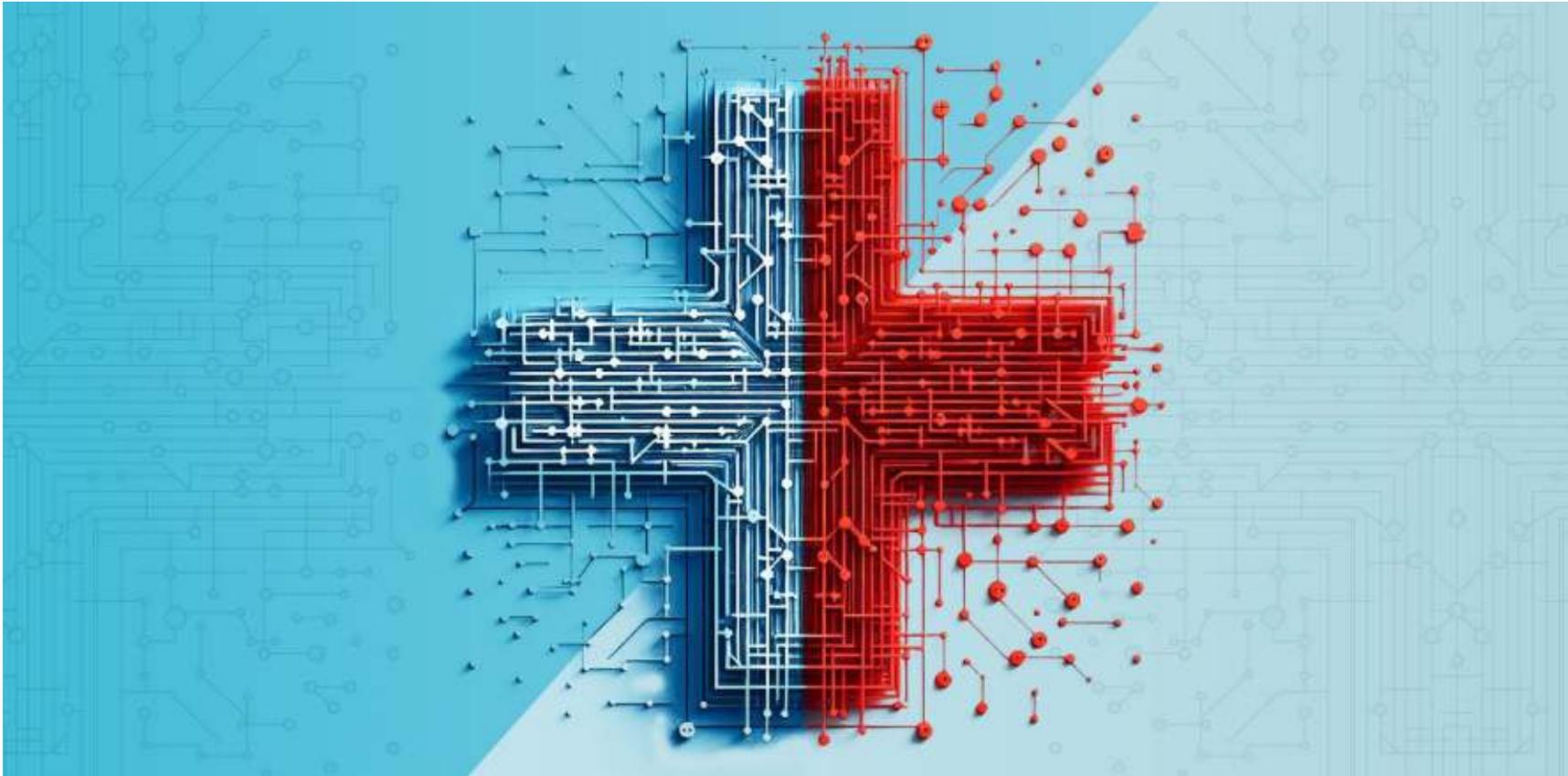
15 MINUTE READ

# Wo finde ich Open Source Modelle



# Apertus – DAS Schweizer Open Source LLM

➤ <https://huggingface.co/swiss-ai/Apertus-8B-Instruct-2509>



# Apertus – DAS Schweizer Open Source LLM

- Architektur
  - Decoder-Only Transformer
  - xIELU Activation Function
  - Context Länge: 65k Tokens
- Trainingsdaten
  - Reproduzierbar: Alle Skripts öffentlich
  - Mehrsprachig: 1811 Sprachen
  - Compliant: Respektiert aktuelle robots.txt, PII entfernt, toxische Daten entfernt
- Gewichte
  - 8B Parameter: <https://huggingface.co/swiss-ai/Apertus-8B-Instruct-2509>
  - 70B Parameter: <https://huggingface.co/swiss-ai/Apertus-70B-Instruct-2509>

# Wie bringe ich OSS Modelle in die Anwendung

# OLLAMA

- Open Source Tool um LLMs lokal zu betreiben



- Einheitliche API gegen aussen

Get up and running with large language models.

- Für die Konsole

```
PS C:\Users\ggm7> ollama run qwen3:latest
>>> what can you help me with?
<think>
Okay, the user asked, "what can you help me with?" I need to provide a helpful and comprehensive response. Let me start by recalling my capabilities. I can assist with a wide range of tasks, so I should list them in a clear and organized way.

First, I should mention general assistance, like answering questions, providing information, and helping with tasks. Then, break it down into categories. For example, academic help, like explaining concepts or helping with homework. Then, creative tasks, such as writing stories or poems. Technical support could include coding help or troubleshooting. Language-related tasks like translation or grammar checks. Also, maybe include things like generating ideas, planning, or even casual conversation.

I need to make sure the response is friendly and inviting. Maybe start with a greeting and then list the areas. Use bullet points or sections for clarity. Also, remind the user they can ask for specific help. Avoid being too technical, keep it conversational. Let me check if I missed any categories. Oh, maybe include things like problem-solving or decision-making advice. Also, note that I can handle multiple languages. Alright, structure the response with an introduction, then list the areas, and end with an invitation to ask for help. Make sure it's concise but covers all key points. Let me put that together now.
</think>
```

# vllm

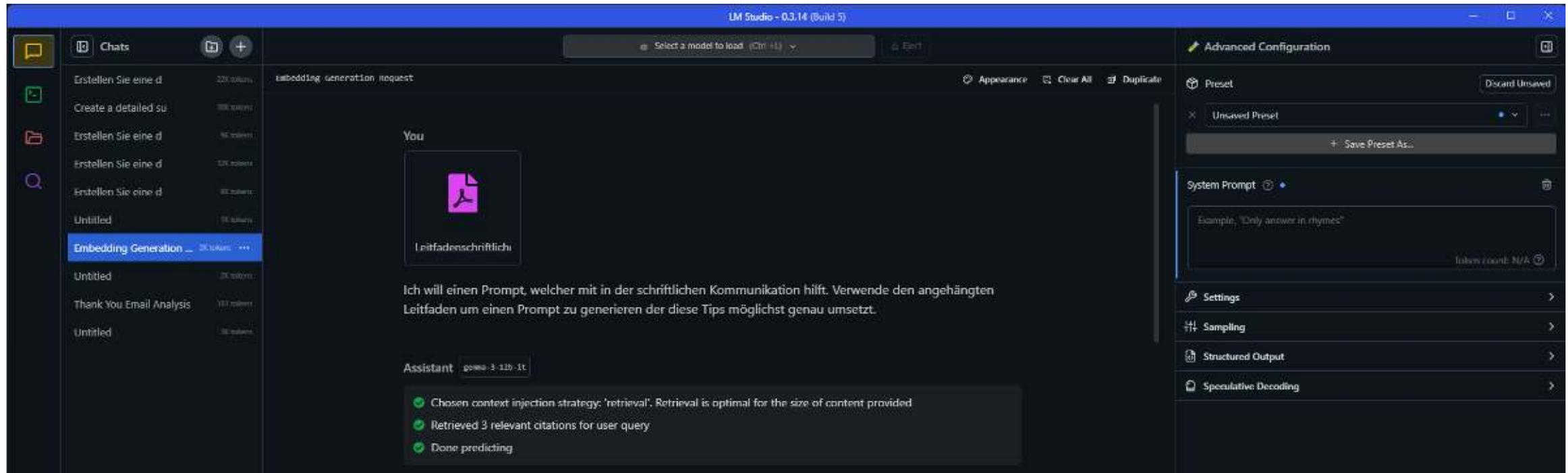
- Open Source Tool um LLMs lokal zu betreiben
- Einheitliche API gegen aussen
- Vor allem für Betrieb auf Servern

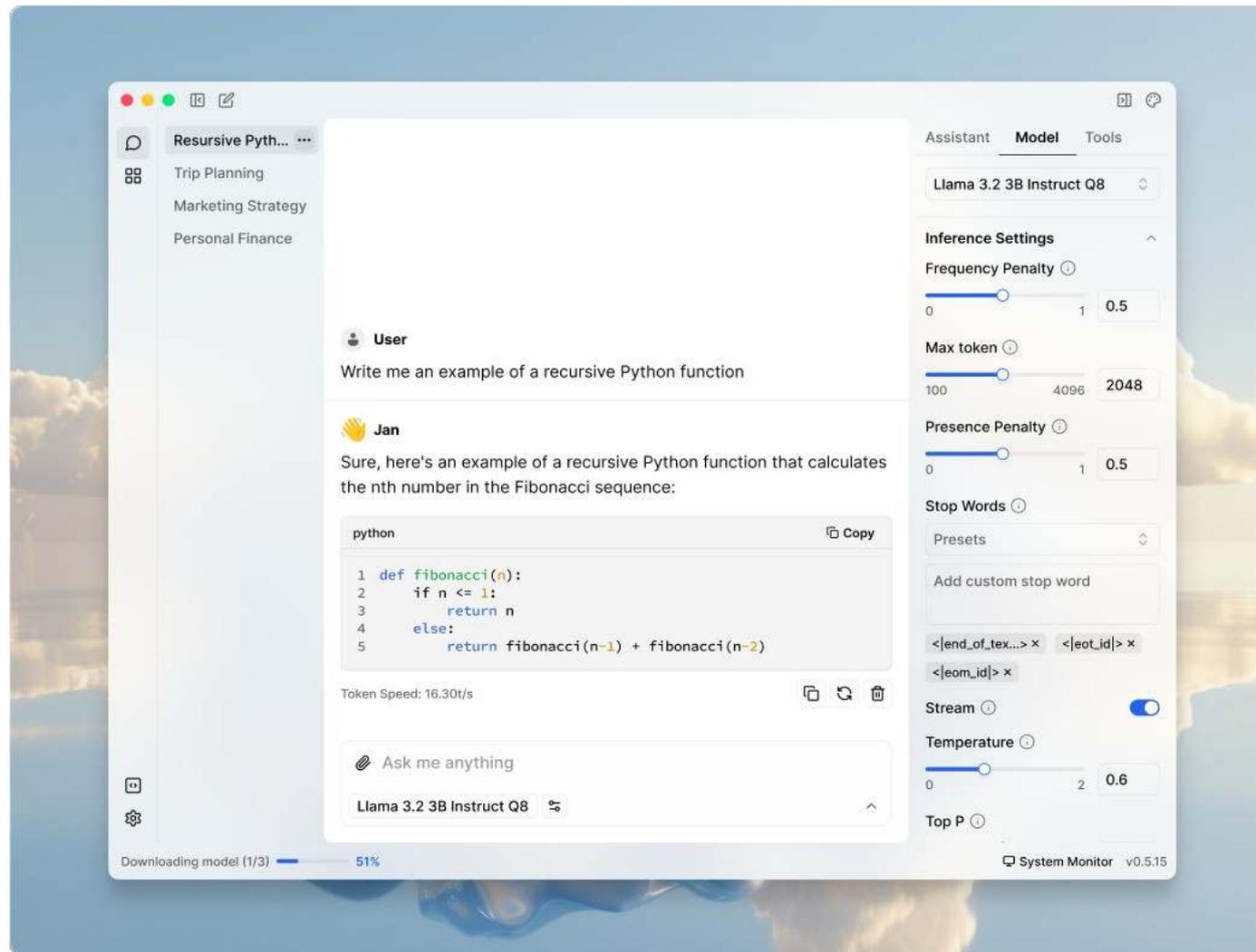


# Lokale Tools

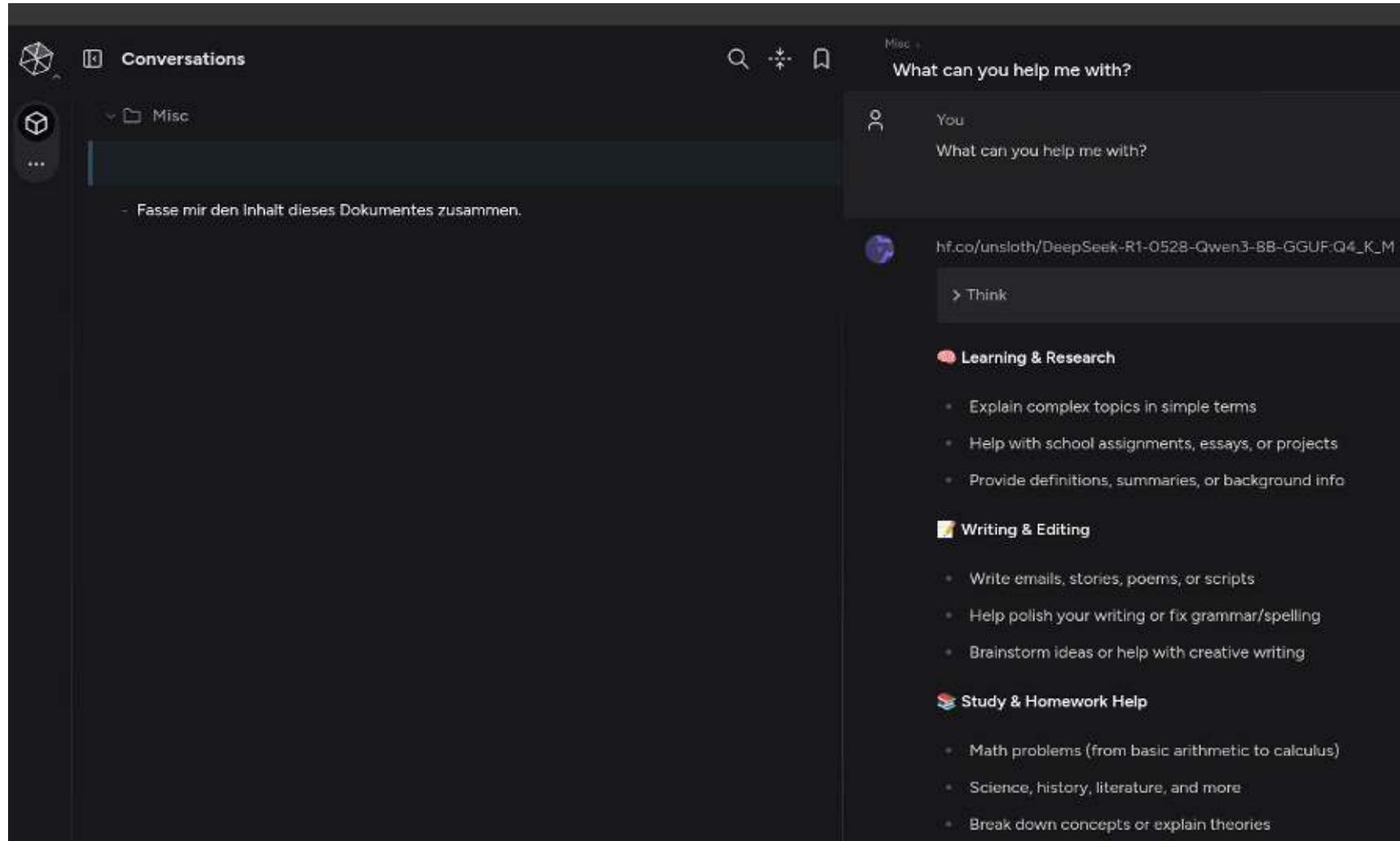
# LM Studio

## ▮ User Experience analog ChatGPT





# Msty



# Besten Dank für die Aufmerksamkeit

