

# RAGtime!

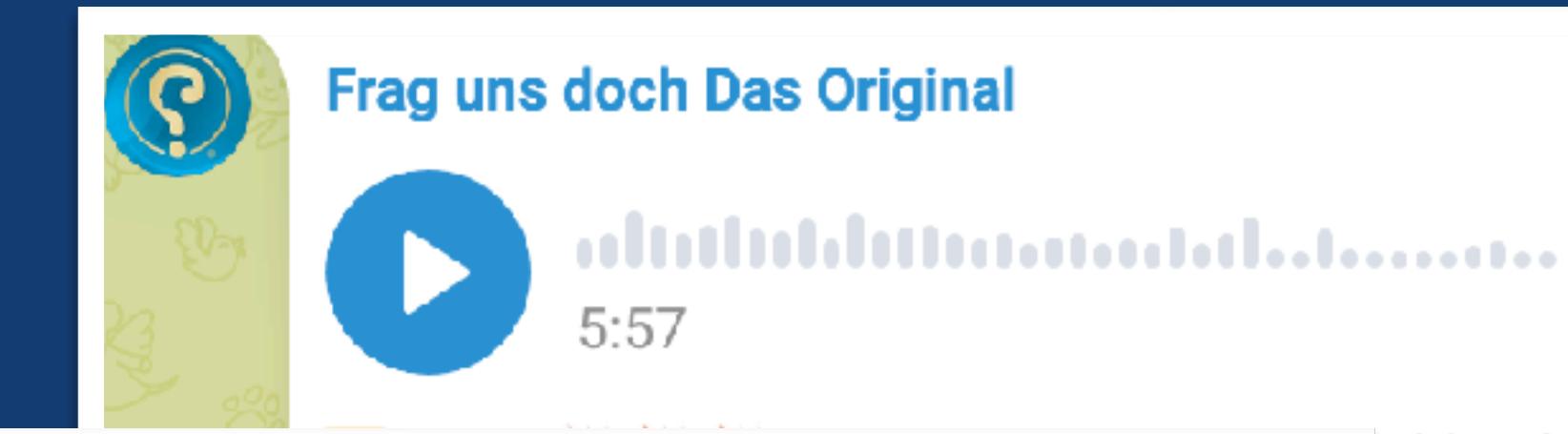
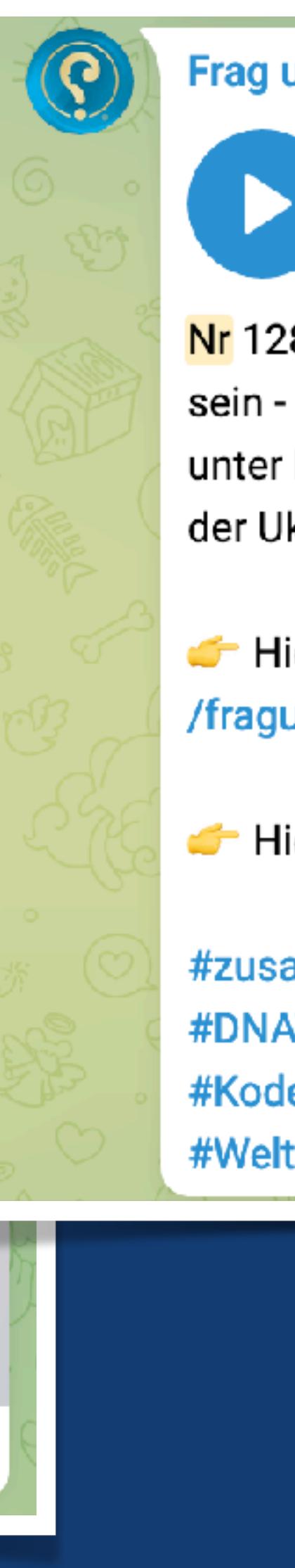
ChatGPT-Assistenten  
und ähnliche KI-  
Ratgeber zum  
Tanzen bringen



# Mein magischer KI-Moment



**hessenschau. 7. Dezember 2022**



schland muss verlieren  
Verschwörungstheorie

nal: <https://t.me>

Suchen  

 @fragunddoch4752 36.900 Abonnenten 93 Videos

Aufgrund der wiederkehrende Einschränkungen stellen wir hier keine weite... >

[ÜBERSICHT](#) [VIDEOS](#) [PLAYLISTS](#) [COMMUNITY](#) [KANÄLE](#) [KANALINFO](#)  >

[Neueste](#) [Beliebt](#) [Älteste](#)



157-2 da ursprüngliches Video defekt ist  
44.193 Aufrufe • vor 2 Jahren



157 Aktuelles vom Tage- Dominion - J. Biden -Sidney Powell - Deep Sta...  
60.425 Aufrufe • vor 2 Jahren



154 Boom-Time beginnt  
53.015 Aufrufe • vor 2 Jahren



149 Boom 2 - Weshalb die aktuellen Coronamaßnahmen -Hintergründe  
31.965 Aufrufe • vor 2 Jahren



148 Boom - Wahlen in USA - Wahlbetrug - Assets  
39.897 Aufrufe • vor 2 Jahren



147 Aktuelles vom Tage - Lockdown -Widersprüche  
37.353 Aufrufe • vor 2 Jahren



145 Interview mit Claudia - Germanische Neue Heilung  
24.885 Aufrufe • vor 2 Jahren



143 Aktuelles vom Tage Irrsinn Corona - Assets #Zusammenstehe...  
36.724 Aufrufe • vor 2 Jahren

video\_liste\_annotiert .XLSX

Datei Bearbeiten Ansicht Einfügen Format Daten Tools Hilfe

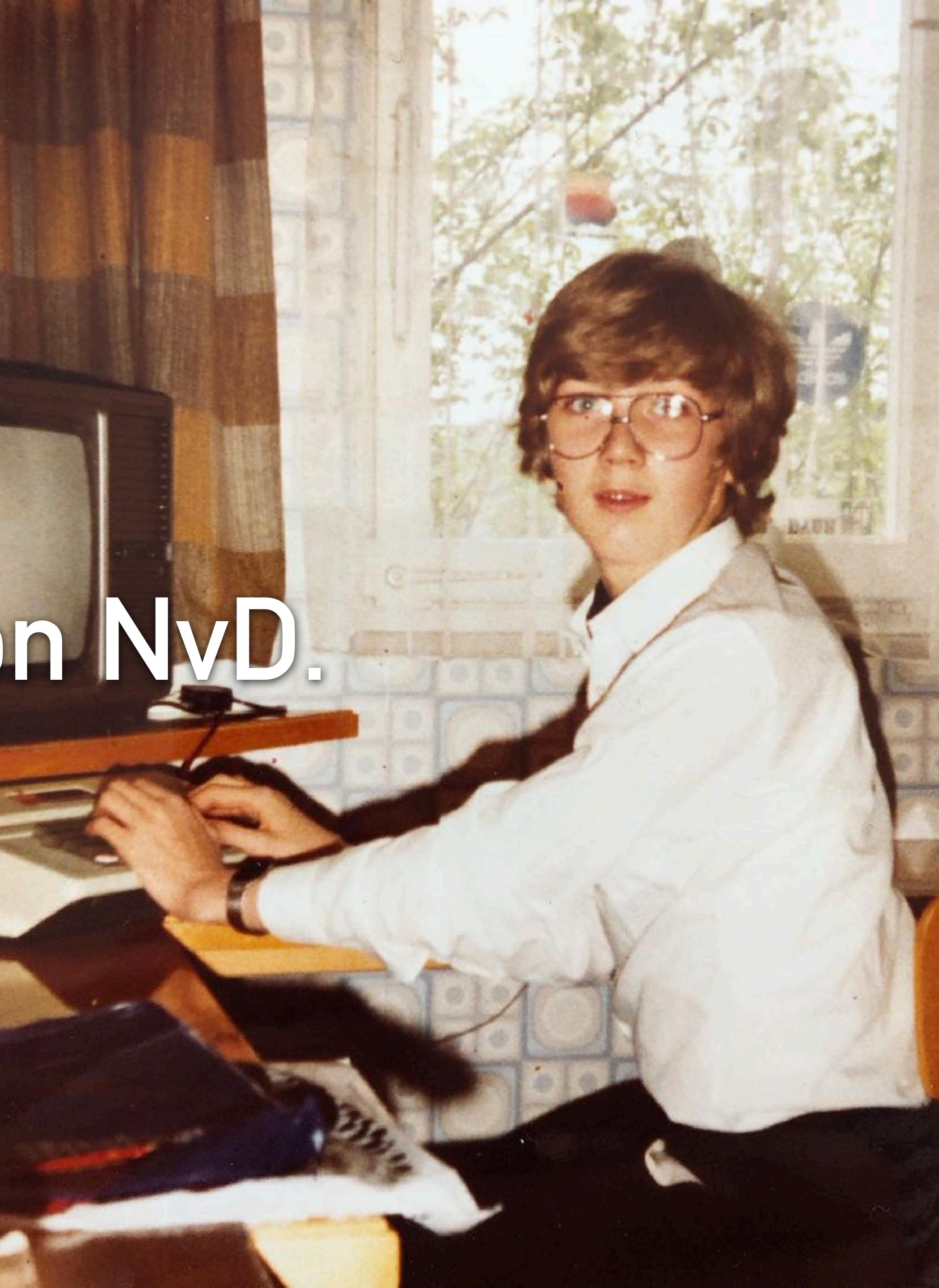
100% Stand... 10 B I A

K1 summary

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		<b>id</b>	<b>upload_date</b>	<b>description</b>	<b>duration</b>	<b>view_count</b>	<b>average_rating</b>	<b>age_limit</b>	<b>categories</b>	<b>tags</b>	<b>summary</b>		
92	C9WaF2 41u9Y	20200327	Wir sprechen über die These der Krise, von Corona als Trigger/Auslöser - dem Werkzeug Krypto-Weltwährung und dem großen Ziel für die Welt und deren Bevölkerung.  Hier geht es zu unserer Telegram-Gruppe: <a href="https://t.me/KanalFragunsdoch">https://t.me/KanalFragunsdoch</a>	2222	5312		0	['News & Politics']	[]		<ul style="list-style-type: none"> <li>Die beiden sprechen über die aktuelle Situation und die Zukunft.</li> <li>Die Deutsche Bank ist nicht in die Pleite gegangen, aber sie hat sich in einer existenzbedrohenden Krise befunden. Das hat das Potenzial, eine Krise auszulösen.</li> <li>Das Buch Cashkiller I ist immer noch eine gute Ausarbeitung und eine Pflichtlektüre für jeden zu lesen. Wir stellen nachher auch noch mal den Link ein, wo man das beziehen kann.</li> <li>Die Corona-Krise hat die Aktienkurse weltweit in einen Sinkflug versetzt. Die Börsen erholen sich gerade wieder, aber die Erholung ist fragil.</li> <li>Die Deutschen lieben Klopapier, die Franzosen Kondome und die Italiener Wein.</li> <li>Weltweit sind mittlerweile drei Milliarden Menschen von Ausgangsbeschränkungen betroffen. Das ist die Hälfte der Weltbevölkerung.</li> <li>Trump spricht von einem Deep State, der Pädophilie, Menschenhandel und Organhandel betreibt.</li> <li>Die katholische Kirche zahlt in Deutschland für sexualisierte Gewalt an Kindern und Jugendlichen zwischen 35.000 und 30.000 Euro.</li> <li>Die beiden Schauspieler Keanu Reeves und Mel Gibson haben in Interviews über den tiefen Sumpf in Hollywood gesprochen.</li> <li>Wer hart gesotten ist, soll sich mit dem, mit dem, mit dem, mit sich selbst beschäftigen. Gleichzeitig sind aber auch die korrupten Eliten dieser Welt und auch die Finanzelite dem Deep State zugeordnet, weil die einem jüdischen zionistischen Finanzsystem sich unterworfen haben und dieses zelebrieren im wahrsten Sinne des Wortes, denn über Schulden regiert man die Bürger. Genau. Das haben wir auch schon in Cash Killer 1 ganz genau beschrieben. Wer die Schulden regiert, regiert die Bevölkerung oder die Bürger.</li> <li>Trump hat in seiner Eröffnungsrede zum Weltwirtschaftsforum in Davos</li> </ul>		

Colab-Notebook: <https://github.com/JanEggers-hr/youtube-scrapers>

Nerdalarm: Immer schon NvD.



# Wo ihr heute durchmüssst:

- Sprachmodell-Technologie verstehen
- Den OpenAI-Playground nutzen
- Diesen wacken Kommandozeilenkram einsetzen
- Lokale Sprachmodelle installieren
- Docker-Container starten



# Was ihr davon habt:

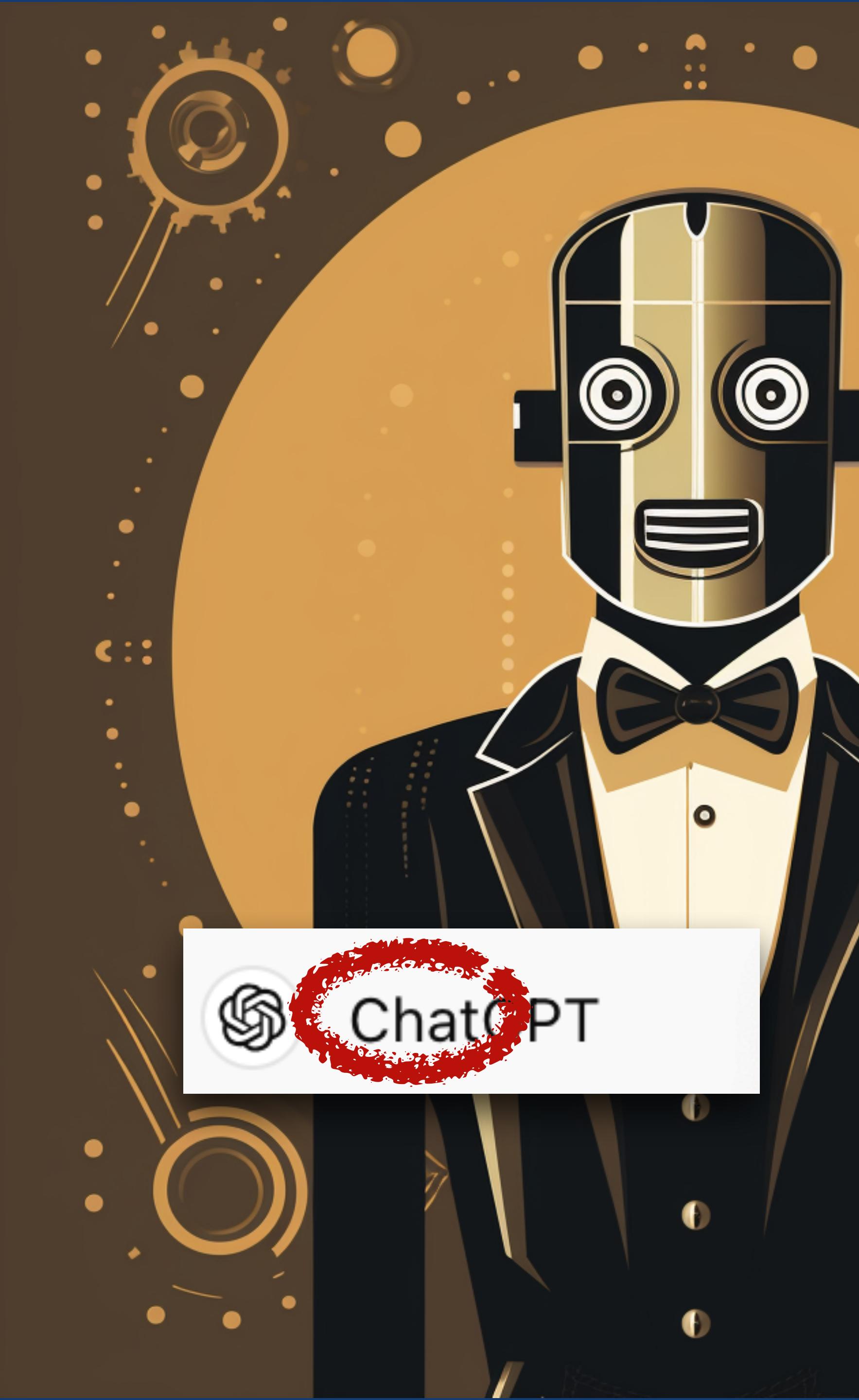
- Killer-Technik, um Recherchen im Griff zu behalten
- Dokumente durchsuchen
- Besser investigativ recherchieren
- Interviews trainieren
- Klügere Bots bauen
- Fakten checken



# Wie redet der eigentlich mit uns?



ChatGPT





Ich  
komme  
heute

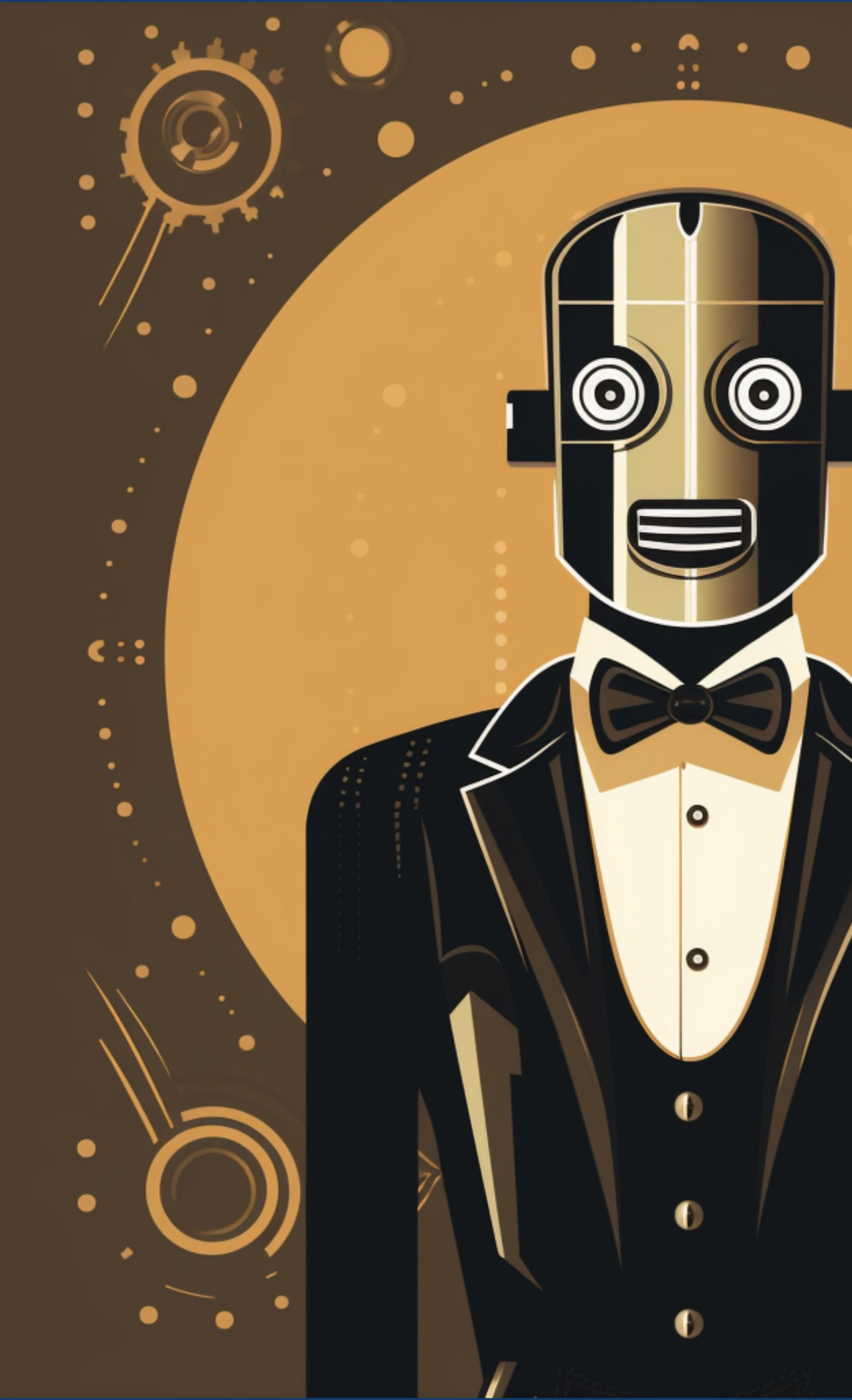
01FA77  
80C3AB  
621181

Abend

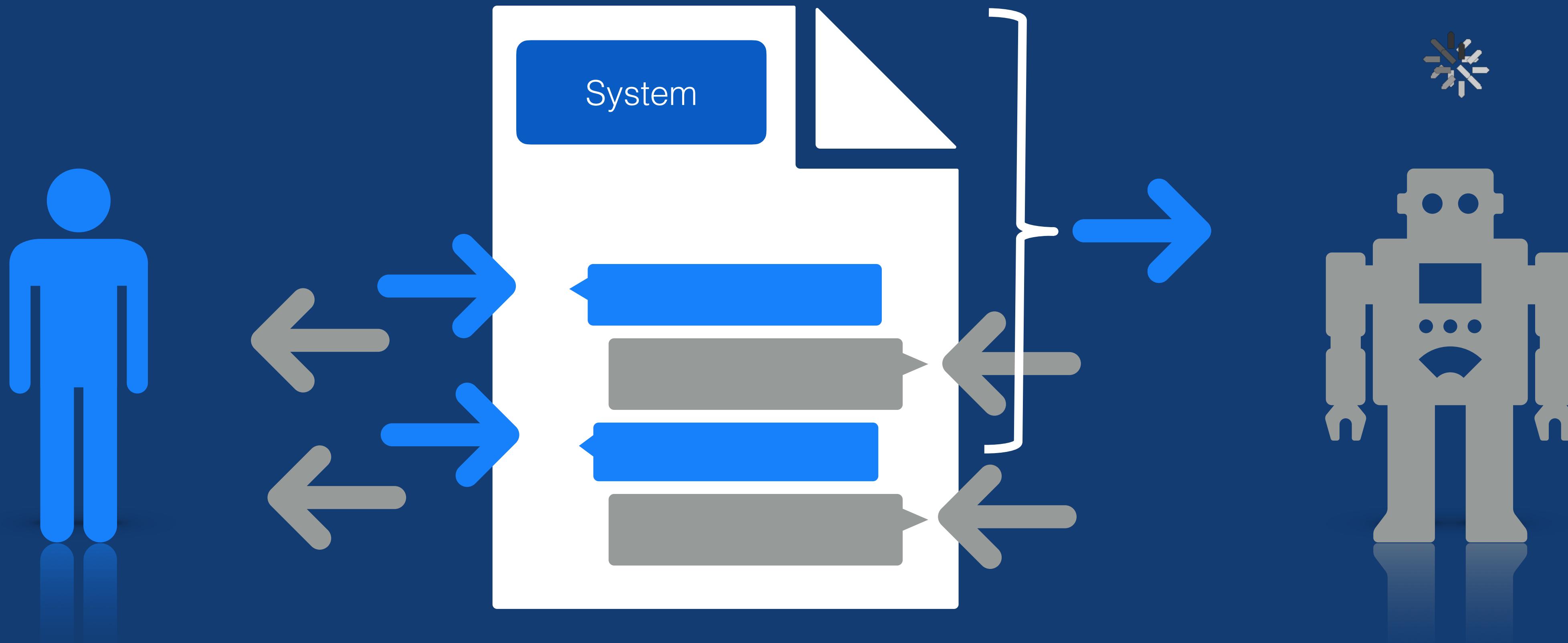
über  
mit  
nicht  
später

# Wie kann die Maschine eigentlich mit uns chatten?

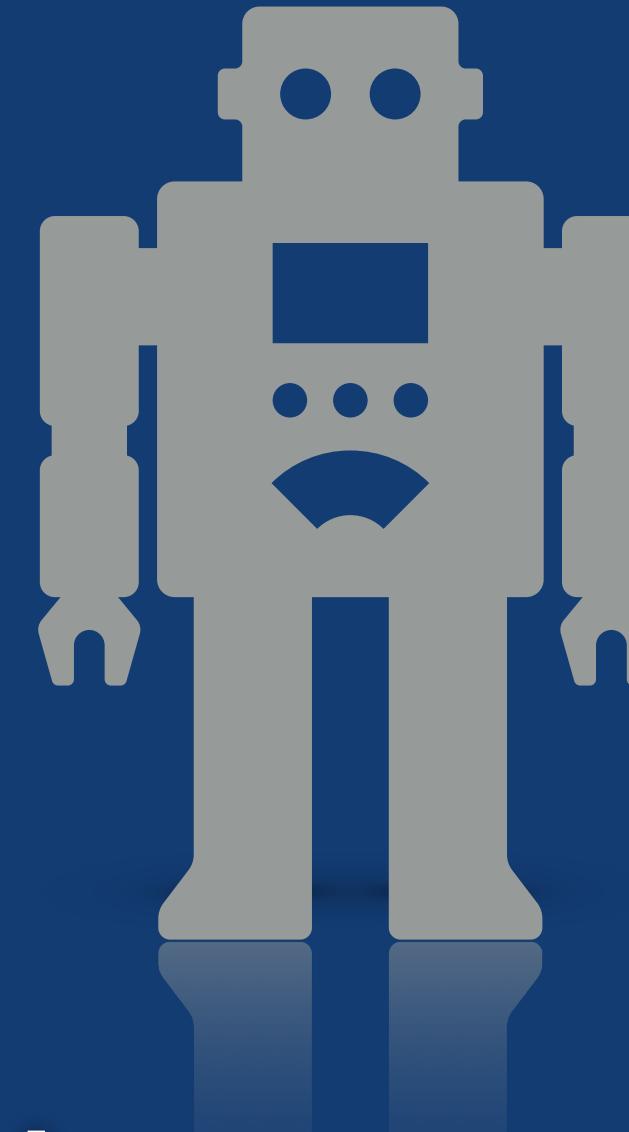
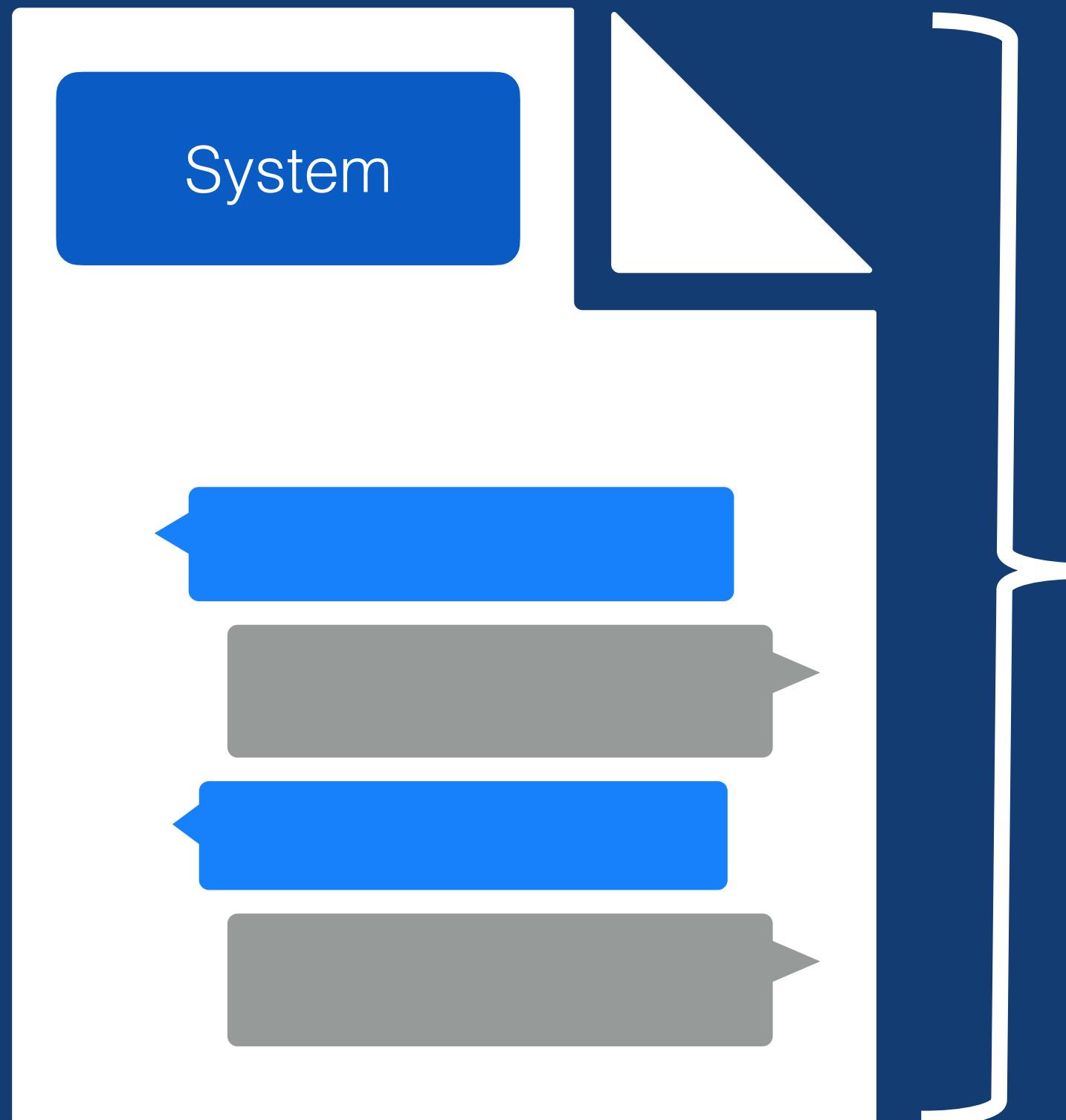
- A. Die KI hat eine Art Empathie-Modell - eine Abbildung der inneren Zustände des Gegenübers
- B. Eine synthetische Persönlichkeit generiert passende Antworten, von denen eine mehr oder weniger zufällig ausgewählt wird
- C. Sie füllt in Wirklichkeit ein Formular aus, auf dem eine Textergänzung eines Dialogformats beantragt wird
- D. Die Maschine ist darauf trainiert, einen Sinn aus unseren Fragen zu extrahieren, und nutzt dann dazu passende Textbausteine



# Der Chat-Trick: Die Maschine vervollständigt!



# LLM können nur beantworten, was auch auf dem Zettel steht.



## Kontextlänge

- GPT-3.5: 16k Token
- GPT-4: 128k Token (ca. 70 Seiten)
- Claude-3: 200k Token
- Gemini-1.5: bis 1 Mio. Token

# Anwendungsfall 1: Krisen-Assistent





Bundesamt  
für Bevölkerungsschutz  
und Katastrophenhilfe



Bundesamt  
für Bevölkerungsschutz  
und Katastrophenhilfe

Ratgeber für Notfallvorsorge und  
richtiges Handeln in Notsituationen

Ka  
tas  
tro  
phen



[https://www.bbk.bund.de/DE/Warnung-Vorsorge/warnung-vorsorge\\_node.html](https://www.bbk.bund.de/DE/Warnung-Vorsorge/warnung-vorsorge_node.html)

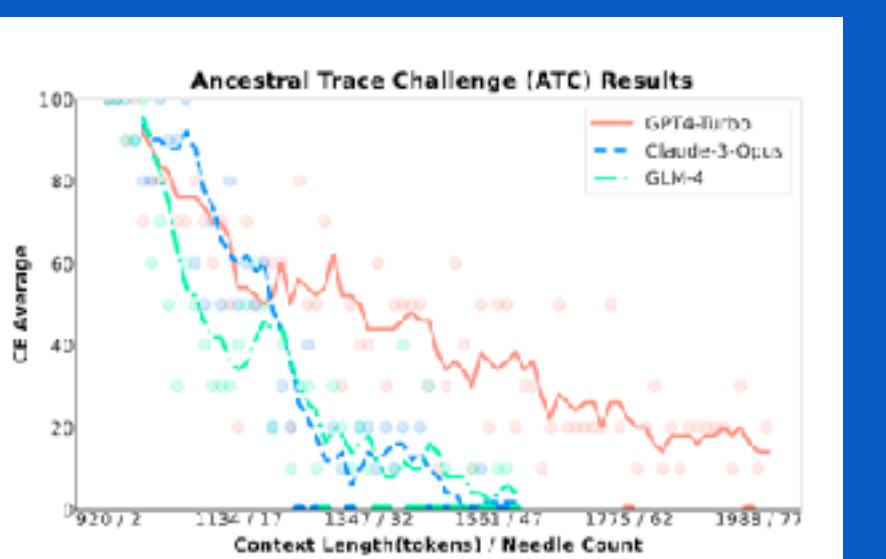
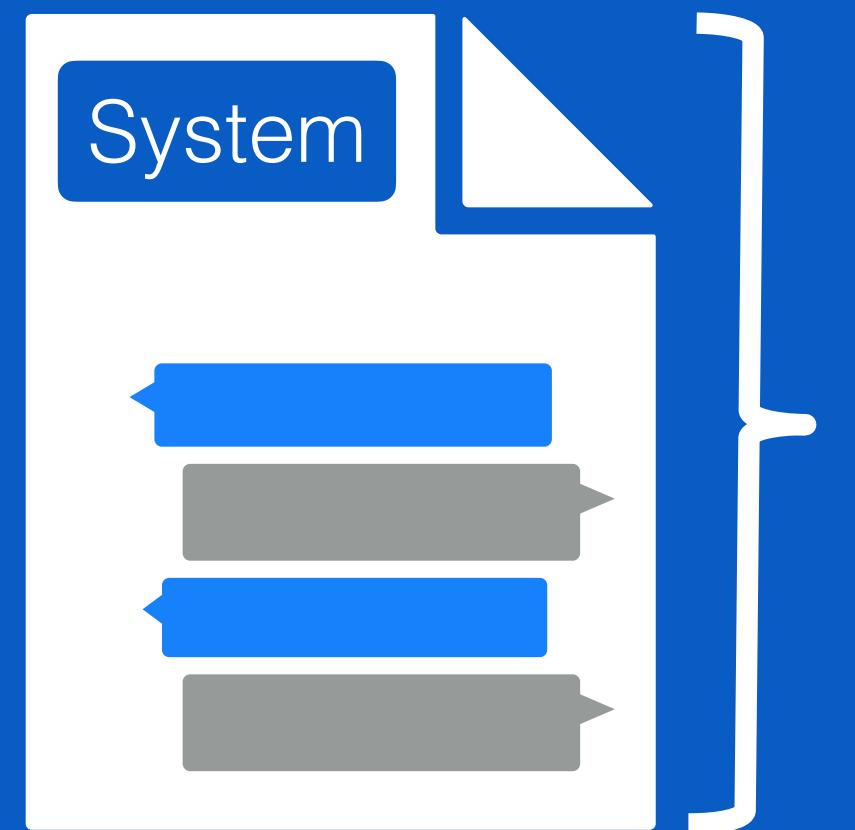
# Einfach bei ChatGPT reinkloppen?



The screenshot shows a web browser window with the URL <https://chatgpt.com>. The page title is "ChatGPT 4o". On the left, there is a sidebar with a list of GPT profiles: ChatGPT, Digitalisierungsbots, ARD-SEO-Redaktions..., ARD-SEO-Bibliothekar..., NDR Info Experte, and "3 mehr". Below this is a section titled "GPTs erkunden" with links to "Gestern" and "Vor...". To the right, there is a large "ChatGPT" icon and a list of four cards: "Einladung eines Freunde...", "Erstelle eine persönliche Website", "Trivia über das Römische Reich", and "Rezept mit Zutaten aus meiner Küche". At the bottom, there is a search bar with the placeholder "Sende eine Nachricht an ChatGPT" and a note: "ChatGPT kann Fehler machen. Überprüfe wichtige Informationen."

# ...keine gute Idee.

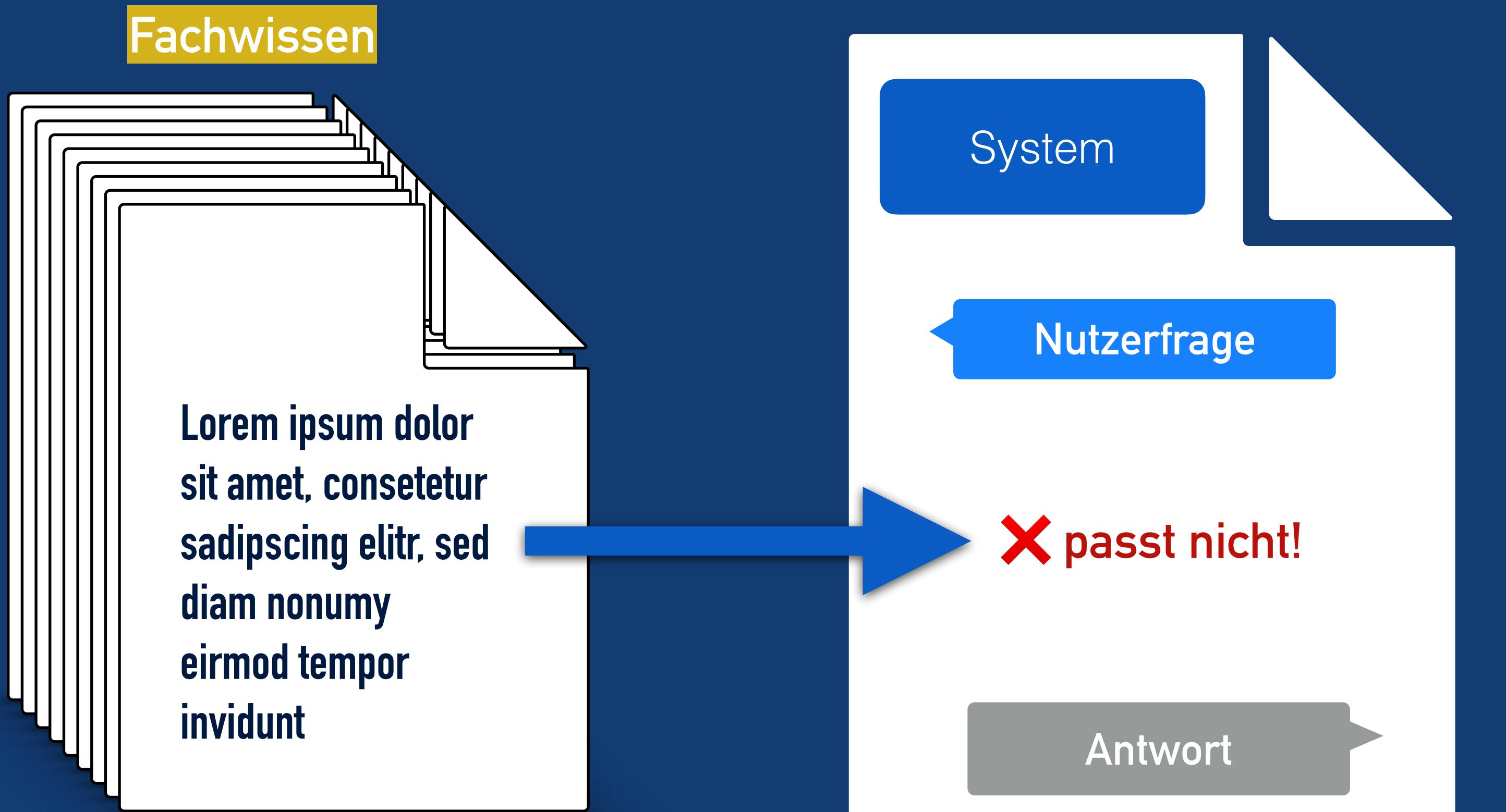
The screenshot shows the ChatGPT 4.0 interface. On the left, there's a sidebar with a list of recent interactions and a "Team-Arbeitsbereich hinzufügen" button. The main workspace has a large input field at the bottom with placeholder text "Sende eine Nachricht an ChatGPT". Above the input field, there are four cards with prompts: "Einladung eines Freundes zur Hochzeit", "Erstelle eine persönliche Website", "Trivia über das Römische Reich", and "Rezept mit Zutaten aus meiner Küche".



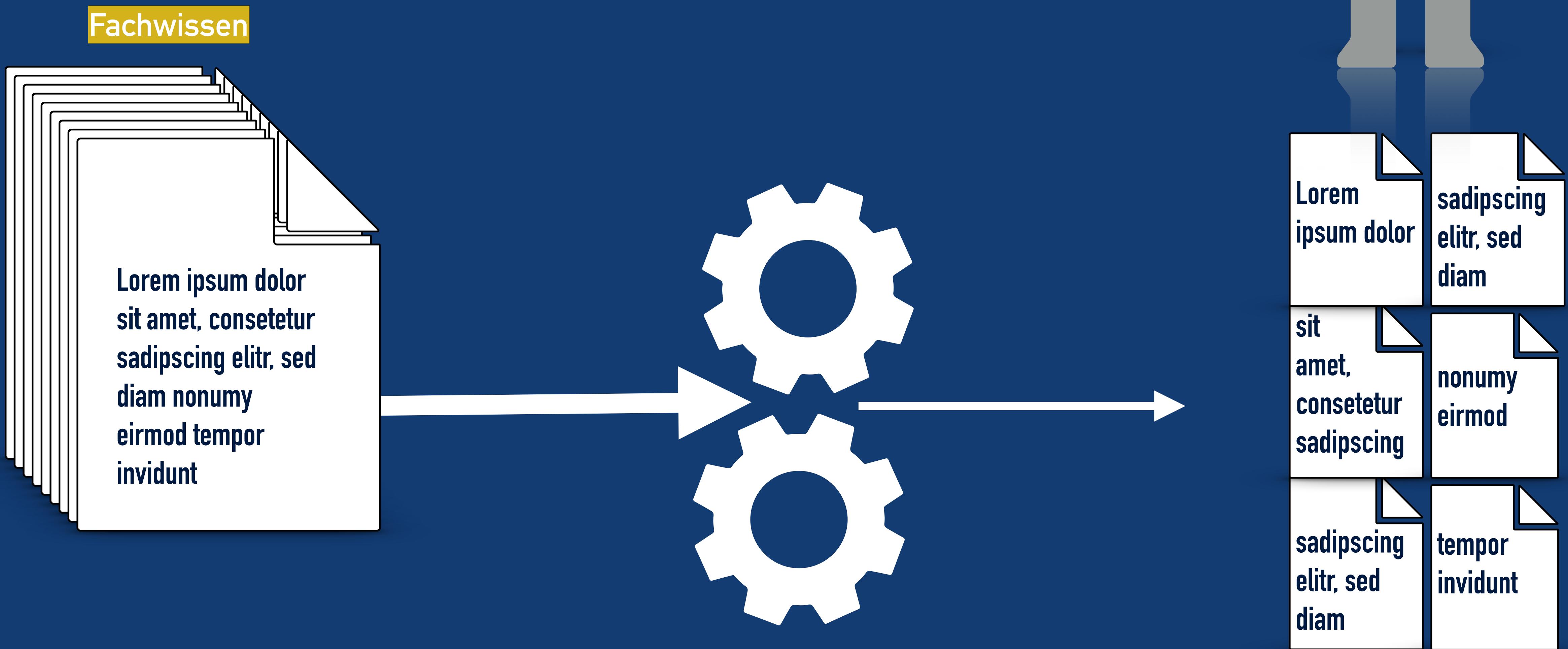
(Li, Zhang et al.:  
NeedleBench: Can LLMs Do Retrieval and  
Reasoning in 1 Million Context Window?  
<https://arxiv.org/abs/2407.11963>)

- Kontext reicht vermutlich nicht
- Lange Kontexte = schlechte Antworten

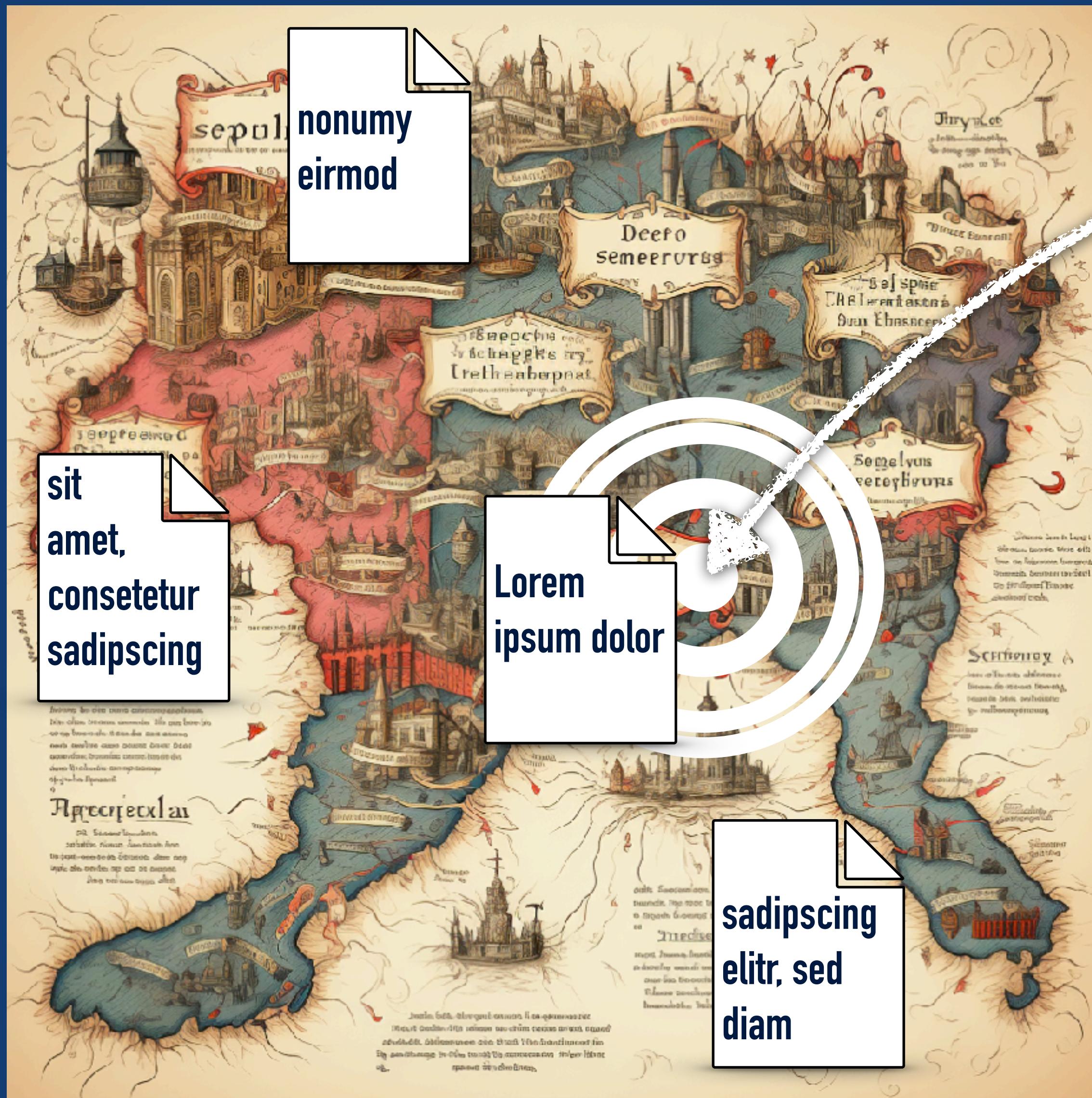
# „Retrieval-augmented Generation (RAG)“:



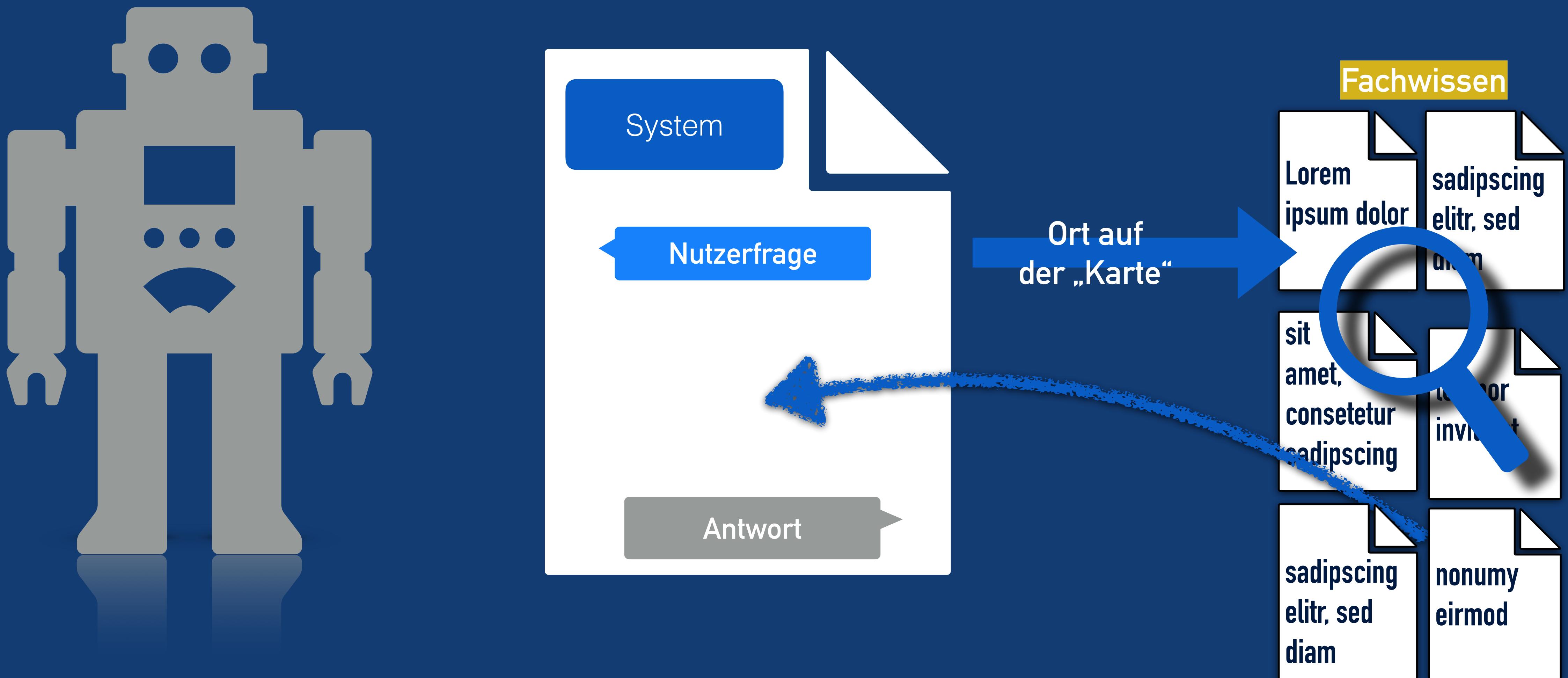
# „Retrieval-augmented Generation (RAG)“



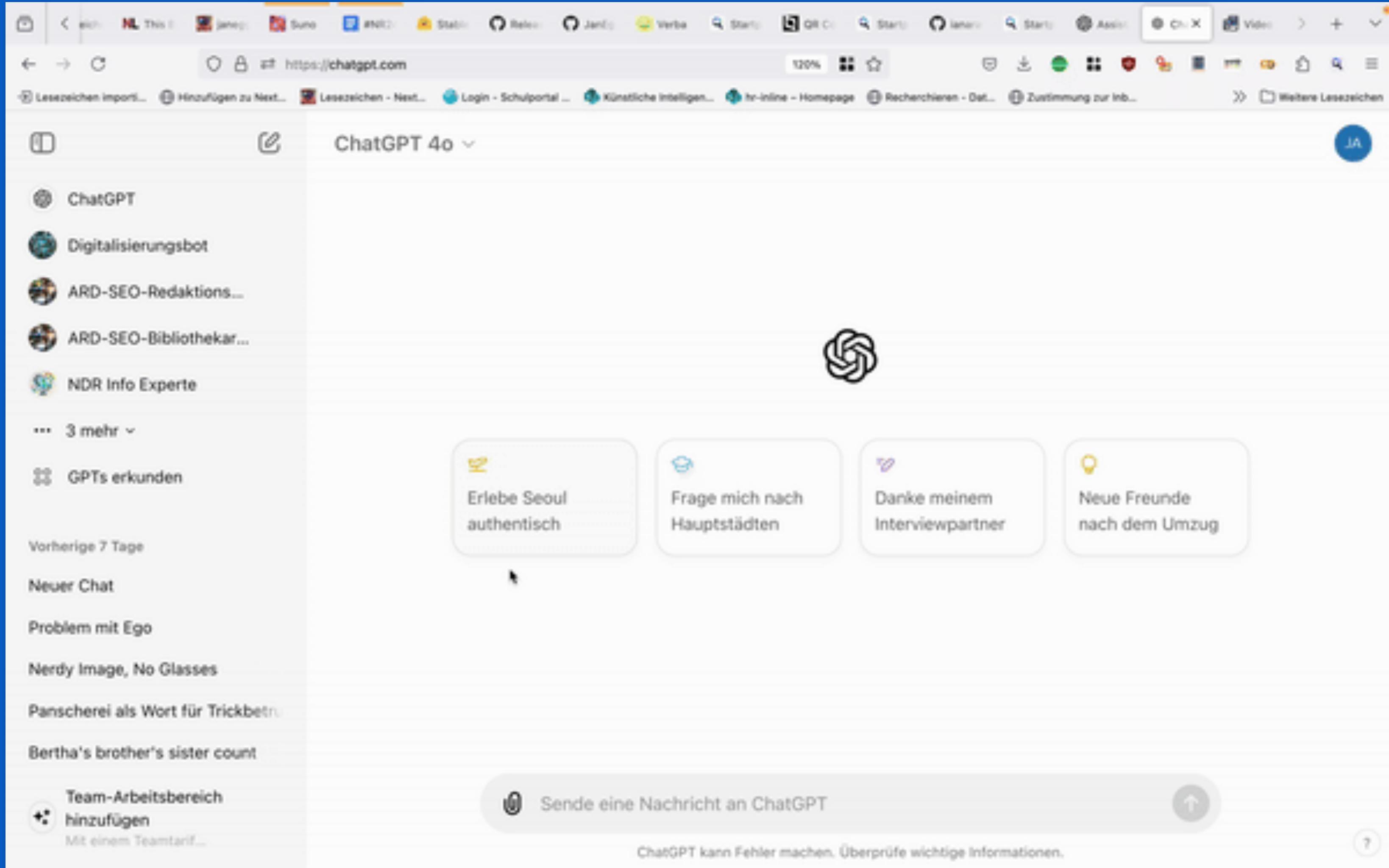
# „Embedding“: KI verortet den Sinn des Textes



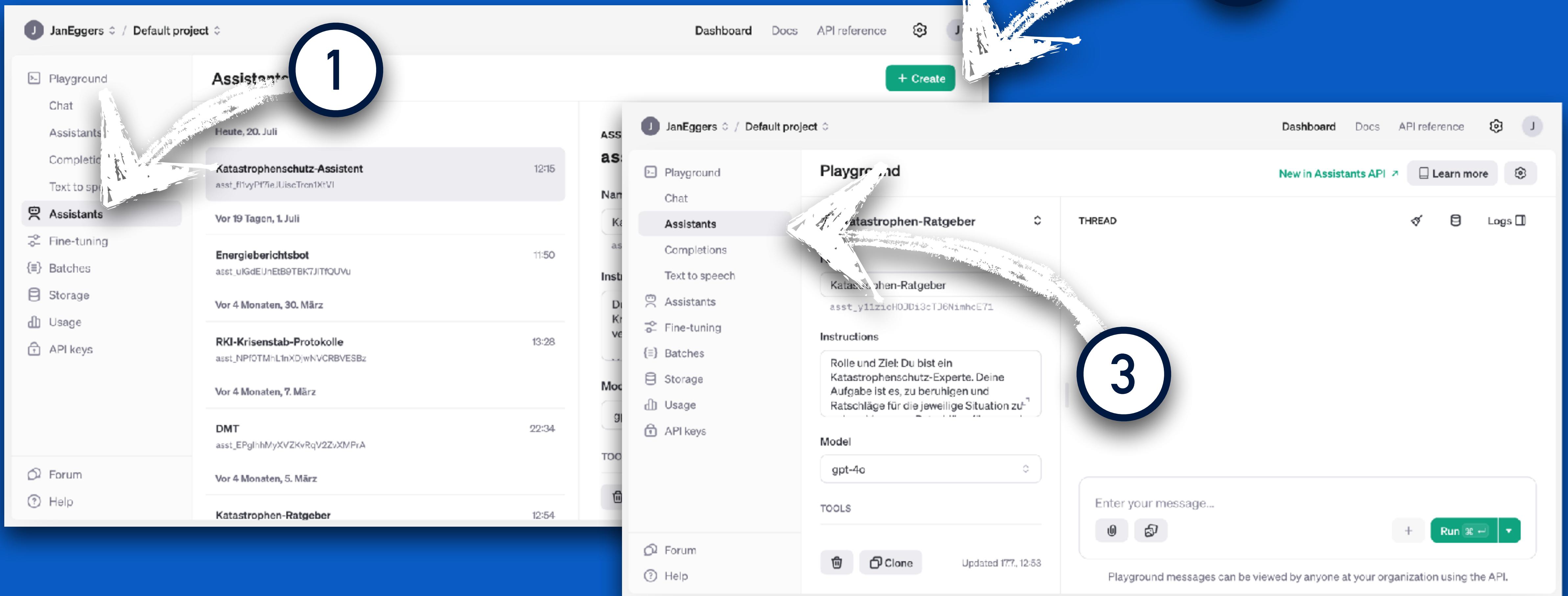
# „Retrieval-augmented Generation (RAG)“:



# ChatGPT Plus bietet „GPTs“ - eigene KI-Assistenten



# Alternative für Menschen ohne ChatGPT-Plus-Bezahlkonto: <https://platform.openai.com>



# Anwendungsfall 2: RKI-Protokolle



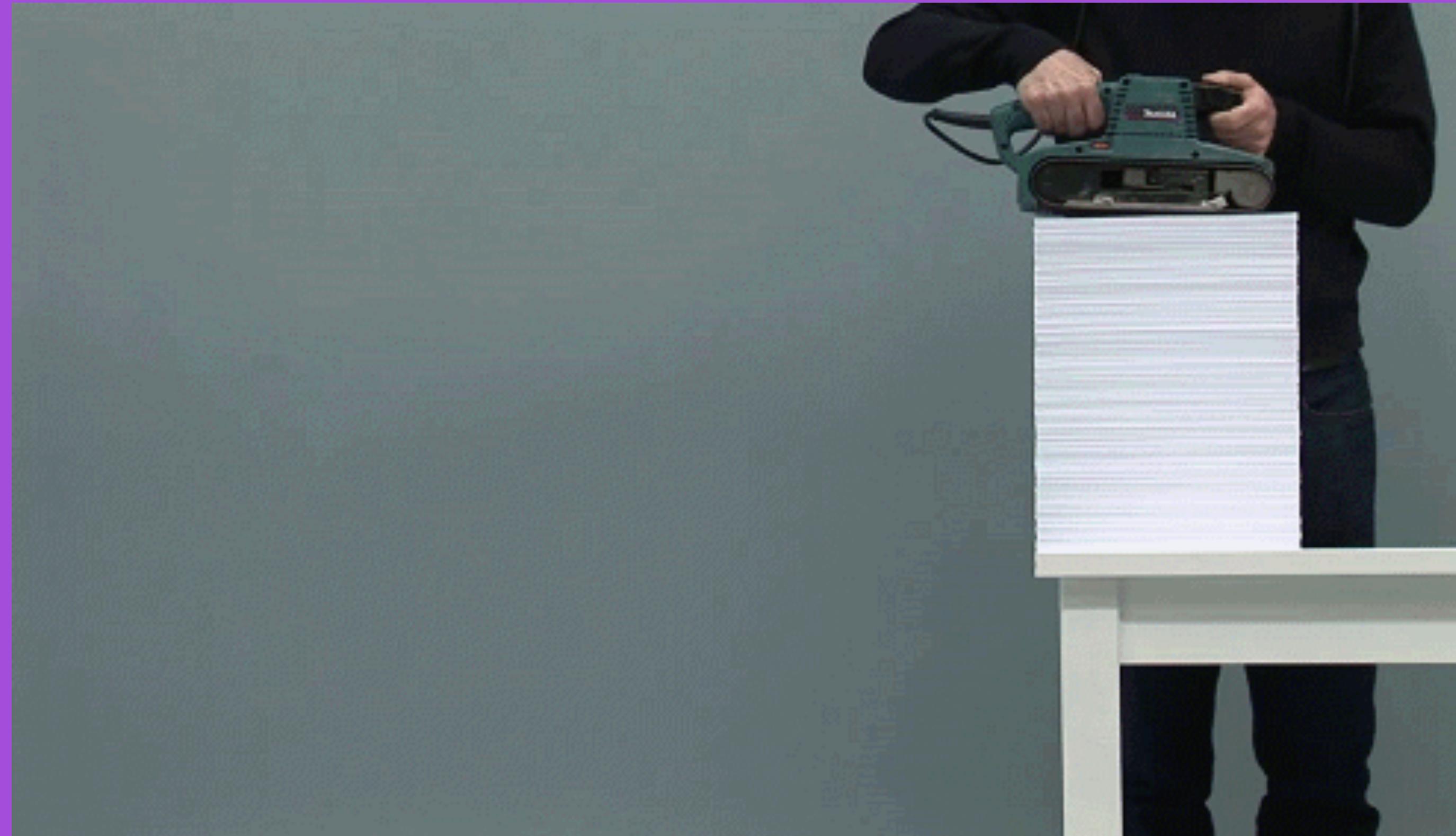


fRAG den Staat

(C) für diesen Kalauer: Jan-Georg Plavec

# 2.200 PDF-Seiten!

TOP	Beitrag/Thema
1	<b>Aktuelle Lage</b> <ul style="list-style-type: none"><li>• Wuhan: 41 Fälle wurden identifiziert. Davon sind 6 aus dem Krankenhaus entlassen worden, 7 in einem kritischen Zustand und 1 Patient ist verstorben. Es gibt mehr als 700 enge Kontaktpersonen, davon gehören ca. 400 zum medizinischen Personal. Der Beginn der Symptome der 41 bestätigten nCoV-Fälle in China reicht vom 08.12.2019 bis zum 02.01.2020.</li><li>• Thailand: Zusätzlich zu den 41 Fällen berichtete die WHO am 13.01.2020 über einen laborbestätigten Fall in Thailand, dabei handelt es sich um eine Chinesin aus Wuhan. In Zusammenhang mit der erkrankten Frau werden derzeit 18 Kontaktpersonen nachverfolgt: 16 Personen im Zusammenhang mit dem Flug, in dem die Patientin symptomatisch reiste, sowie 2 Personen, die denselben Haushalt mit der erkrankten Frau teilen. Sie befand sich nicht auf dem Huanan-Seafood-Markt in Wuhan.</li><li>• Die Inkubationszeit ist nicht bekannt, laut der Falldefinition der WHO wird bis zu 14 Tage angenommen.</li><li>• Laut WHO ist es möglich, dass es eine begrenzte Übertragung von Mensch-zu-Mensch gibt (z. B. im familiären Bereich).</li><li>• [REDACTED]</li><li>• [REDACTED]</li></ul>



[https://www.rki.de/DE/Content/InfAZ/C/COVID-19-Pandemie/COVID-19-Krisenstabsprotokolle\\_Download.pdf?\\_\\_blob=publicationFile](https://www.rki.de/DE/Content/InfAZ/C/COVID-19-Pandemie/COVID-19-Krisenstabsprotokolle_Download.pdf?__blob=publicationFile)

# Prompt aufgesetzt, PDF hochgeladen, und...?

JA

You

Am 17. März 2020?



RKI-Krisenstab-Protokolle

## Relevante Protokollauszüge

Am 17. März 2020 fand keine dokumentierte Sitzung des RKI-Krisenstabs statt, oder es wurden keine spezifischen Entscheidungen oder Diskussionen für diesen Tag in den bereitgestellten Dokumenten vermerkt.

# PDFs sind manchmal sehr meh!

S eite 2 von 5

ROBERT KOCH INSTITUT

SF

VVDSS- I-N NUEUR RN FFSUÜRR T DDGEENN E DBIENRSTAGUEBCRAHUC H

EEiinnssttuuffuunngg aauuffggeeahoobbeenn aamm

1111..0011..22002233 dduurrcchh VVPPrääs s

KKoooorrddinieerrruungnsgsstsetllelele ddeess RRKKI I

AAggeennddaa ddeerr nnCCooVV--LLaaggee--AAGG

po

zs hCETT-”(CtisSCs

?

44 SSuurrvveeiillllaannccaaaannffoorrdeerruunnggeenn

?e\_ EEss ggiibbt bbeerreeitists vvoonn ddeerr WWHHO0 eeiinne

FFaallllddeeffiinnititioon,n , wwaannnn eeiinne

# PDFs umwandeln:

- Zielformat: .txt oder Markdown!
- Python: PyPDF2 -> pdfminer.six -> GroBiD
- Online-Tools: ?



# Anwendungsfall 3: Quizbot

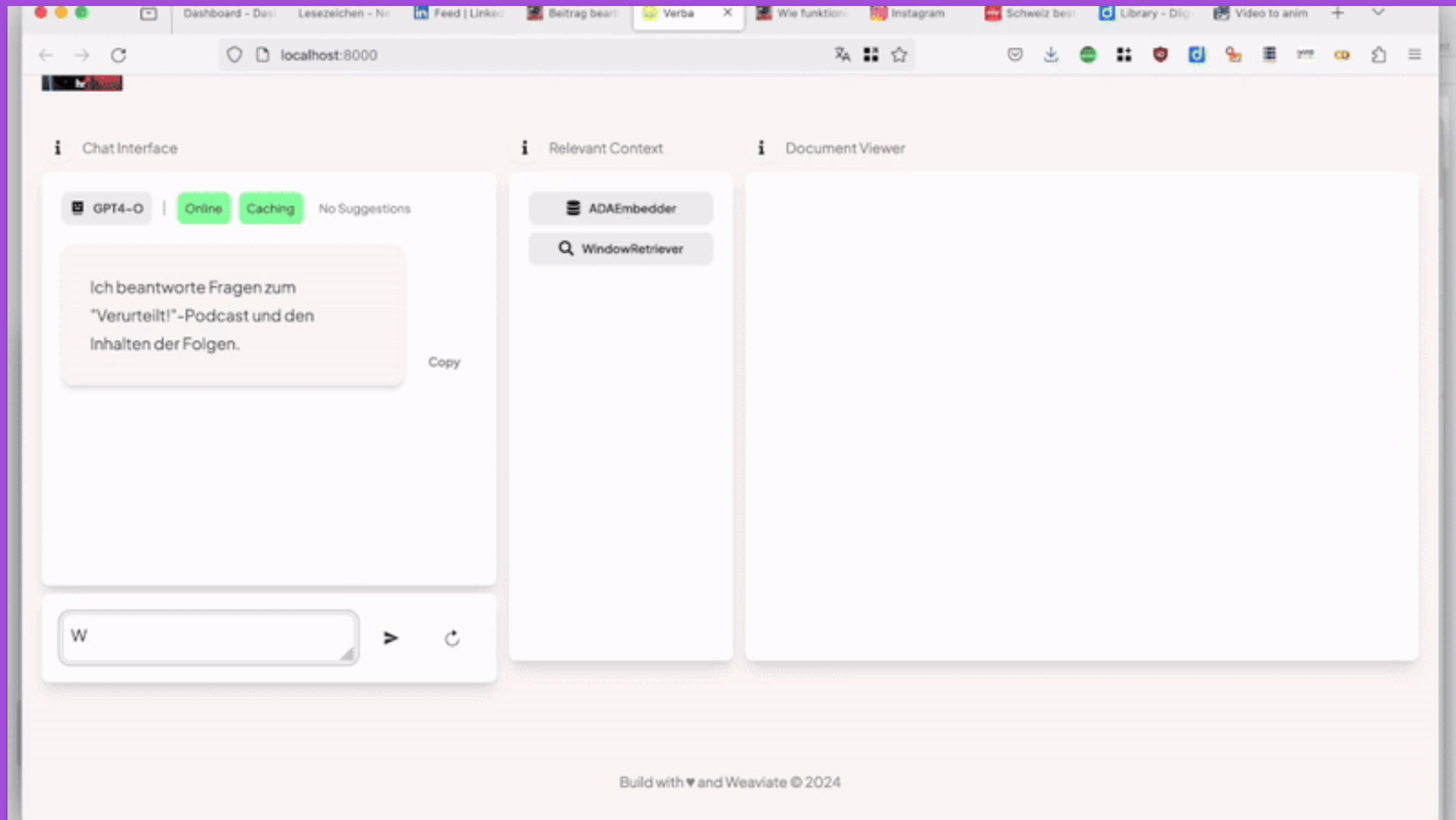


(Geduld, Anwendung im Qualitätsjournalismus folgt!)



# Fragen & Antworten zu 110 Podcast-Folgen

# KI als „Verurteilt“-Quiz-Assistent



Mehr: <https://www.janeggers.tech/li8u> und <https://www.janeggers.tech/dh4b>

# Verba... in Docker-Containern.

[README](#) [BSD-3-Clause license](#)

## Verba

### The Golden RAGtriever

powered by Weaviate ❤️ pip downloads 18k 🚀 D

Welcome to Verba: The Golden RAGtriever, streamlined, and user-friendly interface for easy steps, explore your datasets and extract through LLM providers such as OpenAI, Co

```
pip install goldenverba
```

**docker desktop**

Containers Images Volumes Builds Docker Scout Extensions

Containers Give feedback ⓘ Container CPU usage ⓘ Container memory usage ⓘ Show charts

2.37% / 800% (8 CPUs available) 351.5MB / 7.57GB

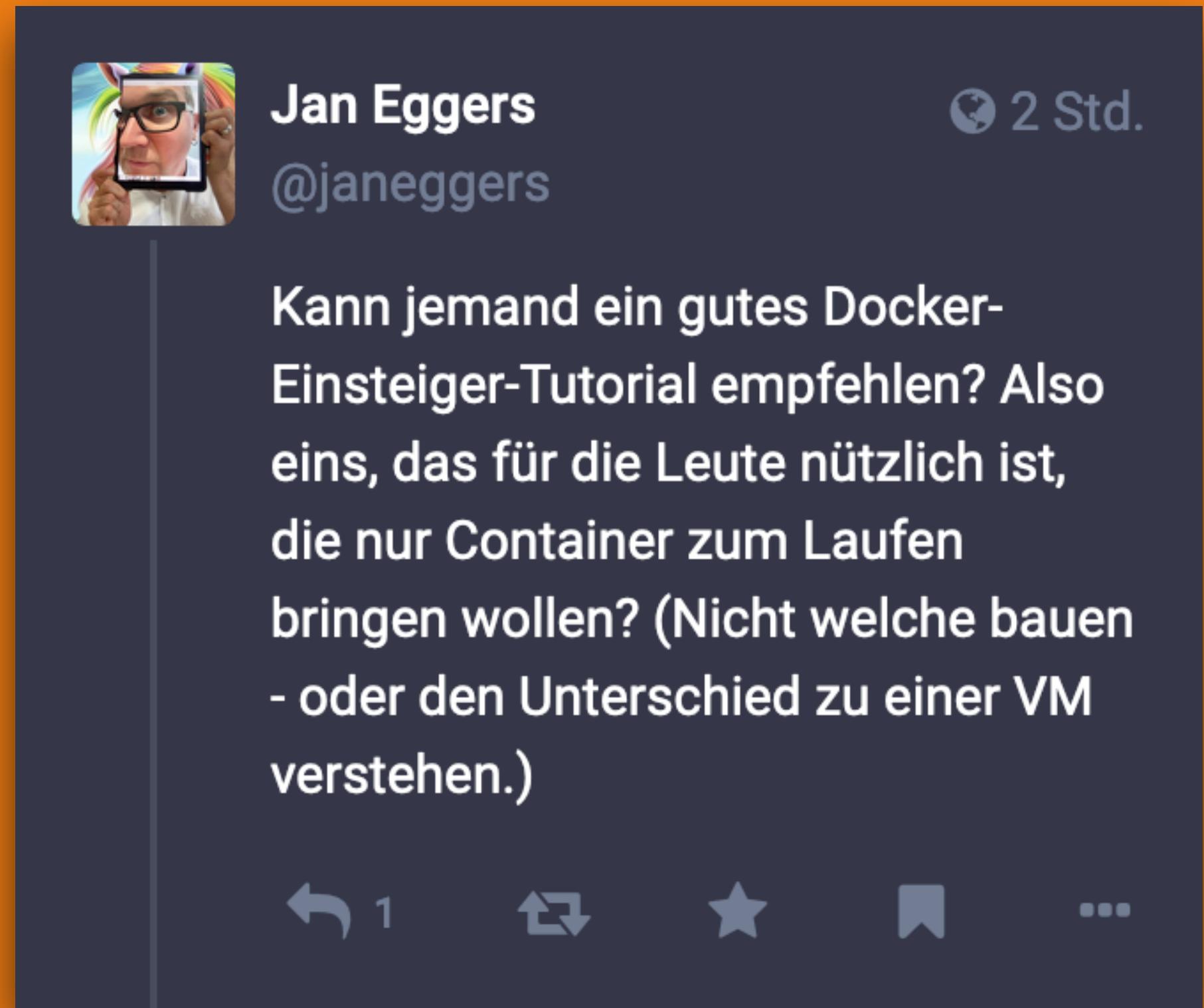
Search Only show running containers Delete ⏪ ⏴ ⏵

<input checked="" type="checkbox"/>	Name	Image	Status	Port(s)	CPU (%)	Last started	Actions
<input checked="" type="checkbox"/>	verba		Running (2/2)		2.37%	27 days ago	⋮ ⏴ ⏵
<input checked="" type="checkbox"/>	verba-1 423e7624	verba-verba	Running	8000:8000 ↗	0.37%	27 days ago	⋮ ⏴ ⏵
<input checked="" type="checkbox"/>	weaviate b1ef082fa	semitechnologies/	Running	3000:8080 ↗ <a href="#">Show all ports (2)</a>	2%	27 days ago	⋮ ⏴ ⏵

Selected 3 of 3

Engine running Kubernetes failed to start RAM 1.92 GB CPU 0.25% Disk 43.98 GB avail. of 62.67 GB 5

# Starthilfe für Docker:



Jan Eggers (@janeggers) · 2 Std.  
Kann jemand ein gutes Docker-Einsteiger-Tutorial empfehlen? Also eins, das für die Leute nützlich ist, die nur Container zum Laufen bringen wollen? (Nicht welche bauen - oder den Unterschied zu einer VM verstehen.)

1 · 0 · 0 · 0 · ...

@janeggers Ich hätte auch Interesse an so etwas.

20. Juli 2024, 12:13 · 0 · Tusky · 0 · 0

# Anwendungsfall 3a: Leak analysieren (Oder eigene Recherche-Notizen)



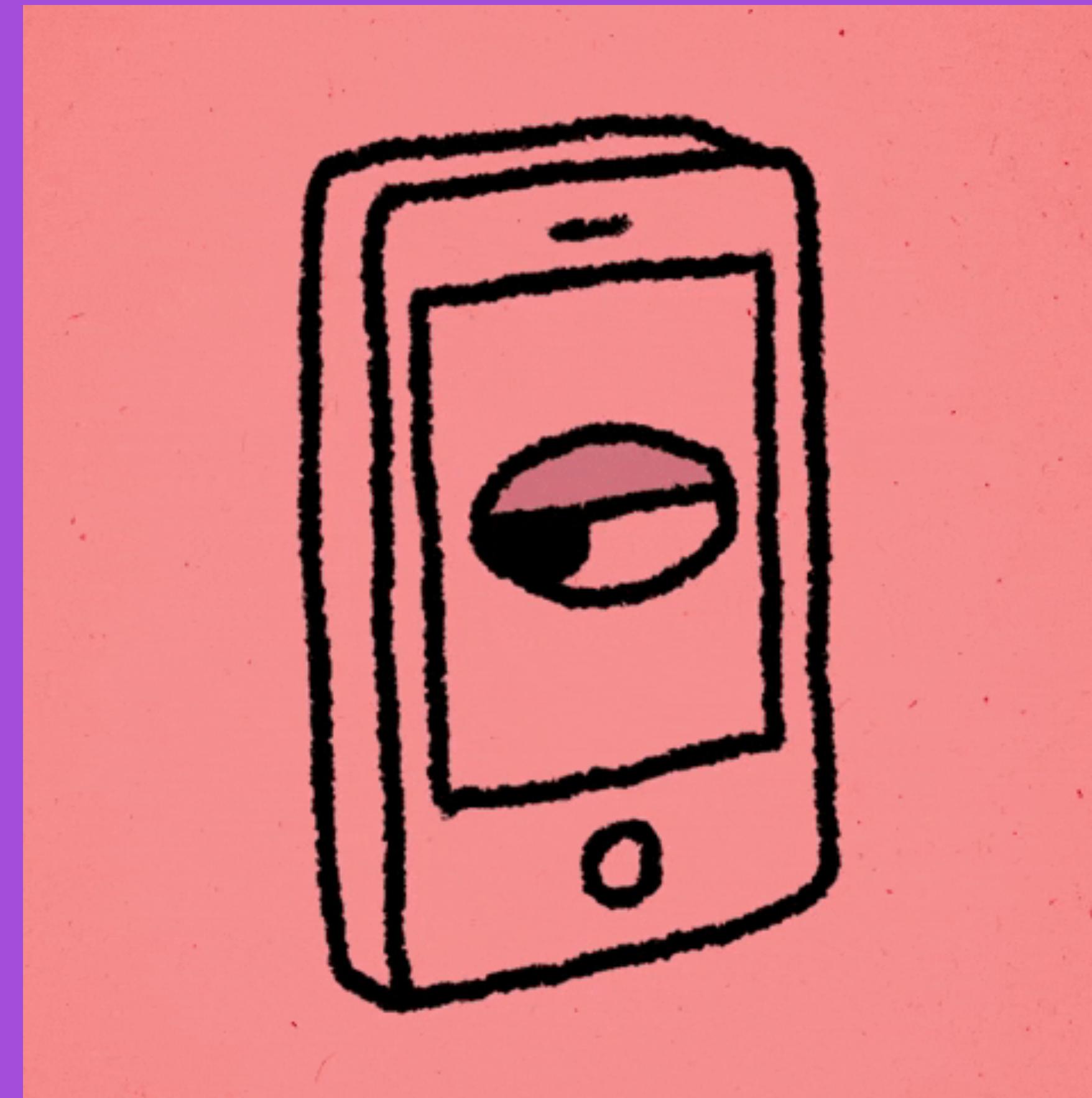
An ICIJ Investigation

## THE PANAMA PAPERS

**Exposing the Rogue Offshore Finance Industry**

A giant leak of more than 11.5 million financial and legal records exposes a system that enables crime, corruption and wrongdoing, hidden by secretive offshore companies.

# ChatGPT & Cie. sind Cloud-Dienste



(und gehören SEHR datenhungrigen Firmen)

# Lokale KI... ...in OLLAMA.



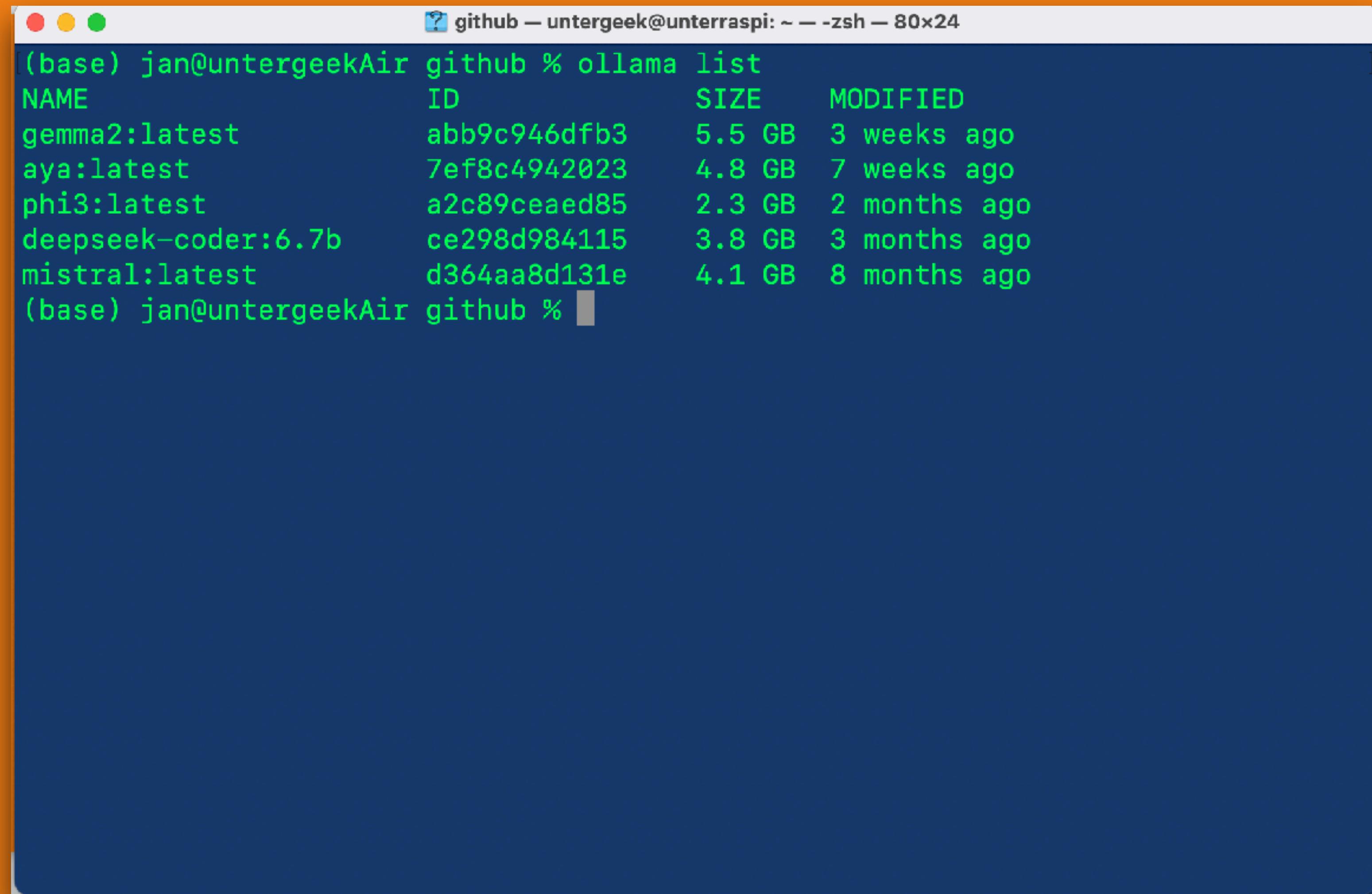
**Get up and running with large  
language models.**

Run [Llama 3](#), [Phi 3](#), [Mistral](#), [Gemma 2](#), and  
other models. Customize and create your own.

[Download ↓](#)

Available for macOS, Linux,  
and Windows (preview)

# gemma2 - Guter Kompromiss aus Größe und Leistung Funktioniert fürs Embedding und die Antworten



The screenshot shows a terminal window with the following text:

```
? github — untergeek@unterra: ~ -- zsh — 80x24
(base) jan@untergeekAir github % ollama list
NAME           ID      SIZE   MODIFIED
gemma2:latest  abb9c946dfb3  5.5 GB  3 weeks ago
aya:latest     7ef8c4942023  4.8 GB  7 weeks ago
phi3:latest    a2c89ceaed85  2.3 GB  2 months ago
deepseek-coder:6.7b ce298d984115  3.8 GB  3 months ago
mistral:latest d364aa8d131e  4.1 GB  8 months ago
(base) jan@untergeekAir github %
```

# Was ihr für ein völlig privates RAG braucht:

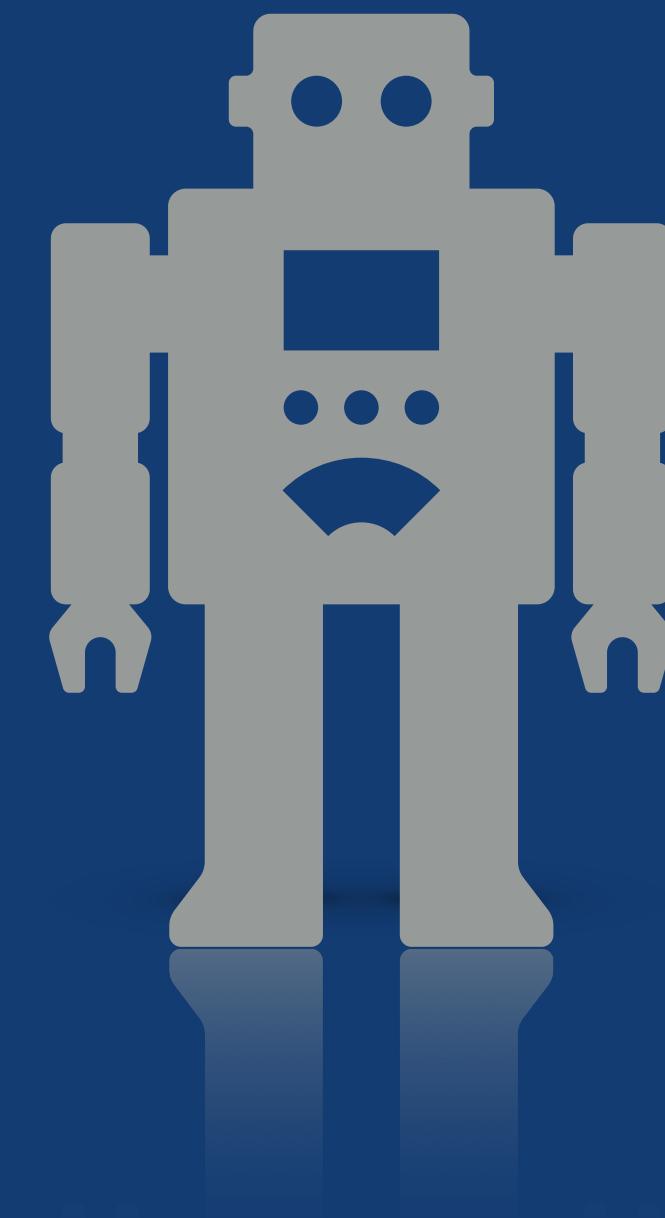
- Docker (oder Rancher) installieren
- OLLAMA installieren
- Rechner, der lokale LLM verkraftet
- Gemma2 ist ein guter Kompromiss
- Frustrationstoleranz
- Geduld



# Was (noch) nicht so gut funktioniert:



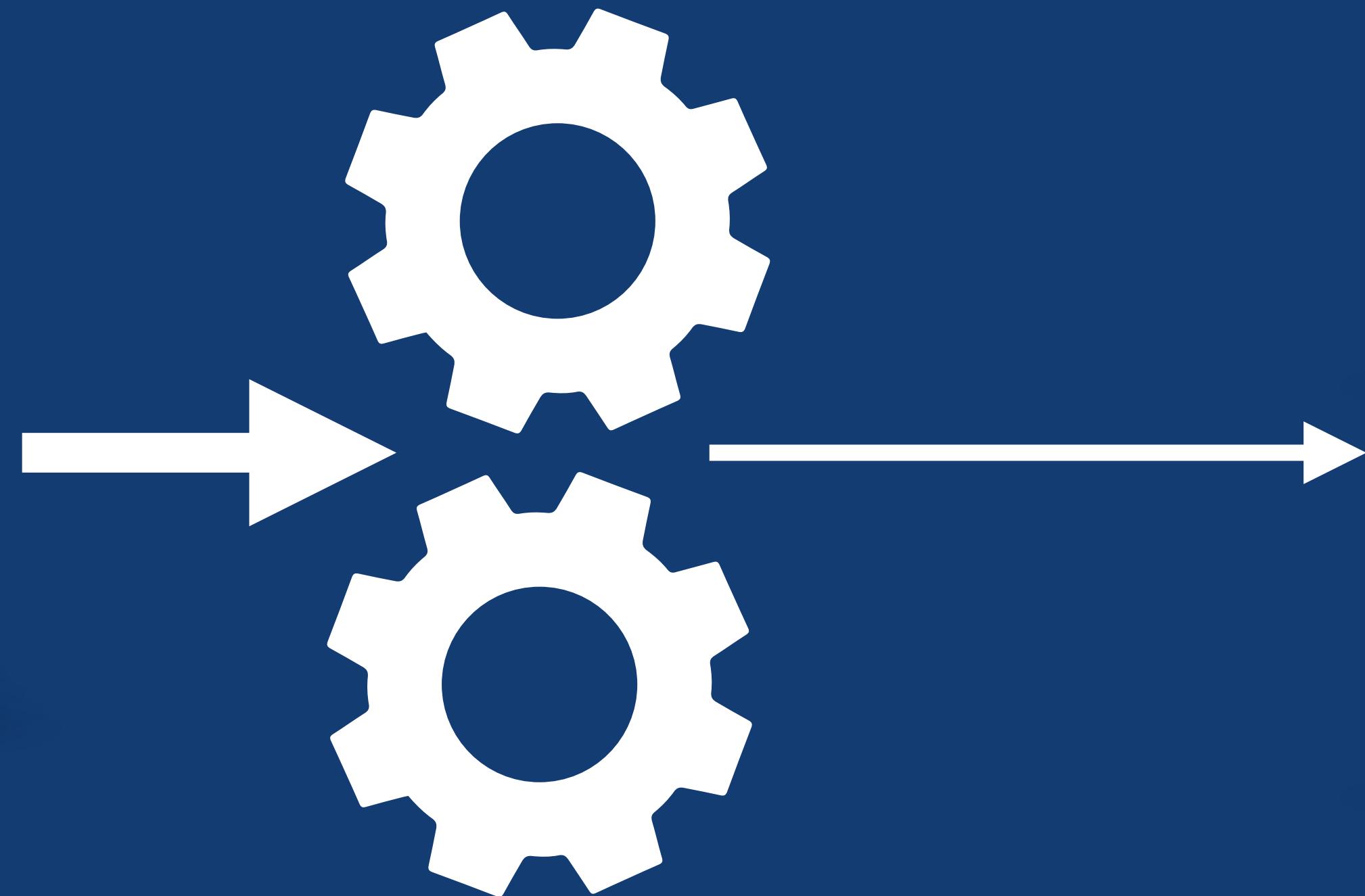
1. Gesamt-Zusammenhang: „Erzähle mir Krieg und Frieden nach“ wird mit RAG nicht funktionieren!
2. Zählen: „Wie oft streitet X mit Y?“ wird RAG nicht beantworten können
3. Ergebnisse manchmal unvollständig: „Alle Stellen, wo X mit Y streitet“ sicherheitshalber lieber mehrfach fragen
4. Quellenangaben bleiben der Pferdefuß von ChatGPT



# Bessere RAGs?



Lorem ipsum dolor  
sit amet, consetetur  
sadipscing elitr, sed  
diam nonumy  
eirmod tempor  
invidunt



Lorem  
ipsum dolor

consetetur  
sadipscing  
elitr,

sed  
diam  
nonumy  
eirmod

tempor  
invidunt.

Lorem ipsum  
dolor sit amet,

consetetur

sadipscing

sed diam  
nonumy eirmod

tempor invidunt.

# Bessere RAGs?

Biographie

**Lorem ipsum  
dolor sit amet,**

Seite 1

Interview 1

**consetetur**

Seite 2

Interview 1

**sadipscing**

Seite 3

Interview 2

**sed diam  
nonumy eirmod**

Seite 4

Interview 2

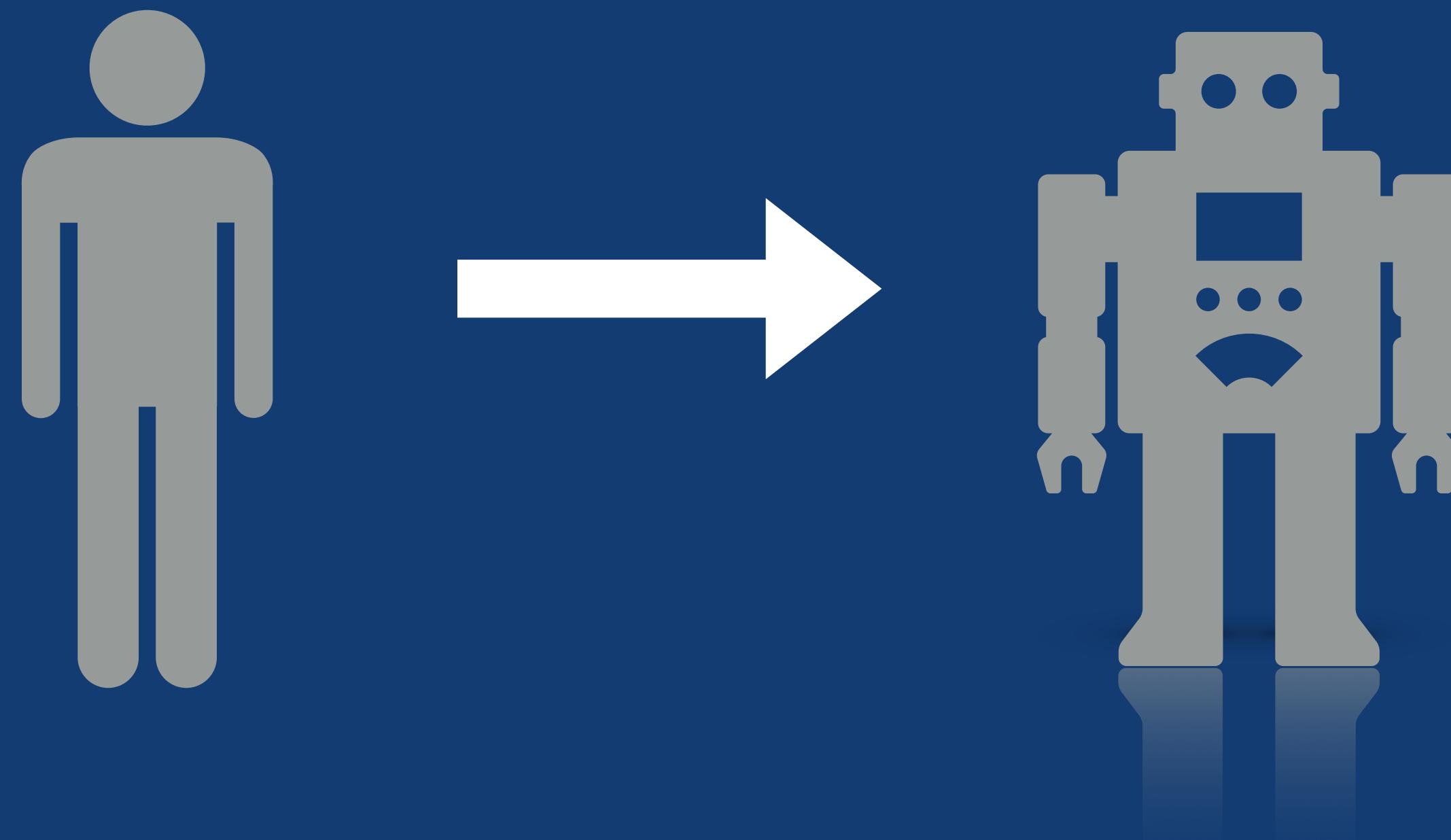
**tempor invidunt.**

Seite 5

# Anwendungsfall: Recherche aufschlüsseln



# Anwendungsfall: Interviewpartner-Trainsdummy



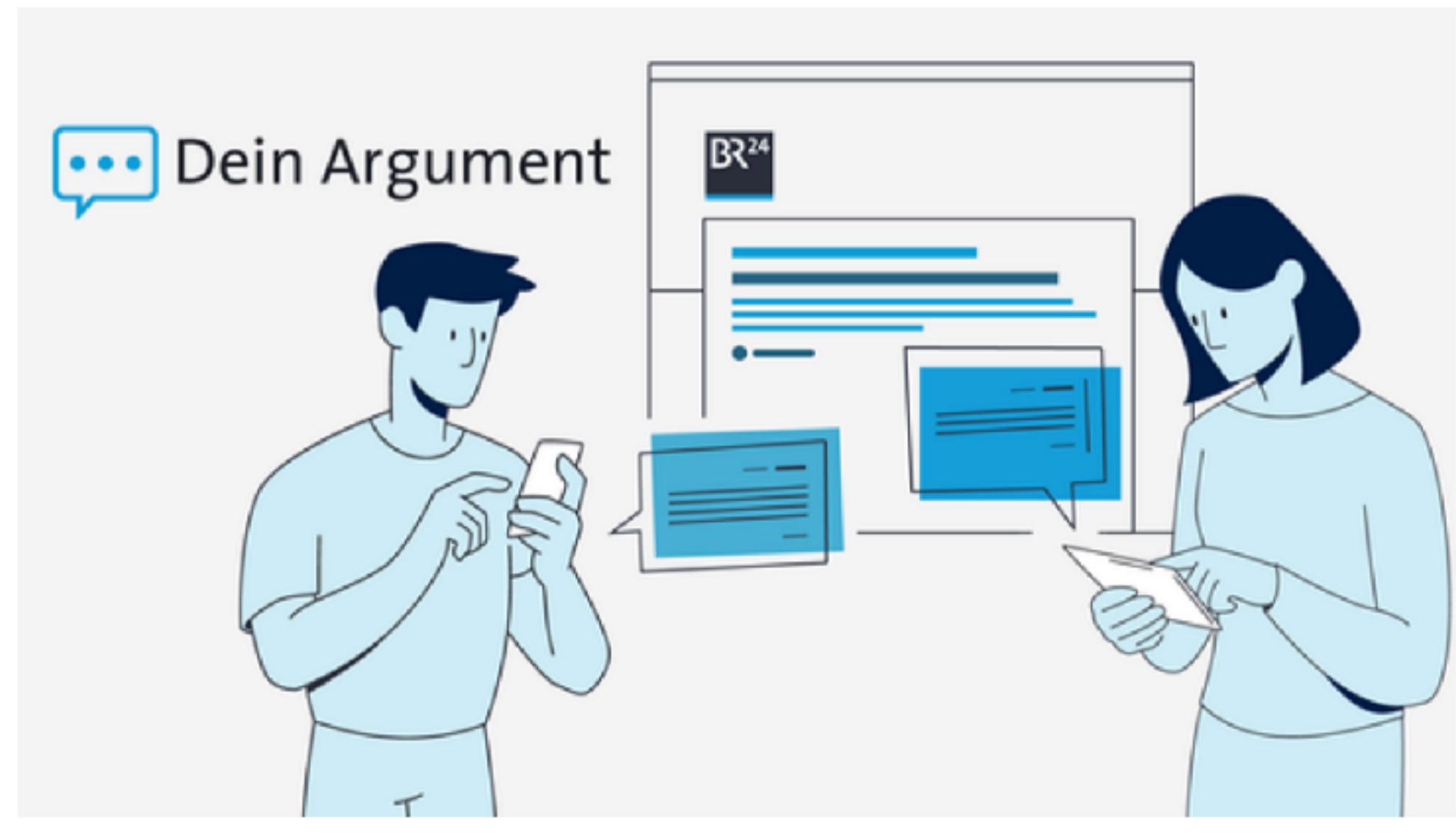
# Anwendungsfall: Faktencheck

## User-Dialog bei BR24

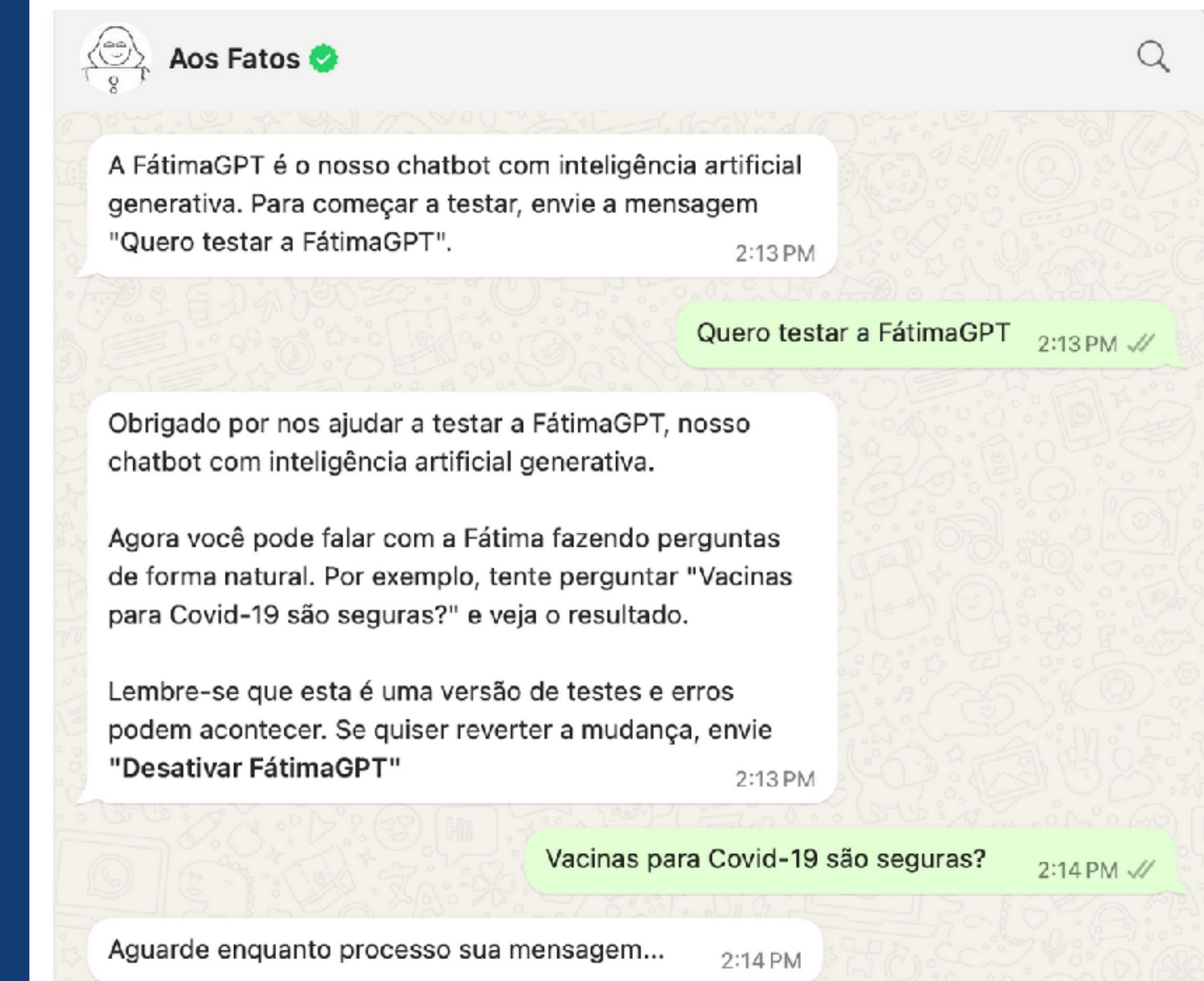
### Wie KI die Arbeit von "Dein Argument" unterstützt

Um verstkt User-Input in die BR24-Berichterstattung einzubinden, sichtet das Team von "Dein Argument" zahlreiche Kommentare. Dabei wird es von Knstlicher Intelligenz (KI) untersttzt. Ihr Einsatz spart bei der Suche nach neuen Argumenten Zeit und untersttzt bei der inhaltlichen Recherche.

Von: Cindy Boden, Teamlead "Dein Argument", und Jrg Pfeiffer, AI + Automation Lab  
Stand: 17.04.2024 10:55 Uhr | [Bildnachweis](#)



# Anwendungsfall: Faktenchecck (2)



This Brazilian fact-checking org uses a ChatGPT-esque bot to answer reader questions

# Chatbots mit KI und „Knowledge Base“

The image displays three screenshots of different chatbot development platforms:

- DMT-Chatbot AI Model configuration:** This screenshot shows the "Set AI" configuration panel. It includes fields for "Antwort beurteilen" (Assess answer), "Data Source" (set to "Knowledge Base"), "Set Variables" (instructions: "Assign whether this information contains enough information to classify the subject is listed in the Digital Media Types."), and "Enter instructions for response (optional)". A "start" button is visible at the bottom.
- Tiledesk Design Studio:** This screenshot shows the "Blocks" section of the Tiledesk Design Studio. It lists blocks such as "start", "defaultFallback", "welcome", "untitled\_block\_1", "untitled\_block\_2", "untitled\_block\_3", and "Special".
- Animal-Mineral-Vegetable-Bot flowchart:** This screenshot shows a flowchart for a bot named "Animal-Mineral-Vegetable-Bot". The flow starts with a "start" block, which leads to a "welcome" block. The "welcome" block has two outgoing paths: one to a "defaultFallback" block (containing the message "Wirklich nicht?") and one to an "untitled\_block\_1" block (containing the message "Alex: Beschreib mir, woran du denkst."). From "untitled\_block\_1", the flow continues through several other blocks, including "untitled\_block\_2" and "untitled\_block\_3".

# Zum Nachlesen und Testen:



<https://github.com/JanEggers-hr/NR24>