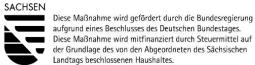


# Funktionsweise von Sprachmodellen

Robert Haase





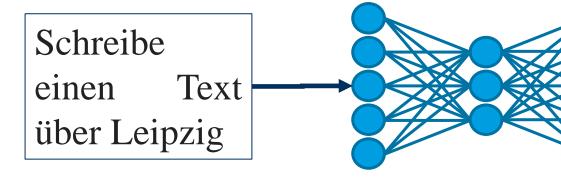






## Große Sprachmodelle

Text-to-text, Text-Generierung

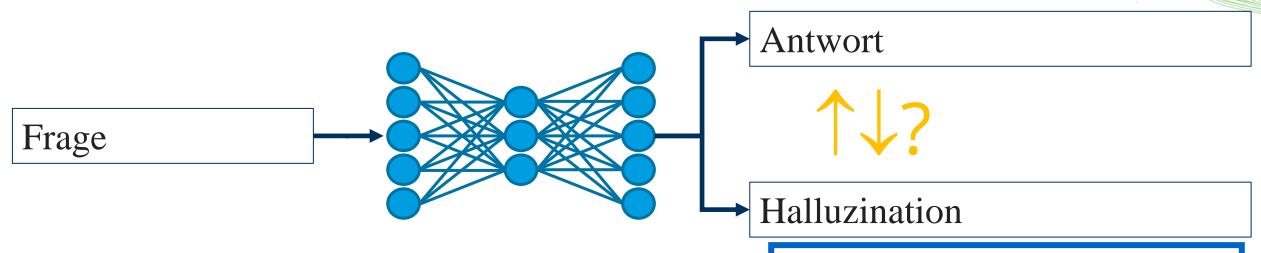


Leipzig, eine Stadt in Sachsen, zeichnet sich durch eine Vielzahl ...



# Große Sprachmodelle

Text-to-text, Wissensextraktion



Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis<sup>†‡</sup>, Ethan Perez<sup>\*</sup>,

Aleksandra Piktus<sup>†</sup>, Fabio Petroni<sup>†</sup>, Vladimir Karpukhin<sup>†</sup>, Naman Goyal<sup>†</sup>, Heinrich Küttler<sup>†</sup>,

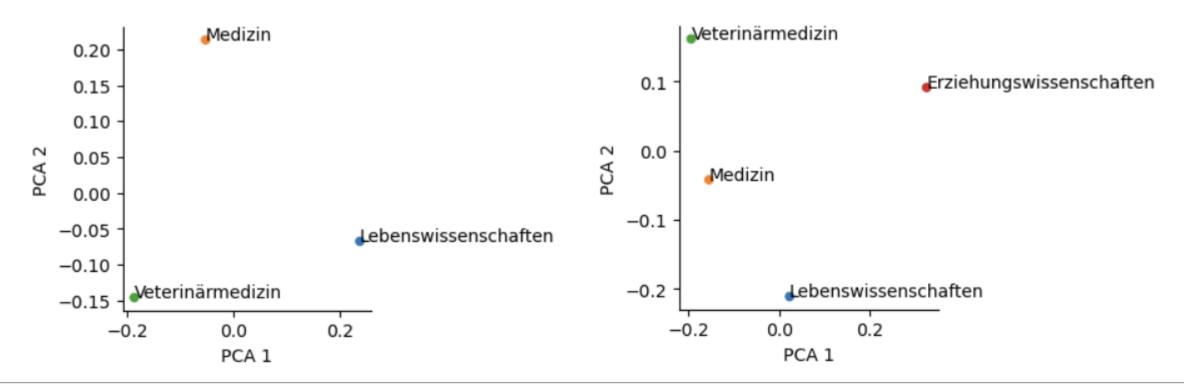
Mike Lewis<sup>†</sup>, Wen-tau Yih<sup>†</sup>, Tim Rocktäschel<sup>†‡</sup>, Sebastian Riedel<sup>†‡</sup>, Douwe Kiela<sup>†</sup>







Embeddings repräsentieren Wörter, Sätze und Texte in From von langen Vektoren von Zahlen. Der Abstand zweier Punkte im Raum beschreibt den Zusammenhang der jeweiligen Dinge.



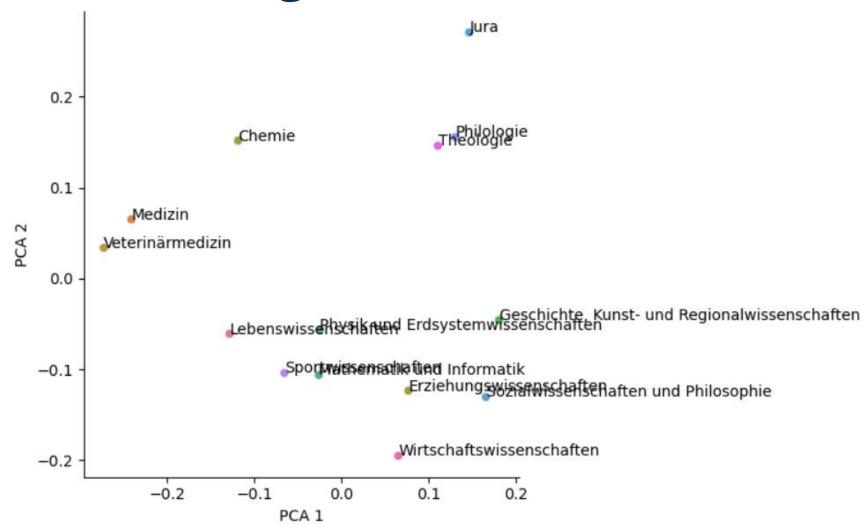






KI-Kompetenzen Robert Haase

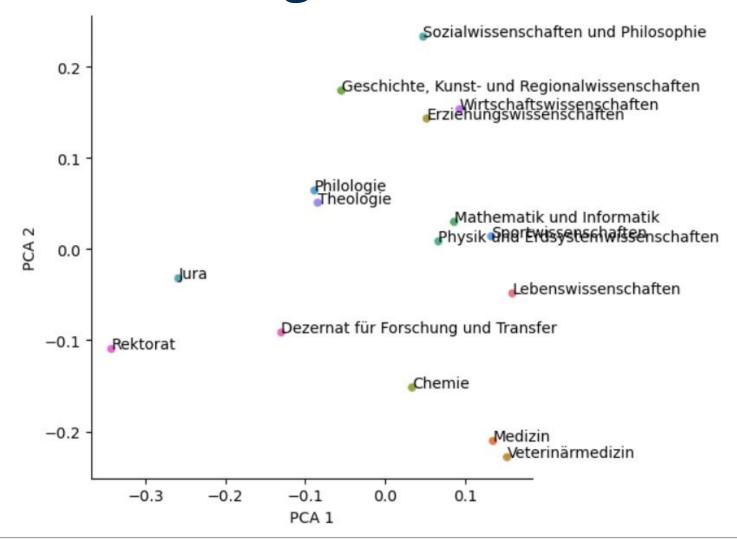
@haesleinhuepf















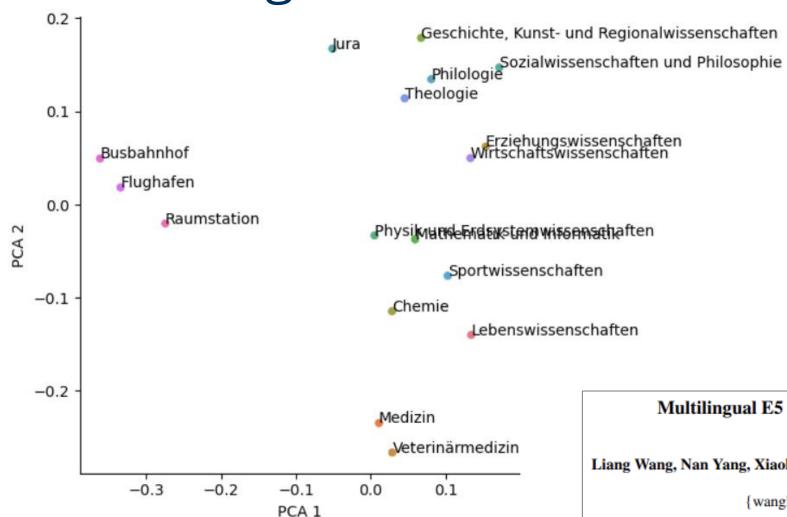


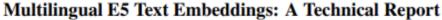
KI-Kompetenzen

@haesleinhuepf

Robert Haase

26. Mai 2025





Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, Furu Wei
Microsoft Corporation
{wangliang,nanya,fuwei}@microsoft.com







Embeddings repräsentieren Wörter, Sätze und Texte in From von langen Vektoren von Zahlen.

#### Generierte Stellenbeschreibungen (Quelle: GPT-40):

	name	research_field	topic	embedding	UMAP0	UMAP1
0	Taylor Reed	Sportwissenschaftliche Fakultät	Effects of different training regimes on athle	[0.03220723941922188, 0.012964349240064621, -0	11.419551	4.258016
1	Jamie Jain	Erziehungswissenschaftliche Fakultät	Erkundungen des Einflusses digitaler Technolog	[0.007033087313175201, -0.003922595642507076,	0.130515	5.945187
2	Avery Garcia	Philologische Fakultät	Analyse des Verwandlungsumfangs der deutschen	[-0.009311092086136341, 0.030486062169075012,	-0.186769	1.813886
3	Skyler Lee	Juristenfakultät	Analyse der Auswirkungen von Vertragsbedingung	[-0.011907541193068027, 0.0018885752651840448,	2.789339	6.239286
4	Robin Smith	Fakultät für Lebenswissenschaften	Die Auswirkungen eines Mangelernährungsstatus	[0.002727826591581106, 0.01623745821416378, -0	12.458349	6.699484

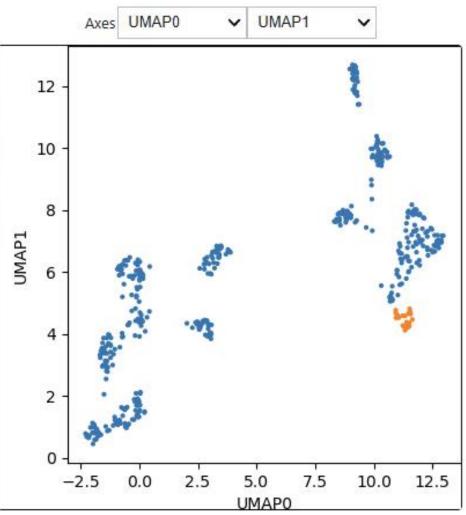






### Übung: Fakultäten raten





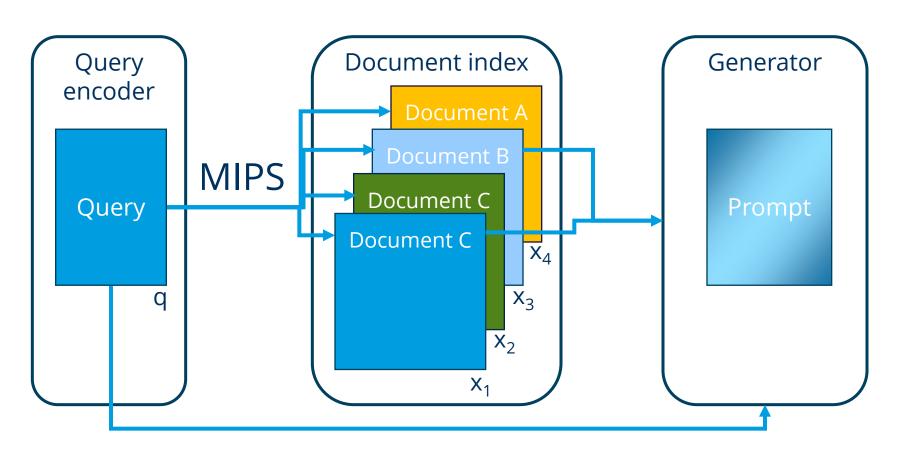






### Retrieval Augmented Generation

Anreichern eines Prompts mit Kontext-spezifischem Inhalt



Maximum inner product search

$$x = \operatorname{argmax}_{x_i \in D} x_i^T q$$



KI-Kompetenzen

@haesleinhuepf 26. Mai 2025

Robert Haase







#### Embeddings - Benchmarks

Leaderboards geben Auskunft über Stärken und Schwächen von Embedding-Modellen

Rank (Bor	Model	Zero-shot	Bitext	Classification	Clustering	Instruction R	Multilabel Class	Pair Classificat	Reranking
1	<pre>gemini-embedding-exp-03-07</pre>	99%	79.28	71.82	54.59	5.18	29.16	83.63	65.58
2	Ling-Embed-Mistral	99%	70.34	62.24	50.60	0.94	24.77	80.43	64.37
3	gte-Qwen2-7B-instruct	▲ NA	73.92	61.55	52.77	4.94	25.48	85.13	65.55
4	multilingual-e5-large- instruct	99%	80.13	64.94	50.75	-0.40	22.91	80.86	62.61
5	SFR-Embedding-Mistral	96%	70.00	60.02	51.84	0.16	24.55	80.29	64.19
6	GritLM-7B	99%	70.53	61.83	49.75	3.45	22.77	79.94	63.78
7	text-multilingual- embedding-002	99%	70.73	64.64	47.84	4.08	22.80	81.14	61.22
8	GritLM-8x7B	99%	68.17	61.55	50.16	2.44	24.43	79.73	62.61
9	e5-mistral-7b-instruct	99%	70.58	60.31	50.57	-0.62	22.20	81.12	63.82
10	Cohere-embed-multilingual- v3.0	▲ NA	<sub>70.56</sub> G	ruppieren vo ähnlichen lı		-1.89	22.74 SC	rtieren von T nach Releva	







#### Embeddings - Benchmarks

Leaderboards geben Auskunft über Stärken und Schwächen von Embedding-Modellen

Rank (Bor	Model	Zero-shot	Memory U	Number of P	Embedding D	. Max Tokens	Mean (T.
1	gemini-embedding-exp-03-07	99%	Unknown	Unknown	3072	8192	68.37
2	<u>Ling-Embed-Mistral</u>	99%	13563	7B	4096	32768	61.47
3	g <u>te-Qwen2-7B-instruct</u>	▲ NA	29040	7B	3584	32768	62.51
4	multilingual-e5-large- instruct	99%	1068	560M	1024	514	63.22
5	SFR-Embedding-Mistral	96%	13563	7B	1996	G :: 0	_
6	GritLM-7B	99%	13813	7B	4096	Größe vo	
7	text-multilingual- embedding-002	99%	Unknown	Unknown	768	odellen h relevan	
8	GritLM-8x7B	99%	20070	<b>/</b> \_	4096	4096	60.49
9	e5-mistral-7b-instruct	99%	Details		4096	32768	60.25
10	Cohere-embed-multilingual-v3.0	▲ NA	kommer: Modelle		1024	Unknown	61.12
4			unbeka	annt	·		

M: Millionen

B: Milliarden (engl. billion)

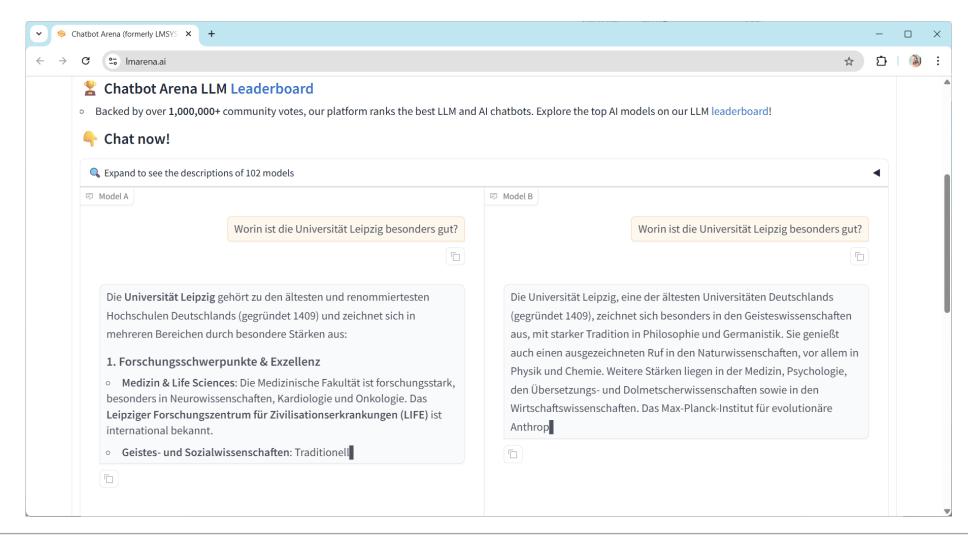
Das Modell ist Klein genug, um es auf meinem Laptop zu benutzen







#### Chatbot Arena Leaderboard





KI-Kompetenzen

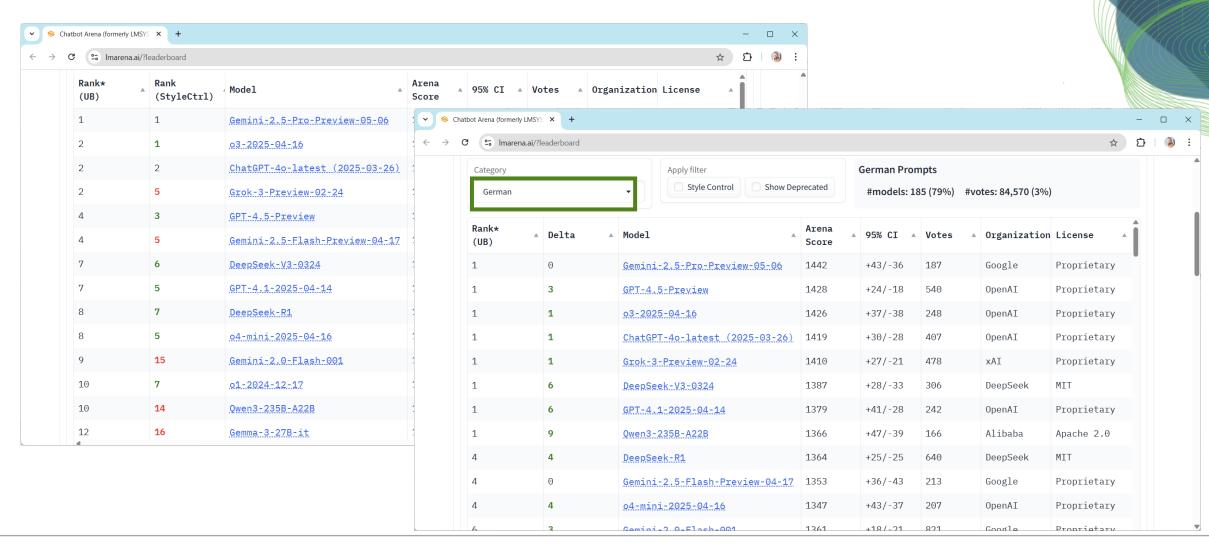
@haesleinhuepf

Robert Haase





#### Chatbot Arena Leaderboard

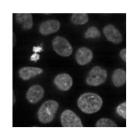








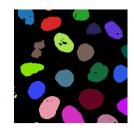
Anwendungsfall: Segmentiere Zellkerne und miss deren Größe















858.04



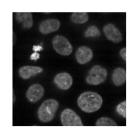


KI-Kompetenzen

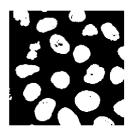
@haesleinhuepf 26. Mai 2025

Robert Haase

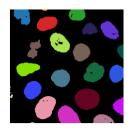
Anwendungsfall: Segmentiere Zellkerne und miss deren Größe















858.04

Pass-rate (n=10):

workflow segmentation measurement sumn

mary	1.0	1.0
	reference	e-3-5-sonnet-20240620

lama3-70b-instruct-q8

phi3-3.8b-mini-instruct-4k-fp16 codellama-70b-instruct-q4 mixtral-8x7b-instruct-v0.1

gemini-pro

lama3-8b-instruct-fp16

command-r-plus-104b-q4\_

KI-Kompetenzen https://www.biorxiv.org/content/10.1101/2024.04.19.590278v1 Robert Haase https://github.com/haesleinhuepf/human-eval-bia @haesleinhuepf 26. Mai 2025 DRESDEN LEIPZIG

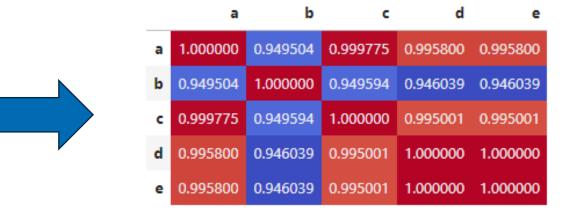




Use-case: compute the correlation matrix

	a	b	c	d	e
0	1.600000	0.100000	1.600000	1.700000	1.700000
1	2.300000	0.200000	2.300000	2.400000	2.400000
2	2.600000	0.300000	2.600000	2.400000	2.400000
3	3.700000	0.300000	3.700000	3.600000	3.600000
4	3.400000	0.400000	3.400000	3.500000	3.500000
5	3.900000	0.400000	3.900000	3.900000	3.900000
6	4.300000	0.400000	4.300000	4.400000	4.400000
7	4.300000	0.500000	4.300000	4.200000	4.200000
8	4.000000	0.500000	4.000000	4.100000	4.100000
9	5.100000	0.500000	5.100000	5.000000	5.000000
10	5.200000	0.600000	5.200000	5.100000	5.100000
11	5.300000	0.600000	5.300000	5.400000	5.400000
12	5.500000	0.600000	5.400000	5.600000	5.600000

pair wise correlation matrix



Pass-rate (n=10):

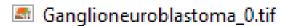




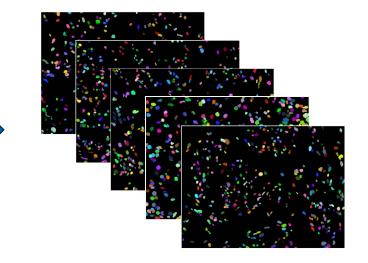


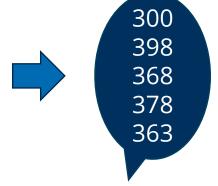


Use case: Count segmented objects in a folder of segmentation results.



- Ganglioneuroblastoma\_1.tif
- Ganglioneuroblastoma\_2.tif
- Ganglioneuroblastoma 3.tif
- Ganglioneuroblastoma\_4.tif





Pass-rate (n=10):



workflow batch process folder count labels



https://github.com/haesleinhuepf/human-eval-bia

Data Source: https://www.ebi.ac.uk/bioimage-

archive/galleries/S-BIAD634-ai.html







Pass-rate (n=10)

reference

89<sup>t</sup>-A<sup>t</sup>Urbor

agrain

egizgiziturde

,Q'

odella

combine\_columns\_of\_tables

create\_umap

t\_test

Statistics / tabular data wrangling

1.0	0.8	0.1	1.0	0.9	0.7	0.1
1.0	0.8	1.0	0.9	1.0	0.8	0.0
1.0	1.0	1.0	0.9	1.0	0.5	0.3

#### Measurements / feature extraction

measure\_intensity\_over\_time
measure\_intensity\_of\_labels
measure\_properties\_of\_regions
count number of touching neighbors

1.0	0.9	0.4	0.1	0.4	0.0	0.1
1.0	0.2	0.4	0.4	0.1	0.0	0.0
1.0	0.4	0.6	0.8	0.2	0.0	0.1
1.0	0.6	0.1	0.2	0.1	0.0	0.0

#### Advanced workflows / big data

	1.0	0.2	0.0	0.0	0.0	0.0	0.0
,	1.0	0.5	0.0	0.9	0.1	0.0	0.0

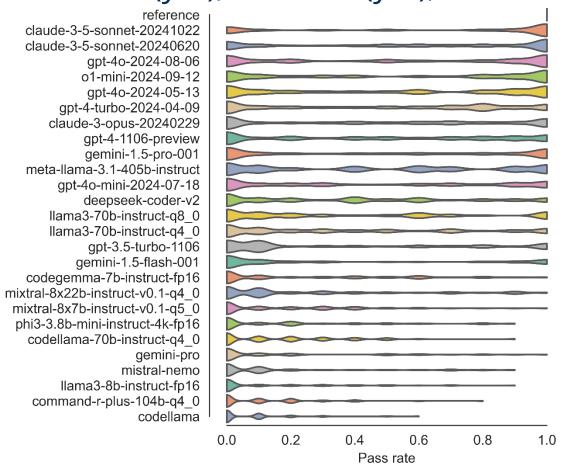
tiled\_image\_processing workflow\_batch\_process\_folder\_measure\_intensity

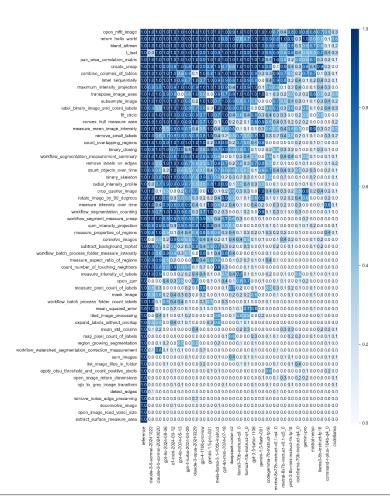






57 use-cases (yet), 26 LLMs (yet), n=10













## Sprachmodelle zur Datenanalyse

Robert Haase

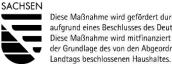
KI-Kompetenzen

@haesleinhuepf

Robert Haase





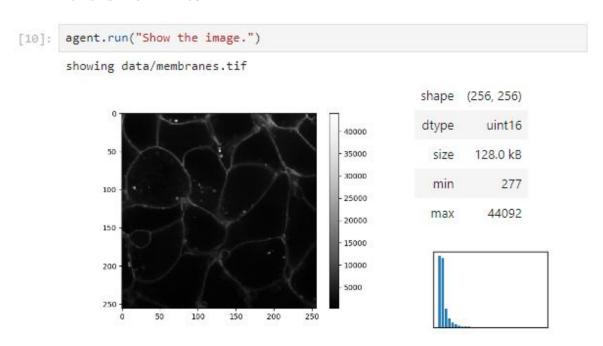




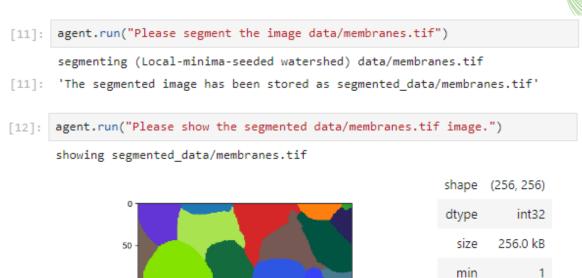


### Datenanalyse-Agenten

KI-Agent: Ein KI-System, das stellvertretend für den Menschen Aktionen ausführt.



Problem: Wir Wissen nicht was konkret im Hintergrund passiert.





DRESDEN LEIPZIG





27

max

150

100 -

150

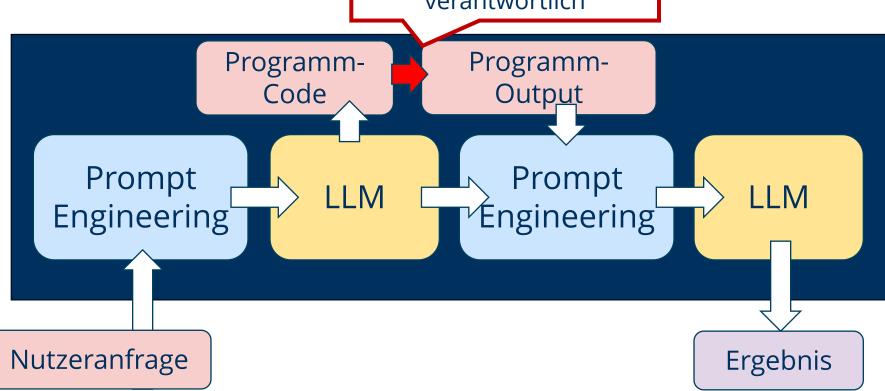
200 -

#### Datenanalyse-Agenten

KI-Agent: ein KI-System, das stellvertretend für einen Menschen

Aktionen ausführt

KI-Betreiber / -Anbieter verantwortlich







Beispiel: Studierendendaten

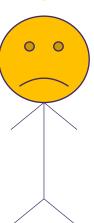
Datenschutz!





Ich möchte bitte die Verteilung des Fachsemesters als Histogram sehen und ein Kuchendiagramm des Studiengangs aus dem die Studierenden sind.

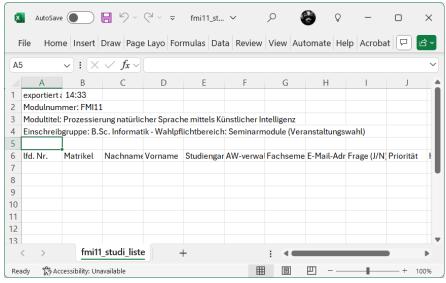




ChatGPT 40 V

Umweg: Datensynthese









Robert Haase

#### Umweg: Datensynthese

- · Reale Namen generieren (Vorname, Nachname)
- Matrikelnummern als fiktive, 7-stellige Zahlen erstellen
- Die Studiengänge "BSc Informatik" und "BSc Digital Humanities" realistisch verteilen
- Fachsemester gemäß deiner Beschreibung zufällig, aber sinnvoll zuweisen (Mittelwert 6, einige auf 2, wenige auf 10)

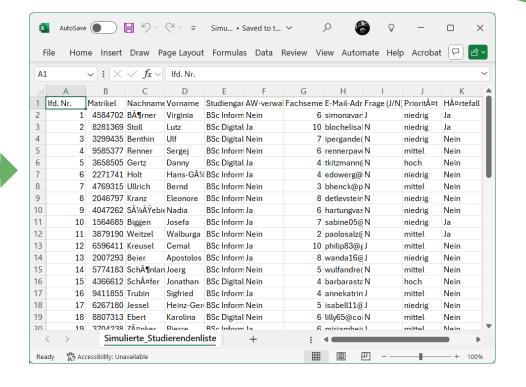
Die übrigen Spalten sinnvoll und zufällig füllen

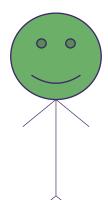
Ich mache das jetzt. [>-]

Simulierte Studierendenliste						
	lfd. Nr.	Matrikel	Nachname			
1	1	4584702	Börner			
2	2	8281369	Stoll			
3	3	3299435	Benthin			
4	4	9585377	Renner			
● ●			▶ .			

Hier ist die mit realistisch wirkenden Daten befüllte Studierendenliste. Wenn du möchtest, kann ich sie dir auch als neue CSV-Datei exportieren oder bestimmte Werte (z.B. mehr Härtefälle oder andere Studiengänge) gezie  $\checkmark$  anpassen. Sag einfach Bescheid! [2-]

Datenschutz







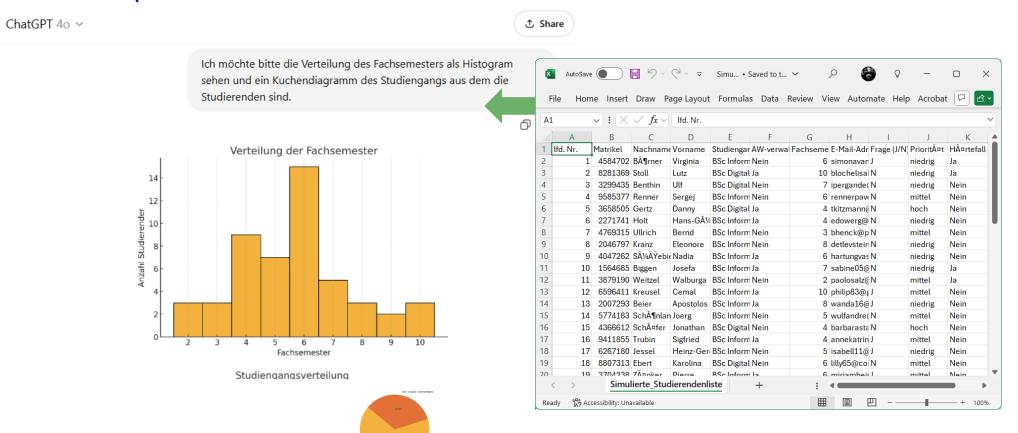
KI-Kompetenzen Robert Haase @haesleinhuepf 26. Mai 2025







Beispiel: Studierendendaten

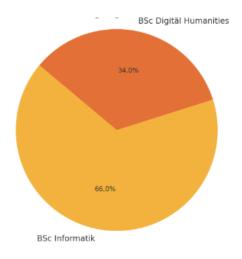








#### Beispiel: Studierendendaten



Spätestens jetzt sind Programmierkenntnisse von Vorteil

#### Hier siehst du:

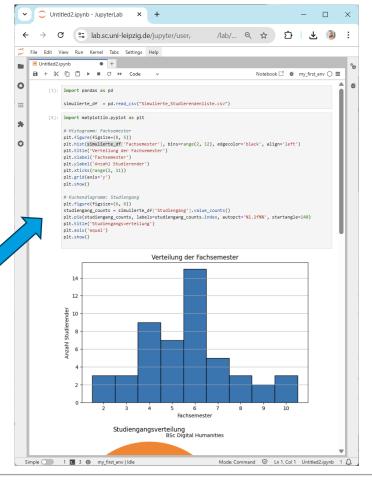
- Ein Histogramm mit der Verteilung der Fachsemester wie gewünscht mit Schwerpunkt im 6. Semester und wenigen Ausreißern.
- Ein Kuchendiagramm zur Verteilung der Studiengänge etwa 70% Informatik, 30% Digital Humanities.

Wenn du weitere Auswertungen oder Diagramme möchtest, sag einfach Bescheid! [>-]

KI-Kompetenzen

@haesleinhuepf

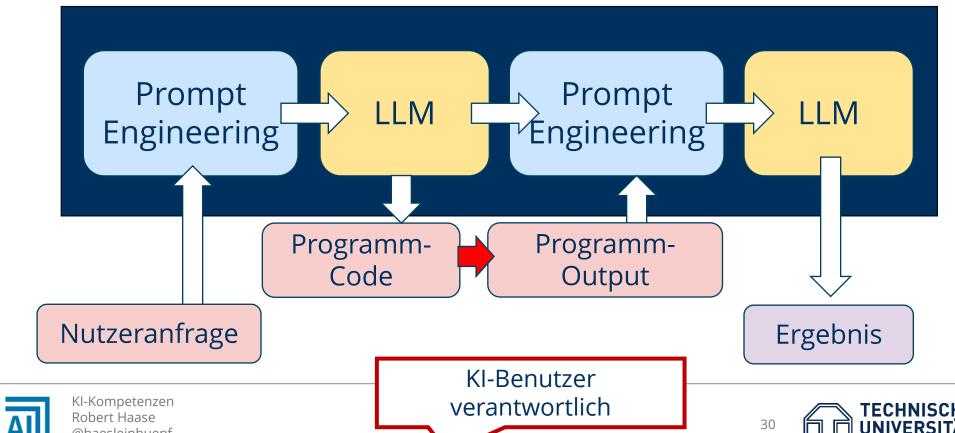
Robert Haase







### Datenanalyse-Agenten





@haesleinhuepf 26. Mai 2025





#### Datenanalyse-Agenten

- Im lokalen IT-System
- Optional mit lokalen
   Sprachmodellen

"bia-bob" (Forschungsprojekt)

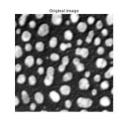


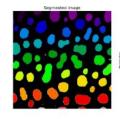
[1]: import bia\_bob

[2]: %%bob
 \* Load blobs.tif from disk
 \* Segment the bright objects on the dark background
 \* Measure area and perimeter of the objects
 \* Show the original image, the segmented image and an

This notebook may contain text, code and images generated by artificial intelligence. Used model: claude-3-5sonnet-20240620, vision model: claude-3-5-sonnet-20240620, endpoint: None, bia-bob version: 0.21.0. Do not enter sensitive or private information and verify generated contents according to good scientific practice. Read more: https://github.com/haesleinhuepf/bia-bob#disclaimer

area versus perimeter plot side-by-side using subfigures.





	label	area	perimeter
0	1	433.0	91.254834
1	2	185.0	53.556349
2	3	658.0	95.698485
3	4	434.0	76.870058
4	5	477.0	83.798990



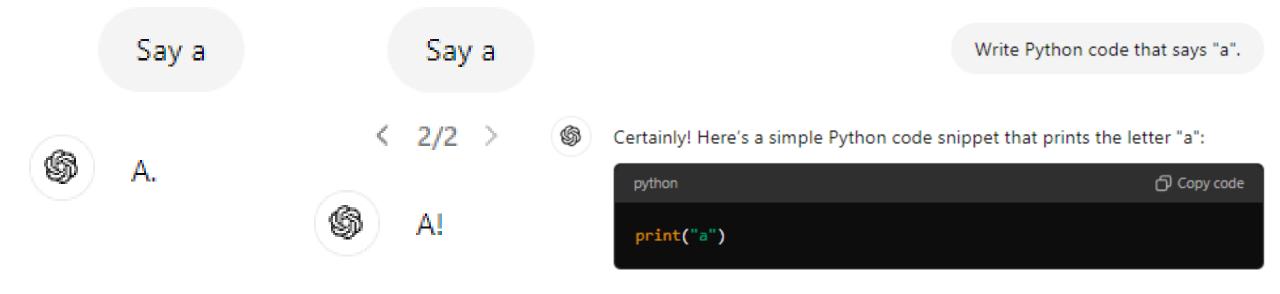
```
import numpy as np
from skimage import io, filters, measure, segmentation
import matplotlib.pyplot as plt
import pandas as pd
import pyclesperanto prototype as cle
import stackview
# Load the image
image = io.imread('blobs.tif')
# Segment the bright objects
thresh = filters.threshold_otsu(image)
binary = image > thresh
# Label the objects
labels = measure.label(binary)
# Measure properties
properties = ['label', 'area', 'perimeter']
df = pd.DataFrame(measure.regionprops_table(labels, properties=properties))
# Create the figure with subfigures
fig, (ax1, ax2, ax3) = plt.subplots(1, 3, figsize=(15, 5))
# Original image
ax1.imshow(image, cmap='gray')
ax1.set_title('Original Image')
ax1.axis('off')
# Seamented image
ax2.imshow(labels, cmap='nipy_spectral')
ax2.set_title('Segmented Image')
ax2.axis('off')
# Scatter plot
ax3.scatter(df['area'], df['perimeter'])
ax3.set xlabel('Area')
ax3.set_ylabel('Perimeter')
ax3.set title('Area vs Perimeter')
plt.tight_layout()
plt.show()
# Display the dataframe
print(df.head())
```





#### Warum Programmcode-Generierung?

 Generierter Text von Sprachmodellen oft nur begrenzt reproduzierbar  Generierter Code wird aber wiederholt die gleichen Ergebnisse liefern









# Sprachmodelle zur Textanalyse

Robert Haase

KI-Kompetenzen

@haesleinhuepf

Robert Haase











#### Informationsextraktion

Selbst unvollständige Informationen sind oft ausreichend.

Hi Tanja und Peter,

wann hättet Ihr Zeit für ein Meeting? Bei mir ginge nächste Woche Donnerstag und Freitag jeweils ab 13 Uhr.

Beste Grüße, Robert







#### Informationsextraktion

Selbst unvollständige Informationen sind oft ausreichend.

#### **Prompt:**

Wann kann das Meeting stattfinden?

{email}

#### Ausgabe:

Das Meeting kann am Donnerstag stattfinden, da sowohl Tanja als auch Robert an diesem Tag Zeit haben. Tanja hat ebenfalls zugestimmt, dass es bei ihr klappt. Peter hat jedoch erwähnt, dass er das Meeting gerne kurz halten würde, da er sehr

beschäftigt ist.

Das Format der Antwort ähnelt dem Format des Prompts. Hi Peter und Robert,

klappt bei mir auch!

VG, Tanja

> Peter schrieb:

> Hi Robert,

> ich bin die Tage total busy und bin unsicher ob es nächste Woche klappt. Wir

> müssen ja auch noch die Vortragsfolien für das Training mit der Stadt fertig

> machen und die zwei neuen Angebote schreiben. Als mir wäre es lieb, wenn wir

> das Meeting kurz halten könnten, und ich kann nur am Donnerstag. Am Freitag

> habe ich Urlaub.

> Viele Grüße,

> Peter

>> Robert schrieb:

>> Hi Tanja und Peter,

>> wann hättet Ihr Zeit für ein Meeting? Bei mir ginge nächste Woche Donnerstag

>> und Freitag jeweils ab 13 Uhr.

>>

>> Beste Grüße,

>> Robert







#### Informationsextraktion

Selbst unvollständige Informationen sind oft ausreichend.

#### **Prompt:**

Heute ist Freitag der 21. März. Gib den ersten möglichen Termin an, wo das Meeting stattfinden kann.

Halte Dich extrem kurz.

{email}

#### **Ausgabe:**

Donnerstag, 27. März, ab 13 Uhr.

KI-Kompetenzen Robert Haase

@haesleinhuepf 26. Mai 2025 Hi Peter und Robert,

klappt bei mir auch!

VG,

Tanja

- > Peter schrieb:
- >
- > Hi Robert,
- > ich bin die Tage total busy und bin unsicher ob es nächste Woche klappt. Wir
- > müssen ja auch noch die Vortragsfolien für das Training mit der Stadt fertig
- > machen und die zwei neuen Angebote schreiben. Als mir wäre es lieb, wenn wir
- > das Meeting kurz halten könnten, und ich kann nur am Donnerstag. Am Freitag
- > habe ich Urlaub.
- >
- > Viele Grüße,
- > Peter
- >
- >
- >> Robert schrieb:
- >>
- >> Hi Tanja und Peter,
- >>
- >> wann hättet Ihr Zeit für ein Meeting? Bei mir ginge nächste Woche Donnerstag
- >> und Freitag jeweils ab 13 Uhr.
- >>
- >> Beste Grüße,
- >> Robert

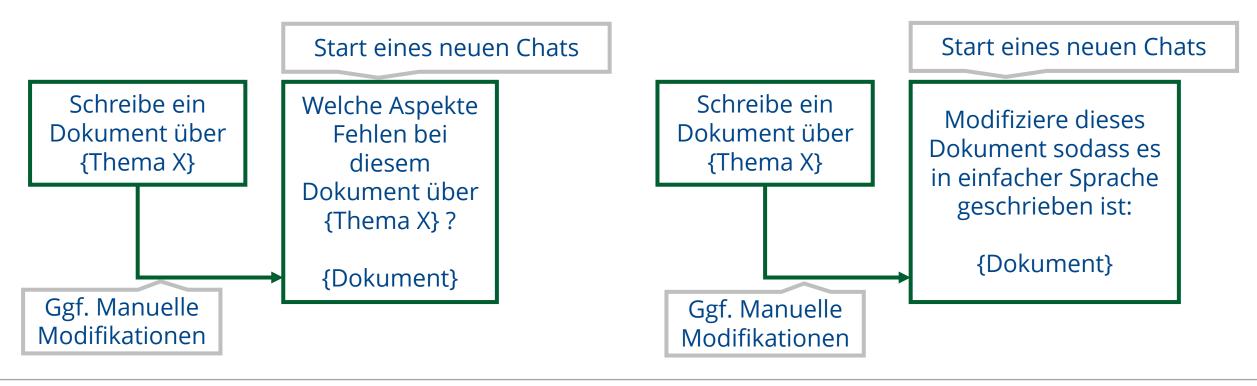






### Weitere Prompt-Engineering Tipps

- Reflektion: Dem Chat-Modell seinen eigenen Output zurückgeben um nach Verbesserungsmöglichkeiten zu fragen
- Tipp: Starten Sie einen neuen Chat, insbesondere nach langen Diskussionen.





KI-Kompetenzen Robert Haase

@haesleinhuepf





#### Bias

Bei manchen Prompts ist es erforderlich explizit Bias zu unterbinden, bspw. im Kontext Diversität

Prompt "Generate an image of a scientist explaining a highly complex topic to a

group of colleagues"

KI-Kompetenzen

@haesleinhuepf

Robert Haase









### Weitere Prompt-Engineering Tipps

Feedback einholen, bspw. zum Ende von Brainstorming Meetings

Meeting Protokoll Meeting Protokoll Brainstorming {Thema} Brainstorming {Thema} Welche Aspekte Welche Aspekte Fehlen Fehlen bei bei diesem Offenheit diesem Sitzungsprotokoll? Sitzungsprotokoll Beachte Eignung {Richtlinien/Aspekte}. {Protokoll} {Protokoll} spezifisches Feedback Generisches VS.

- Diversität
- Fachliche Korrektheit /
- Gerechtigkeit
- Transparenz









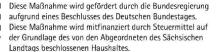
#### Gruppenarbeit

### Übungen Robert Haase









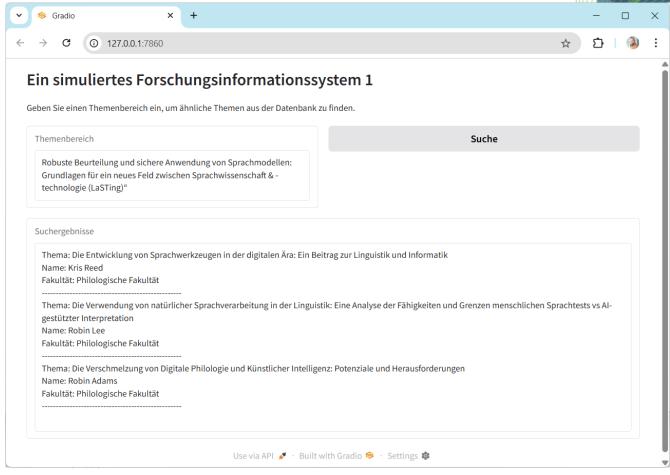




# Übung: Forschungsinformationssysteme

Sie bekommen Zugriff auf 3 Forschungsinformationssysteme um Forschende zu finden, die für konkrete Projekte / Calls in Frage kommen. Welches FIS...

- gibt zufällige Antworten?
- nutzt Embeddings?
- nutzt ein Sprachmodell direkt?



User-Interface / Programmcode wurden mit Cursor KI-generiert.

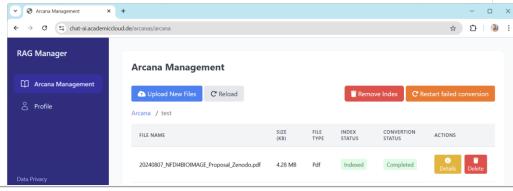


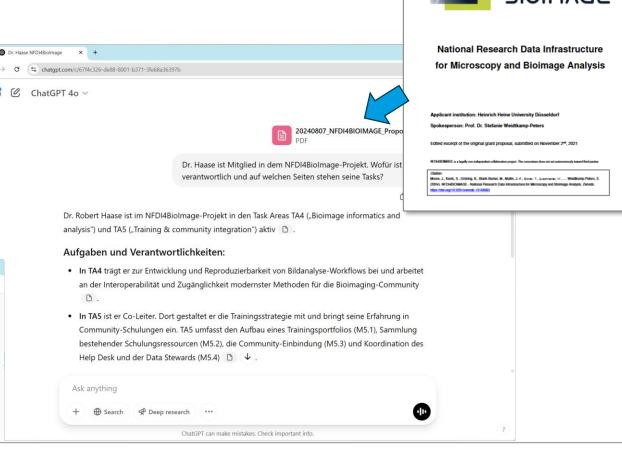




### Übung: Durchsuchen von Fördermittelanträgen

- Laden Sie das Projektproposal NFDI4Biolmage herunter und übergeben Sie es an ChatGPT.
- Vergleichen Sie mit dem Arcanas-System von GWDG / Kisski.







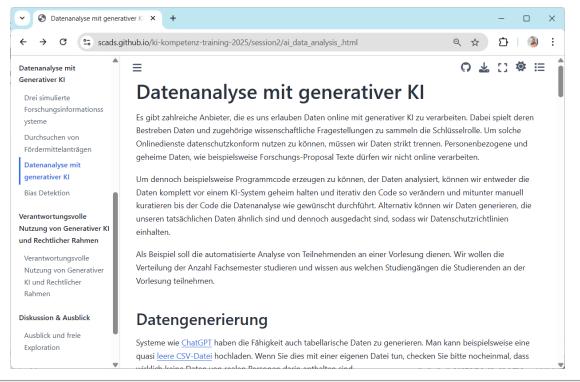


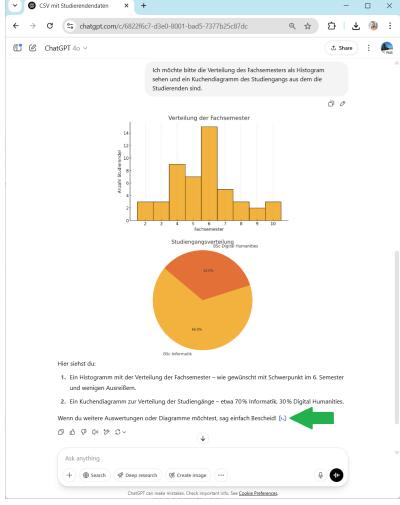


NFDI4

# Übung: Datenanalyse mit generativer Kl

Generieren Sie synthetische Daten von Studierenden und analysieren Sie diese.













## Übung: Bias Detektion

Holen Sie Feedback bzgl. eines Meetingprotokolls von ChatGPT ein.

- Generisches Feedback
- Spezifisches Feedback
  - Diversität
  - Fachliche Eignung der Sprechenden

Hinweis: Sie können mit dieser Methode auch versehentlich Bias etablieren.

#### **Sitzungsprotokoll - Planung Symposium**

Datum: 3. April 2025, 10:00 – 10:35 Uhr (Zoom)

Teilnehmer: Dr. Thomas Becker, James Müller, William Schubert

#### Besprechungspunkte:

- 1. Thema & Zielsetzung
- Konsens über eine interdisziplinäre Ausrichtung.
- Vorschlag von Schubert angenommen: "Moderne Verwaltung mit ChatGPT"

#### 2. Mögliche Sprecher Vorschläge (Müller):

- Dr. Richard Müller (Stadtverwaltung Zürich)
- Dr. Anton Berg (KI in der Klimaforschung, UFZ Leipzig)
- Prof. Josef Angermann (Verwaltungsdirektor Uniklinik Dresden)

Becker betont, dass alle angefragten Sprecher international bekannt sein sollen.

Entscheidung: Kontaktaufnahme mit allen dreien bis nächste Woche.

Nächstes Treffen: 14. April 2025, 10:00 Uhr (online)

Ende der Sitzung: 10:35 Uhr

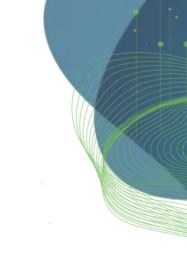






#### Feedback-Runde

Was hat gut funktioniert? Was nicht?





#### **Fazit**

#### Zur Datenanalyse mit Sprachmodellen haben wir 3 Optionen

Analyse von Zusammenhängen zwischen Texten mittels Embeddings

Programmcode-Generierung

Informationsextraktion direkt mit dem Sprachmodell

#### Nutzerfreundlichekeit

#### Erforderliche Programmierkenntnisse

Wir helfen gerne!



Bester Ansatz?





