Henri Wagner und Max Althaus

Data Challenges 2024:

NLP auf dem Corpus Nummorum Datensatz historischer Münzen

M-DS-ADS, bei Dr. Karsten Tolle

Übersicht

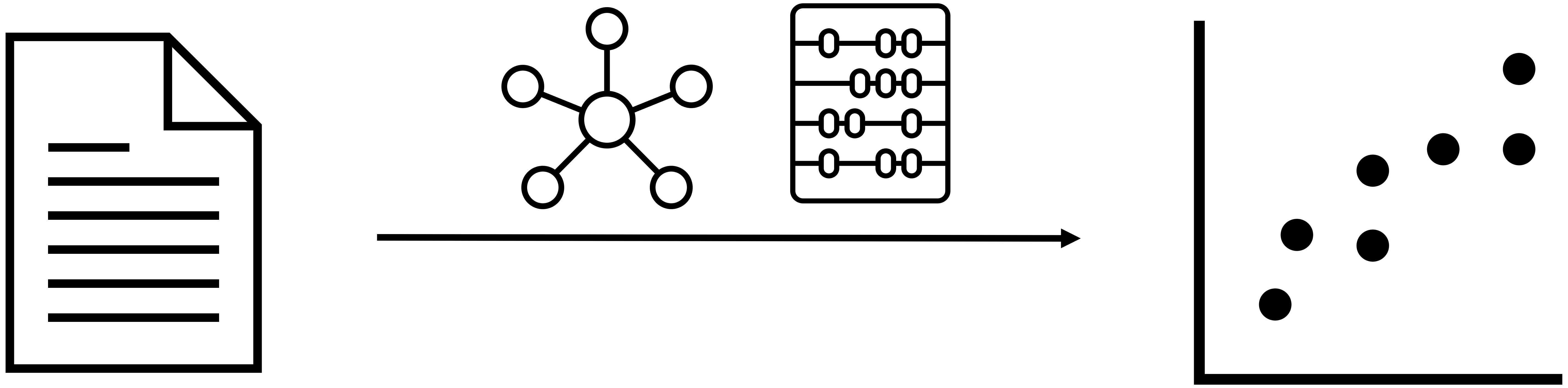
- Hintergrund
- Pipeline
- Previous Work
- Vergleich verschiedener Modelle
- Auswertung
- Demo
- Fazit

Hintergrund

Was sind Embedding und wie können sie hilfreich sein?

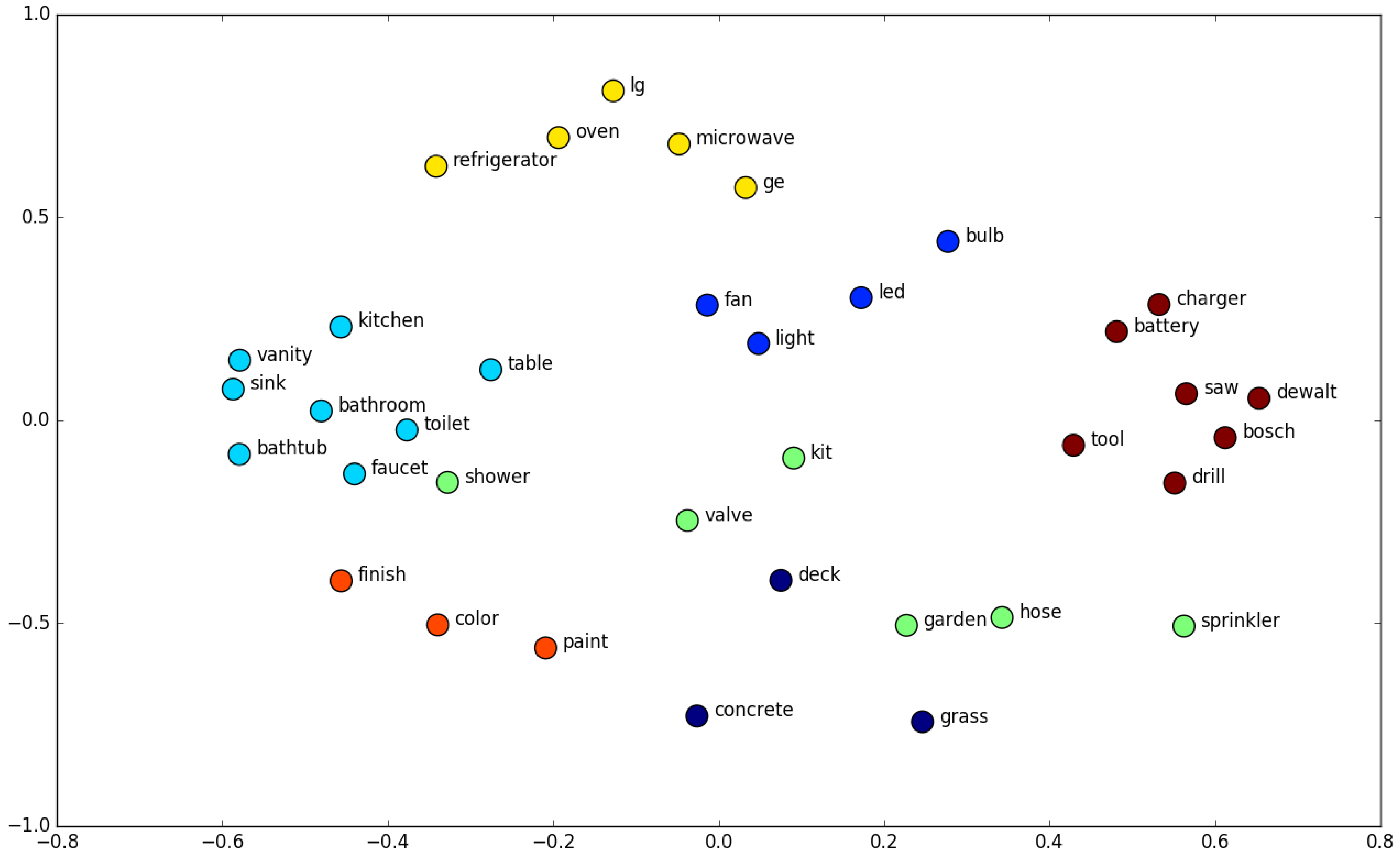
Hintergrund

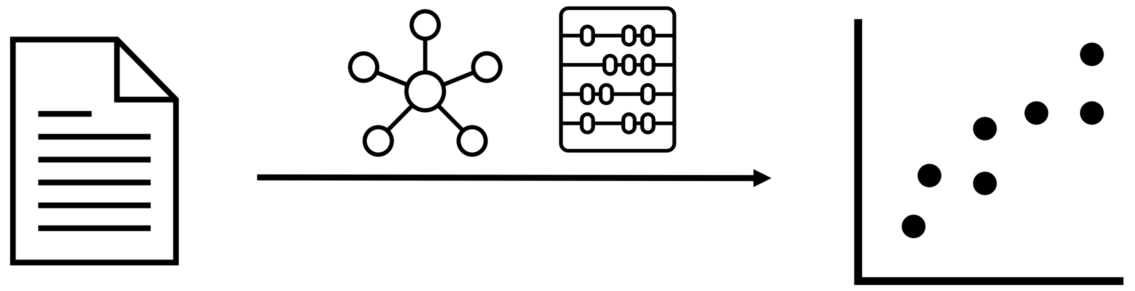
Was sind Embedding und wie können sie hilfreich sein?



Hintergrund

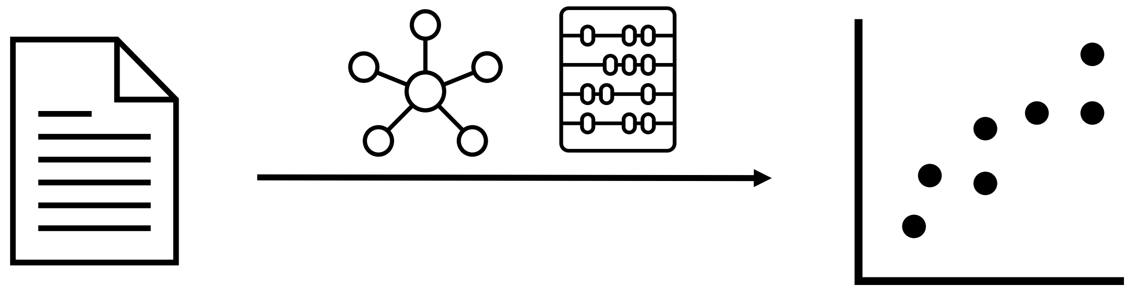
Was sind Embedding und wie können sie hilfreich sein?



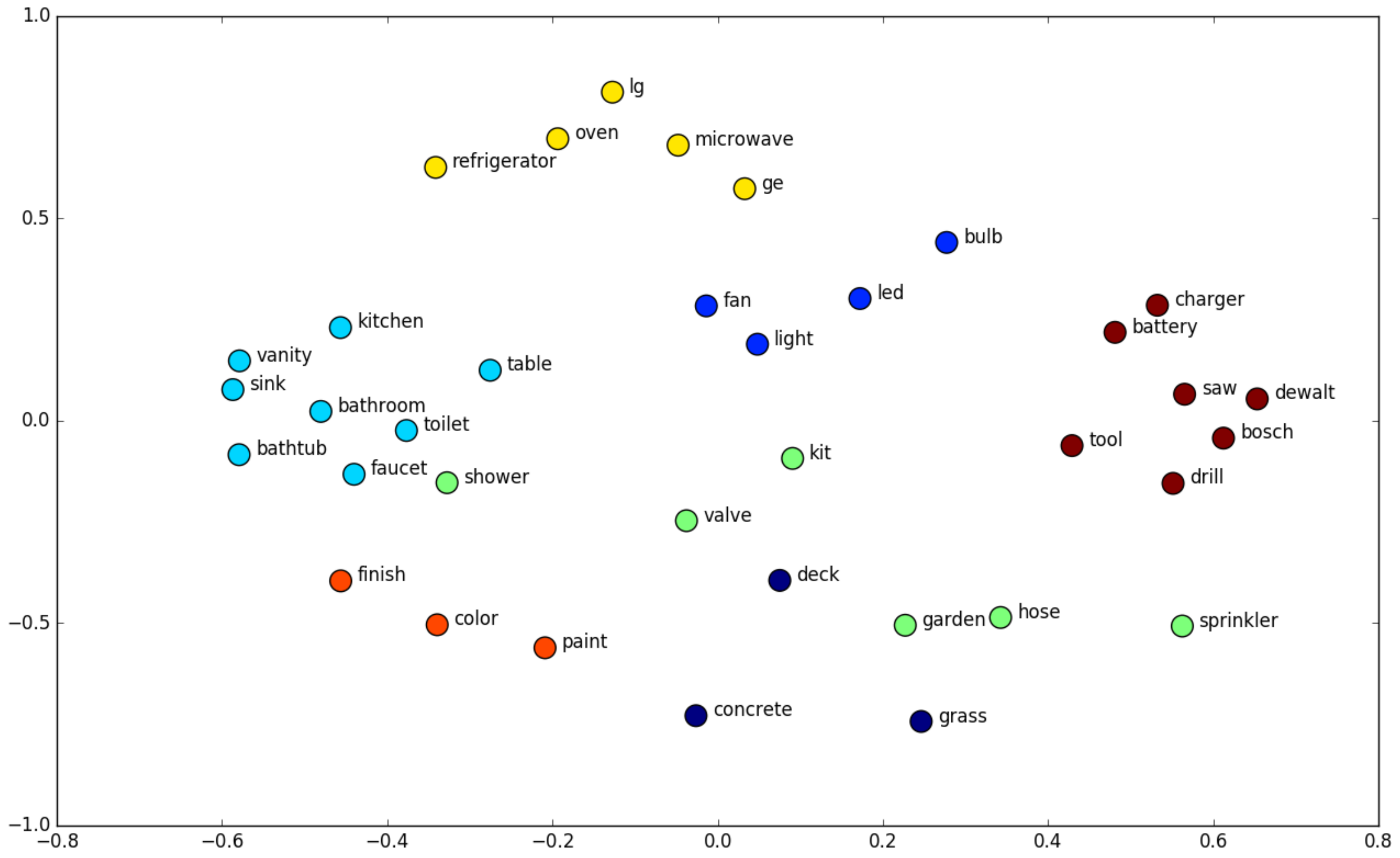


Hintergrund

Was sind Embedding und wie können sie hilfreich sein?



- **Definition:** Embeddings sind numerische Darstellungen von Objekten (wie Wörter, Sätze, Bilder) in einem kontinuierlichen Vektorraum.
- **Zweck:** Erfassung semantischer Beziehungen => ähnliche Objekte liegen nah beieinander
- **Berechnung:** Transformer basierte Sprachmodelle wie BERT
- **Verbreitung:** Google verwendet Embeddings seit 2018

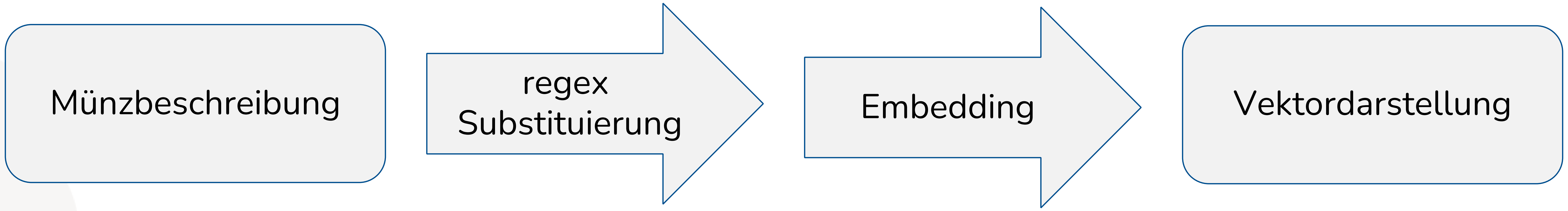


Pipeline

Part 1: Preprocessing, Part 2: Suchanfragen

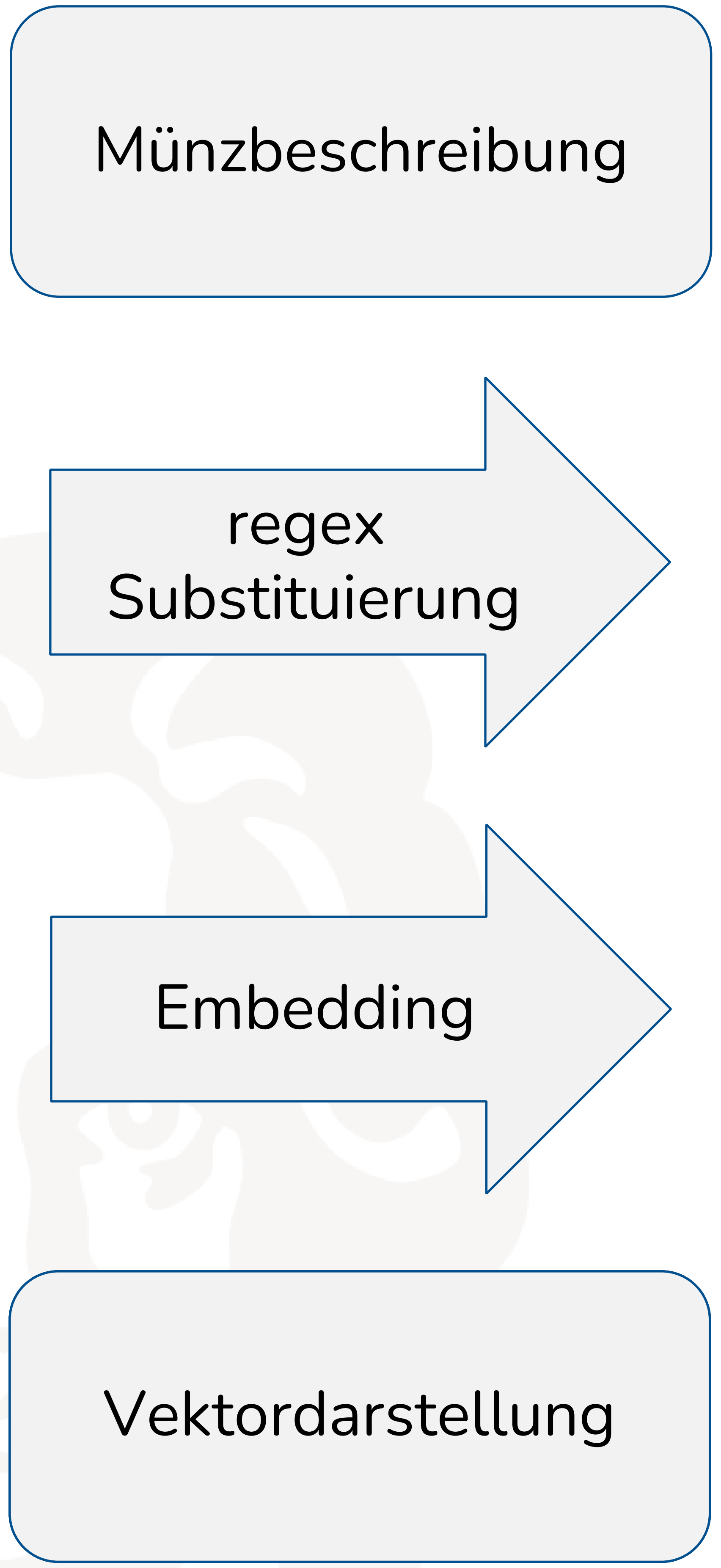
Pipeline

Part 1: Preprocessing



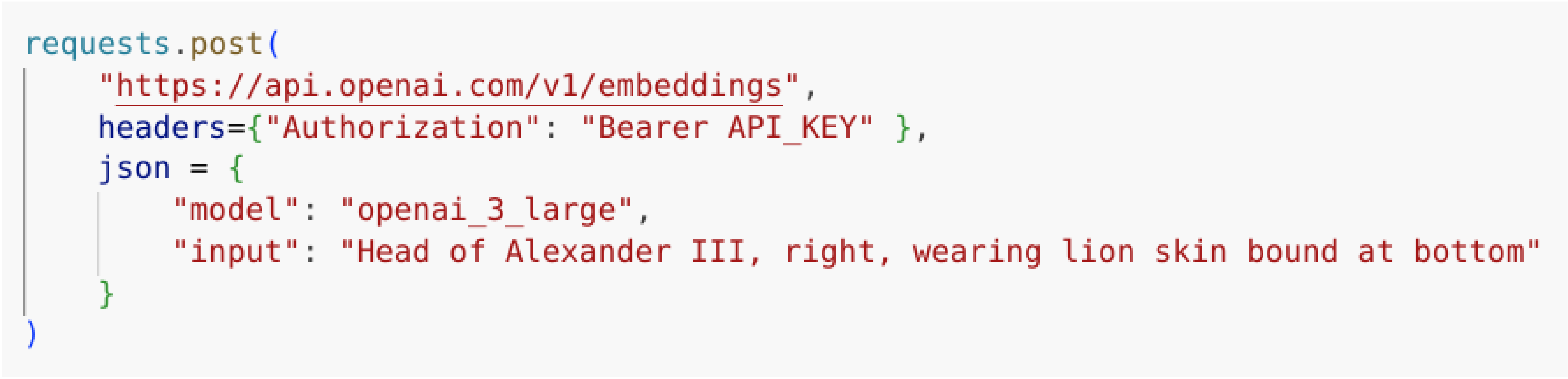
Pipeline

Part 1: Preprocessing



Head of Alexander the Great, right, wearing lion skin bound at bottom

Head of Alexander III, right, wearing lion skin bound at bottom

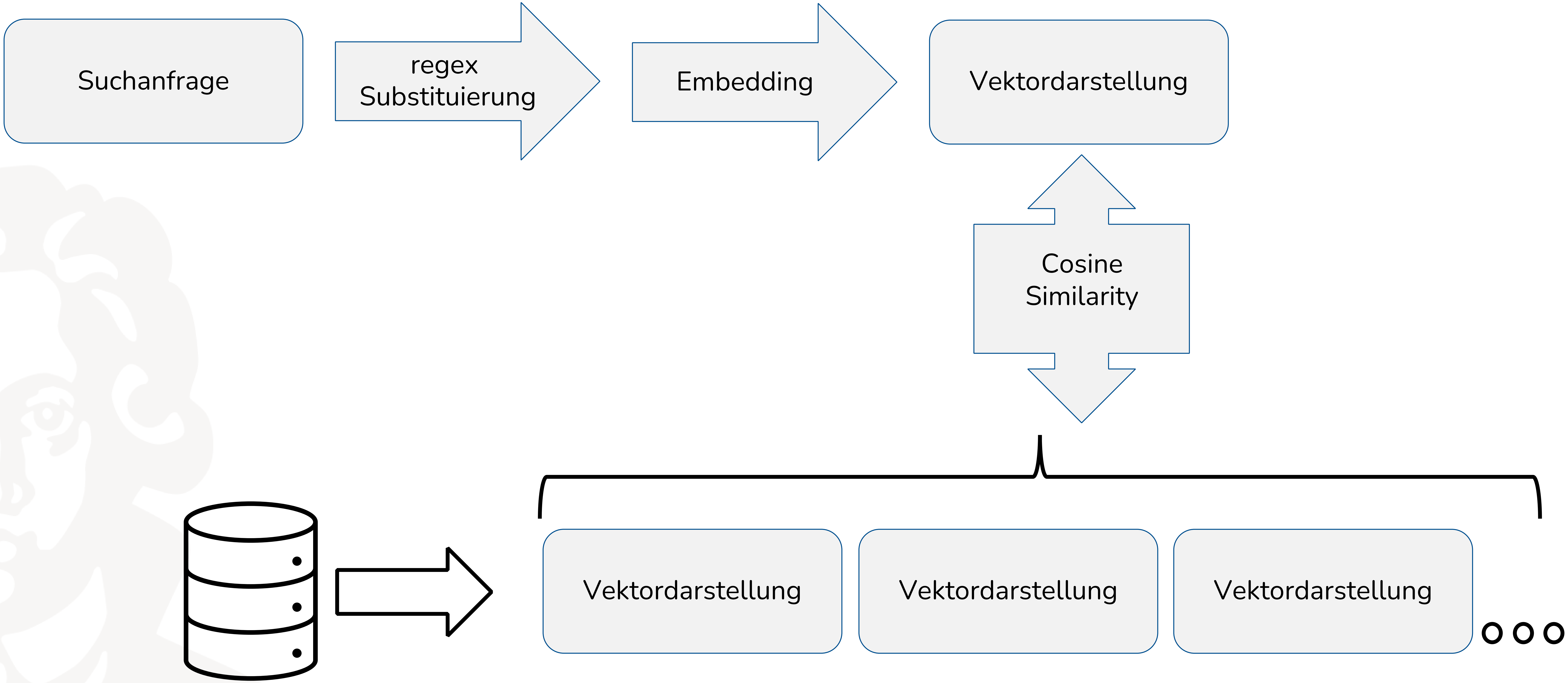


```
requests.post(
    "https://api.openai.com/v1/embeddings",
    headers={"Authorization": "Bearer API_KEY" },
    json = {
        "model": "openai_3_large",
        "input": "Head of Alexander III, right, wearing lion skin bound at bottom"
    }
)
```

```
array([
  0.01386353,  0.00959512, ..., -0.01570495, 0.00167738
])
length: 3072
```

Pipeline

Part 2: Suchanfrage

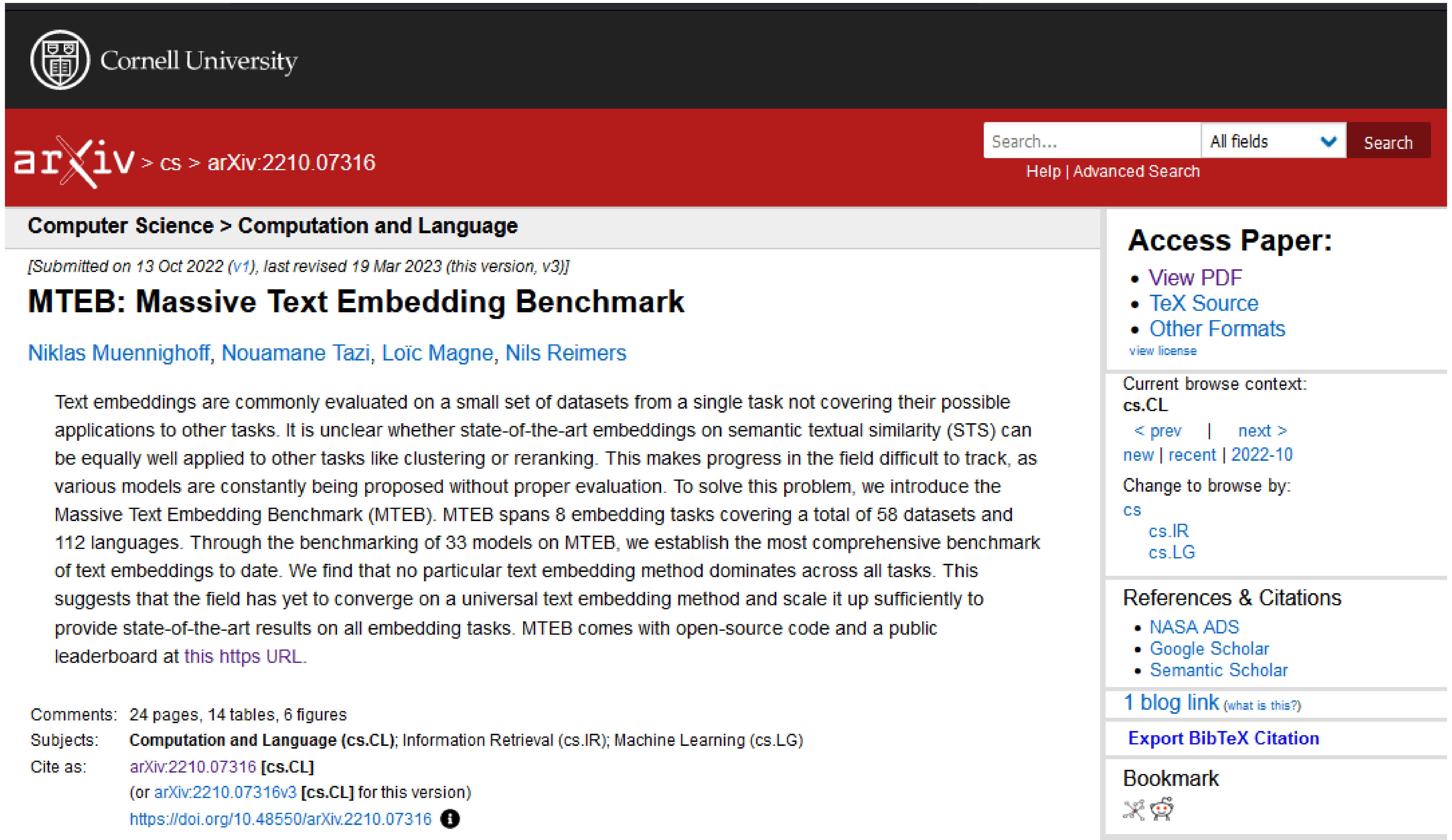


Previous Work

Massive Text Embedding Benchmark (MTEB)

Previous Work

Massive Text Embedding Benchmark (MTEB)



Computer Science > Computation and Language

[Submitted on 13 Oct 2022 (v1), last revised 19 Mar 2023 (this version, v3)]

MTEB: Massive Text Embedding Benchmark

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, Nils Reimers

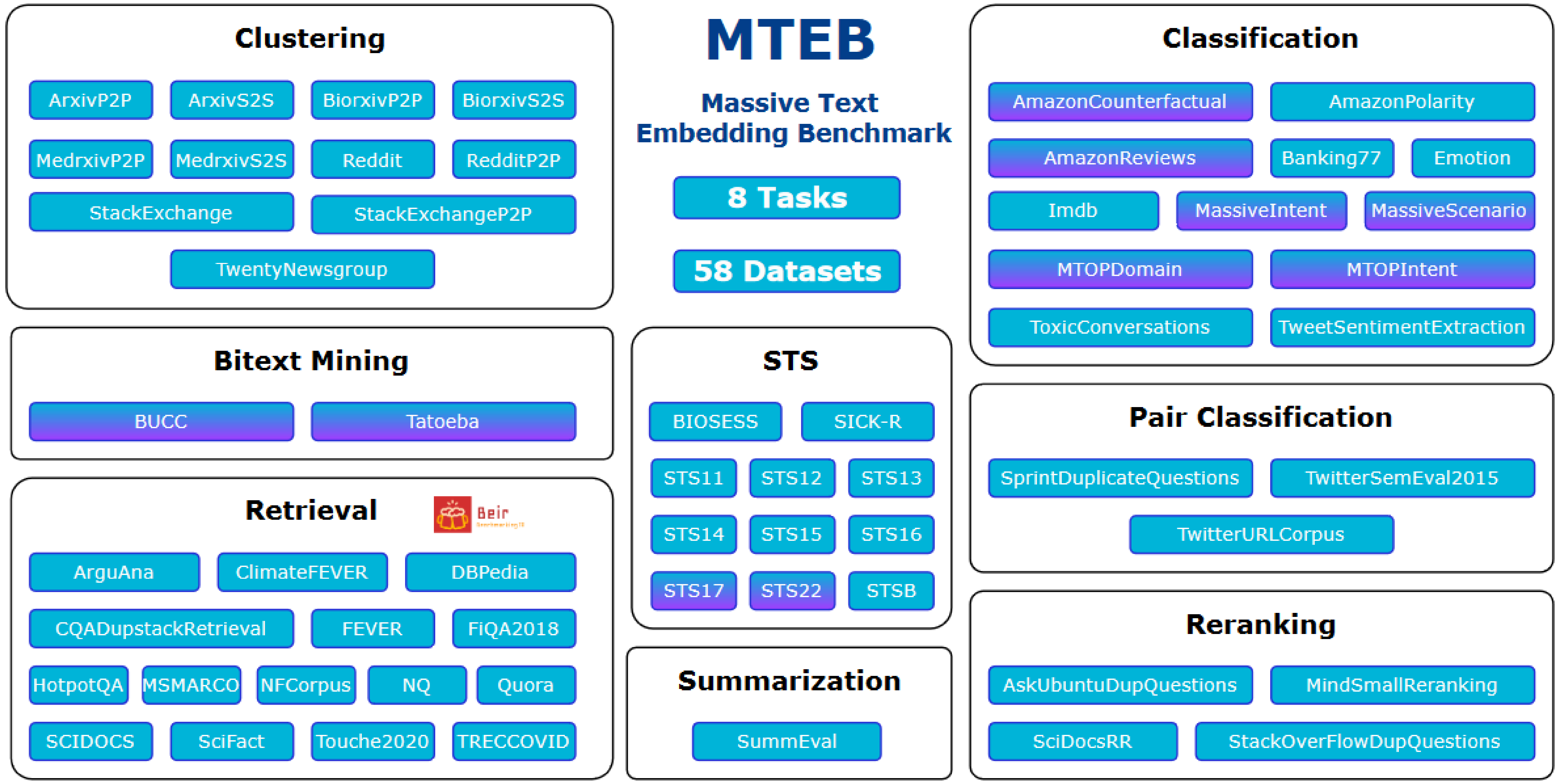
Text embeddings are commonly evaluated on a small set of datasets from a single task not covering their possible applications to other tasks. It is unclear whether state-of-the-art embeddings on semantic textual similarity (STS) can be equally well applied to other tasks like clustering or reranking. This makes progress in the field difficult to track, as various models are constantly being proposed without proper evaluation. To solve this problem, we introduce the Massive Text Embedding Benchmark (MTEB). MTEB spans 8 embedding tasks covering a total of 58 datasets and 112 languages. Through the benchmarking of 33 models on MTEB, we establish the most comprehensive benchmark of text embeddings to date. We find that no particular text embedding method dominates across all tasks. This suggests that the field has yet to converge on a universal text embedding method and scale it up sufficiently to provide state-of-the-art results on all embedding tasks. MTEB comes with open-source code and a public leaderboard at this https URL.

Comments:	24 pages, 14 tables, 6 figures
Subjects:	**Computation and Language (cs.CL)**; Information Retrieval (cs.IR); Machine Learning (cs.LG)
Cite as:	arXiv:2210.07316 **[cs.CL]** (or arXiv:2210.07316v3 **[cs.CL]** for this version) https://doi.org/10.48550/arXiv.2210.07316

Quelle: https://arxiv.org/abs/2210.07316

Previous Work

Massive Text Embedding Benchmark (MTEB)



MTEB

Massive Text Embedding Benchmark

8 Tasks

58 Datasets

Clustering

ArxivP2P ArxivS2S BiorxivP2P BiorxivS2S MedrxivP2P MedrxivS2S Reddit RedditP2P StackExchange StackExchangeP2P TwentyNewsgroup

Bitext Mining

BUCC Tatoeba

Retrieval

Beir

ArguAna ClimateFEVER DBPedia CQADupstackRetrieval FEVER FiQA2018 HotpotQA MSMARCO NFCorpus NQ Quora SCIDOCS SciFact Touche2020 TRECCOVID

STS

BIOSESS SICK-R STS11 STS12 STS13 STS14 STS15 STS16 STS17 STS22 STSB

Summarization

SummEval

Classification

AmazonCounterfactual AmazonPolarity AmazonReviews Banking77 Emotion Imdb MassiveIntent MassiveScenario MTOPDomain MTOPIntent ToxicConversations TweetSentimentExtraction

Pair Classification

SprintDuplicateQuestions TwitterSemEval2015 TwitterURLCorpus

Reranking

AskUbuntuDupQuestions MindSmallReranking SciDocsRR StackOverFlowDupQuestions

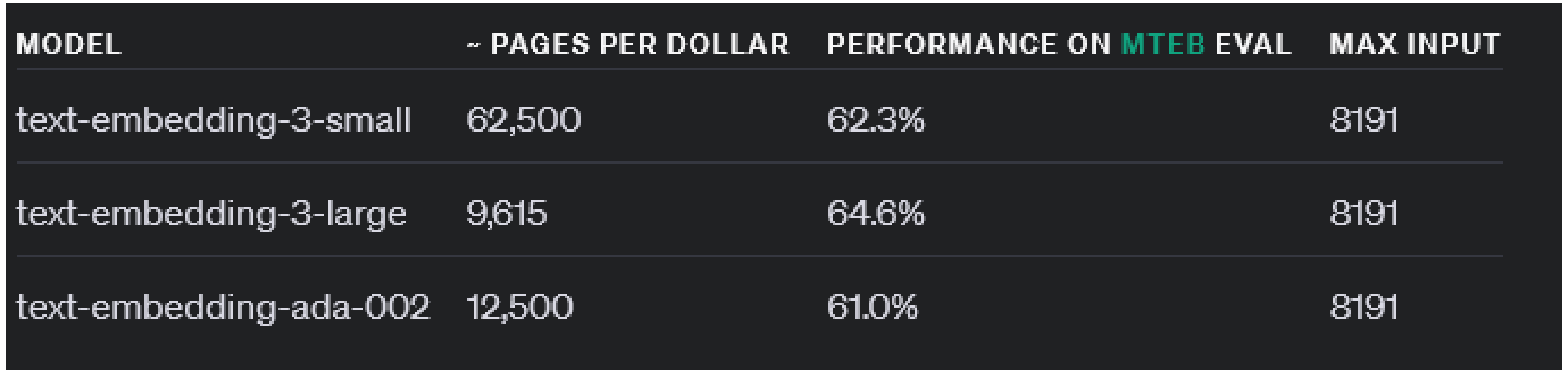
Figure 1: An overview of tasks and datasets in MTEB. Multilingual datasets are marked with a purple shade.

Modelle

Embedding 3	- OpenAI
MiniLM	- UKP Lab
MP Net	- Microsoft
BGE	- BAAI (Alibaba)
Embedding E5	- Microsoft
mistral-embed	- Mistral

Modelle

OpenAI - Embedding 3



MODEL	~ PAGES PER DOLLAR	PERFORMANCE ON MTEB EVAL	MAX INPUT
text-embedding-3-small	62,500	62.3%	8191
text-embedding-3-large	9,615	64.6%	8191
text-embedding-ada-002	12,500	61.0%	8191

Embedding-Größe: small - **1536**
large - **3072**

Parameter: ? (GPT-3: 175B)

Lizenz: closed-source - Zugriff nur durch OpenAI API

Trainingsdaten?

Modelle

UKP Lab - MiniLM

Entstanden aus einer Forschungsarbeit von Microsoft[1]

Weiterentwickelt durch UKP Lab (Darmstadt)
Huggingface Challange:
Train the Best Sentence Embedding Model
*Ever with **1B Training Pairs***

UBIQUITOUS KNOWLEDGE PROCESSING

Embedding-Größe: **384**
Parameter: **33M**

Trainingsdaten: Reddit Comments, WikiAnswers, Stack Exchange

Lizenz: Apache-2.0 (Open-Source)

MTEB Score: 112

[1] Wenhui Wang, Furu Wie (2020) MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers

Modelle

Microsoft - MPNet

Platz 1 in MTEB (Okt 2022)
MTEP Score (heute): **103**

Embedding-Größe:	**768**
Parameter:	**109M**

Trainingsdaten: Sentiment 140,
European Parliament Proceedings Parallel Corpus
...

Lizenz: Apache-2.0 (Open-Source)

[1] Niklas Muennighoff, Nouamane Taz (2022) MTEB: Massive Text Embedding Benchmark

Modelle

Alibaba (Beijing Academy of Artificial Intelligence) - BGE

MTEP Score (heute): **28**

Embedding-Größe: **768**
Parameter: **335M**

Trainingsdaten: Sentiment 140,
European Parliament Proceedings Parallel Corpus
…

Lizenz: Apache-2.0 (Open-Source)

[1] Niklas Muennighoff, Nouamane Taz (2022) MTEB: Massive Text Embedding Benchmark

Modelle

Microsoft - Embedding E5

MTEP Score (heute): **68**

Embedding-Größe:	**1024**
Parameter:	**335M**

Trainingsdaten: CCPairs

Lizenz: Apache-2.0 (Open-Source)

[1] Niklas Muennighoff, Nouamane Taz (2022) MTEB: Massive Text Embedding Benchmark

Modelle

Mistral - mistral-embed

MTEP Score (heute): **55**

Embedding-Größe: **1024**
Parameter: ?

Trainingsdaten: ?
Lizenz: closed-source

Auswertung

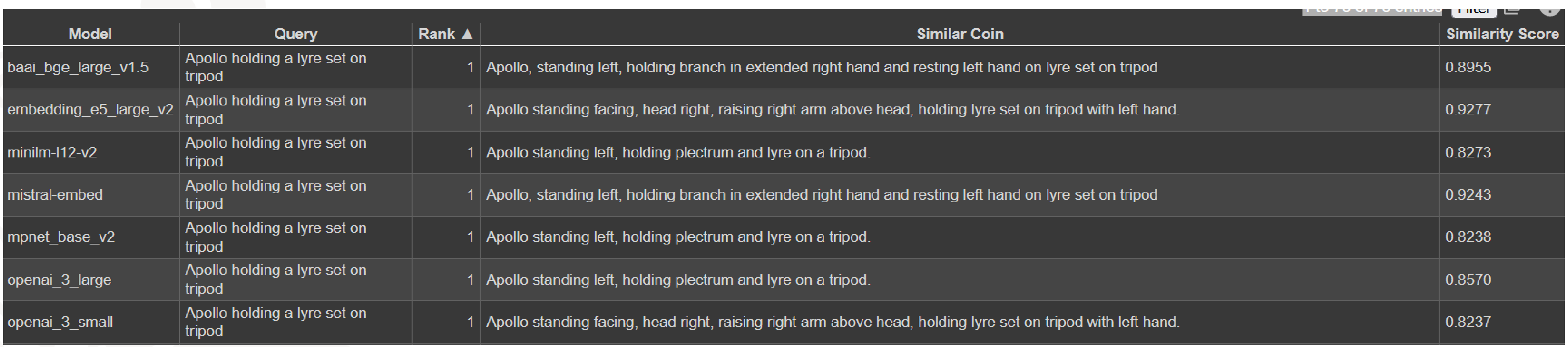
Beispiele, Benchmarks, Visualisierungen

Auswertung

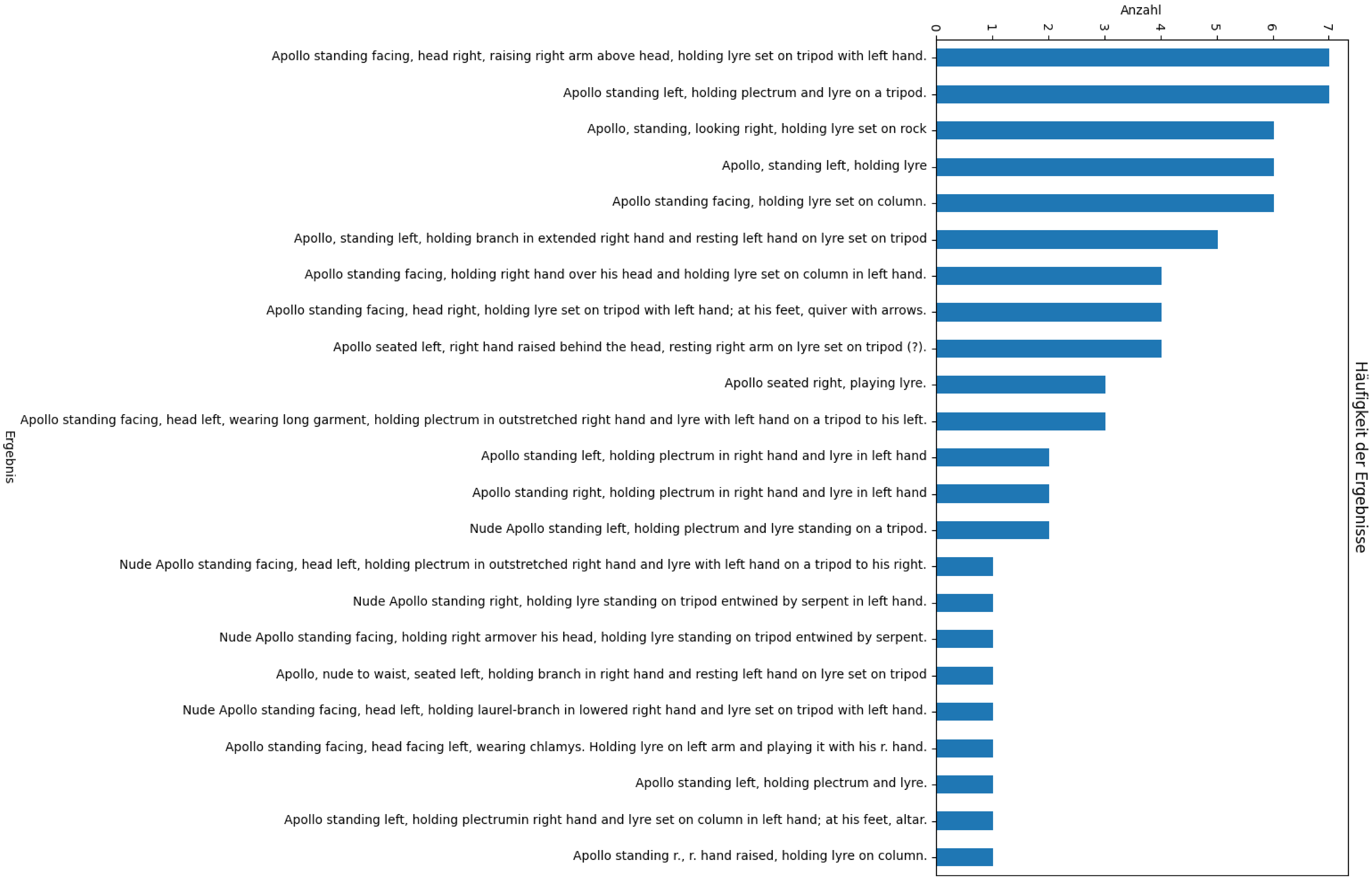
Beispiele

Prompt: Apollo Holding Lyre

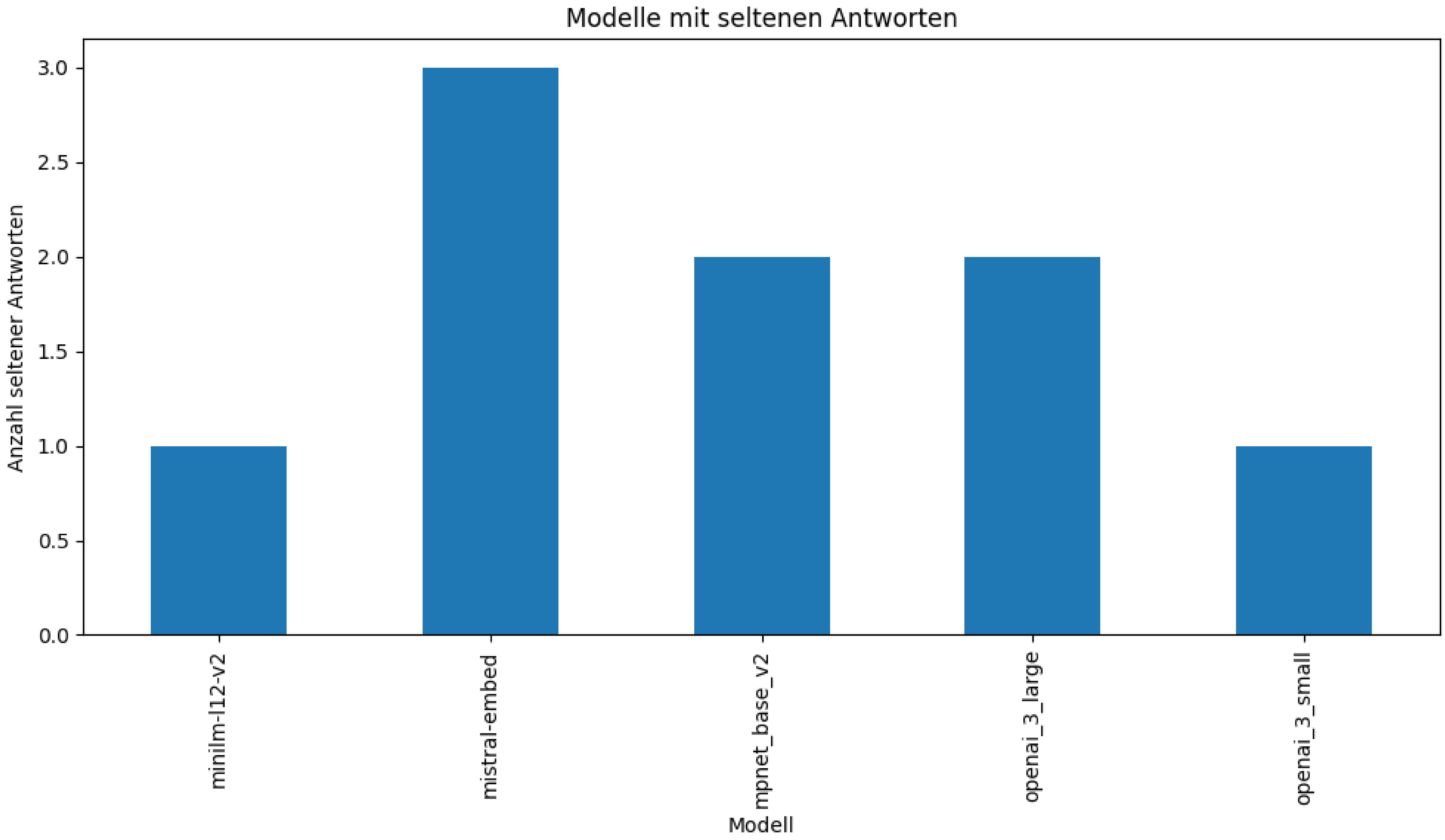
Score: alle Top-5 Ergebnisse richtig



Model	Query	Rank ▲	Similar Coin	Similarity Score
baai_bge_large_v1.5	Apollo holding a lyre set on tripod	1	Apollo, standing left, holding branch in extended right hand and resting left hand on lyre set on tripod	0.8955
embedding_e5_large_v2	Apollo holding a lyre set on tripod	1	Apollo standing facing, head right, raising right arm above head, holding lyre set on tripod with left hand.	0.9277
minilm-l12-v2	Apollo holding a lyre set on tripod	1	Apollo standing left, holding plectrum and lyre on a tripod.	0.8273
mistral-embed	Apollo holding a lyre set on tripod	1	Apollo, standing left, holding branch in extended right hand and resting left hand on lyre set on tripod	0.9243
mpnet_base_v2	Apollo holding a lyre set on tripod	1	Apollo standing left, holding plectrum and lyre on a tripod.	0.8238
openai_3_large	Apollo holding a lyre set on tripod	1	Apollo standing left, holding plectrum and lyre on a tripod.	0.8570
openai_3_small	Apollo holding a lyre set on tripod	1	Apollo standing facing, head right, raising right arm above head, holding lyre set on tripod with left hand.	0.8237


Häufigkeit der Ergebnisse
Anzahl
0
1
2
3
4
5
6
7
Ergebnis
Apollo standing facing, head right, raising right arm above head, holding lyre set on tripod with left hand.
Apollo standing left, holding plectrum and lyre on a tripod.
Apollo, standing, looking right, holding lyre set on rock
Apollo, standing left, holding lyre
Apollo standing facing, holding lyre set on column.
Apollo, standing left, holding branch in extended right hand and resting left hand on lyre set on tripod
Apollo standing facing, holding right hand over his head and holding lyre set on column in left hand.
Apollo standing facing, head right, holding lyre set on tripod with left hand; at his feet, quiver with arrows.
Apollo seated left, right hand raised behind the head, resting right arm on lyre set on tripod (?).
Apollo seated right, playing lyre.
Apollo standing facing, head left, wearing long garment, holding plectrum in outstretched right hand and lyre with left hand on a tripod to his left.
Apollo standing left, holding plectrum in right hand and lyre in left hand
Apollo standing right, holding plectrum in right hand and lyre in left hand
Nude Apollo standing left, holding plectrum and lyre standing on a tripod.
Nude Apollo standing facing, head left, holding plectrum in outstretched right hand and lyre with left hand on a tripod to his right.
Nude Apollo standing right, holding lyre standing on tripod entwined by serpent in left hand.
Nude Apollo standing facing, holding right armover his head, holding lyre standing on tripod entwined by serpent.
Apollo, nude to waist, seated left, holding branch in right hand and resting left hand on lyre set on tripod
Nude Apollo standing facing, head left, holding laurel-branch in lowered right hand and lyre set on tripod with left hand.
Apollo standing facing, head facing left, wearing chlamys. Holding lyre on left arm and playing it with his r. hand.
Apollo standing left, holding plectrum and lyre.
Apollo standing left, holding plectrumin right hand and lyre set on column in left hand; at his feet, altar.
Apollo standing r., r. hand raised, holding lyre on column.

Test - Apollo Holding Lyre



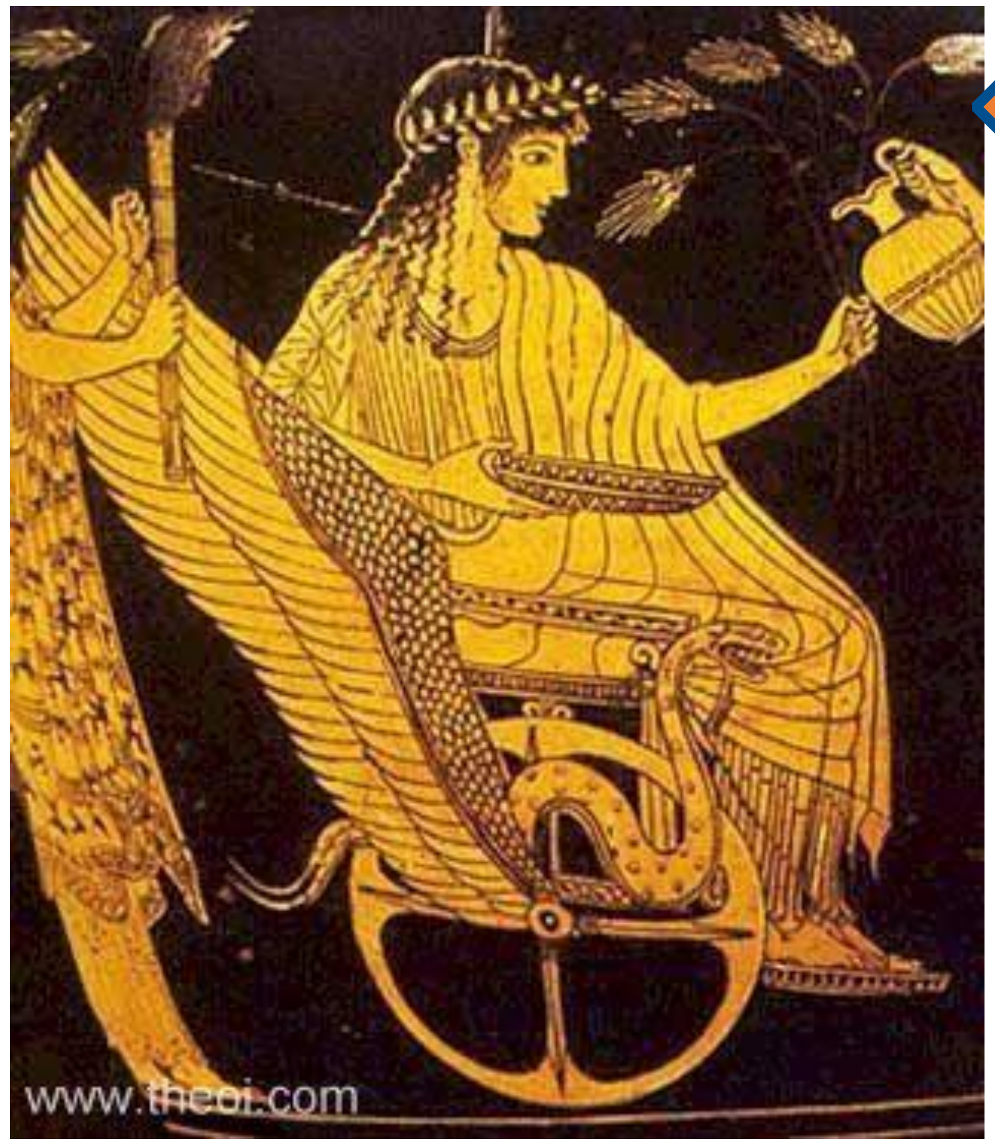
Auswertung

Beispiele

Prompt: Demeter holding a grain ear

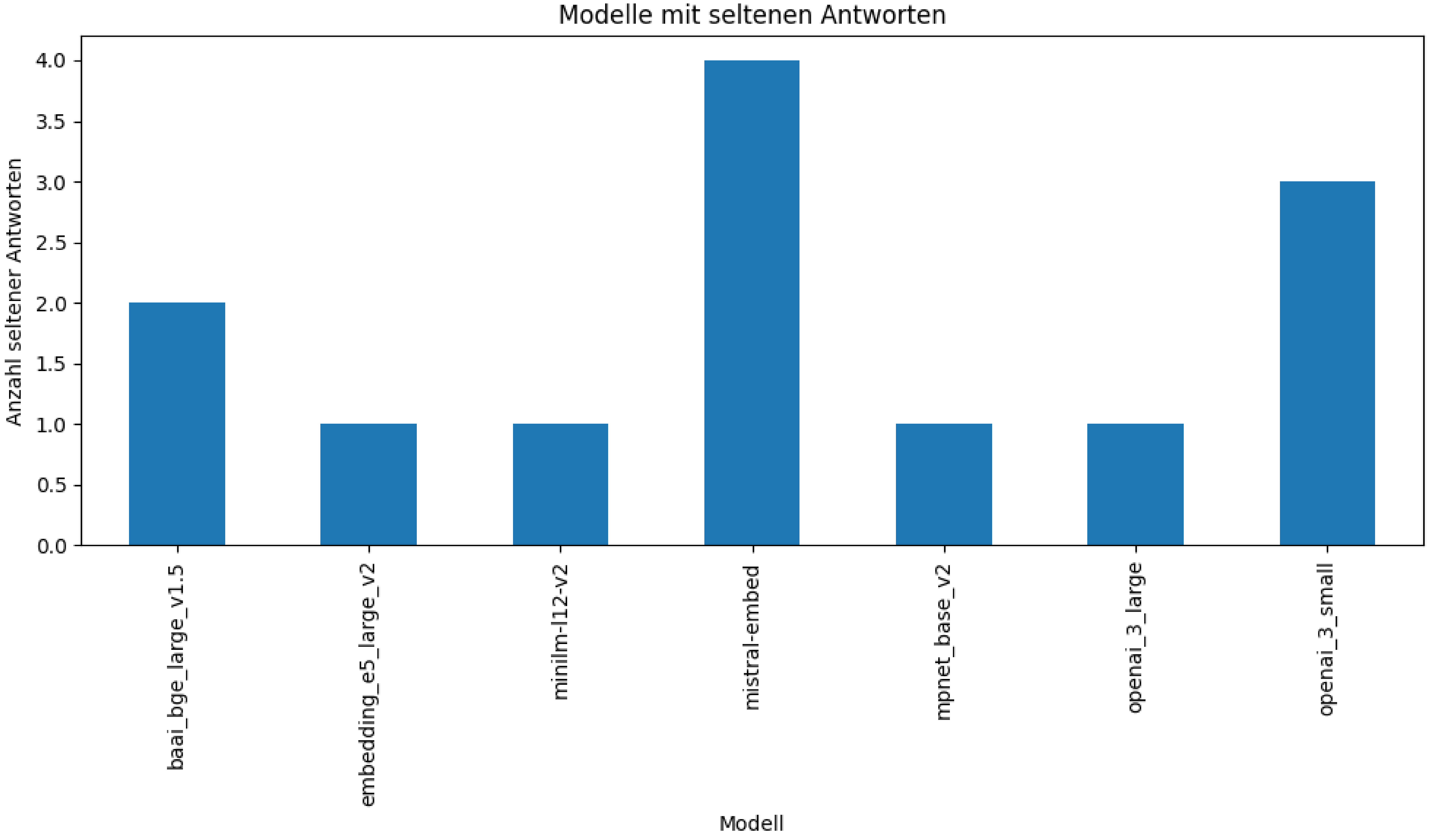
Die Beschreibung mit grain ear kommt nur einmal im Datensatz vor.

Score: Mini-LM: 3
OpenAI large: 2
Mistral: garnicht



Model	Query	Rank ▲	
baai_bge_large_v1.5	Demeter holding a grain gar	1	Demeter standing left, holding ears of corn in her r. hand.
embedding_e5_large_v2	Demeter holding a grain gar	1	Demeter (or Kore?) with a wreath of grain-ears crouching left above a tunny left, holding two ears of grain in her raised right hand ar
minilm-l12-v2	Demeter holding a grain gar	1	Demeter standing facing, head left, wearing long garment, holding ears of corn in raised right hand and long torch in raised left arm
mistral-embed	Demeter holding a grain gar	1	Demeter ? standing left, holding patera and uncertain object.
mpnet_base_v2	Demeter holding a grain gar	1	Head of Demeter, right, wearing earrings and necklace; below, grain.
openai_3_large	Demeter holding a grain gar	1	Demeter standing right, holding basket with ears of corn.
openai_3_small	Demeter holding a grain gar	1	Demeter standing right, holding basket with ears of corn.

Auswertung

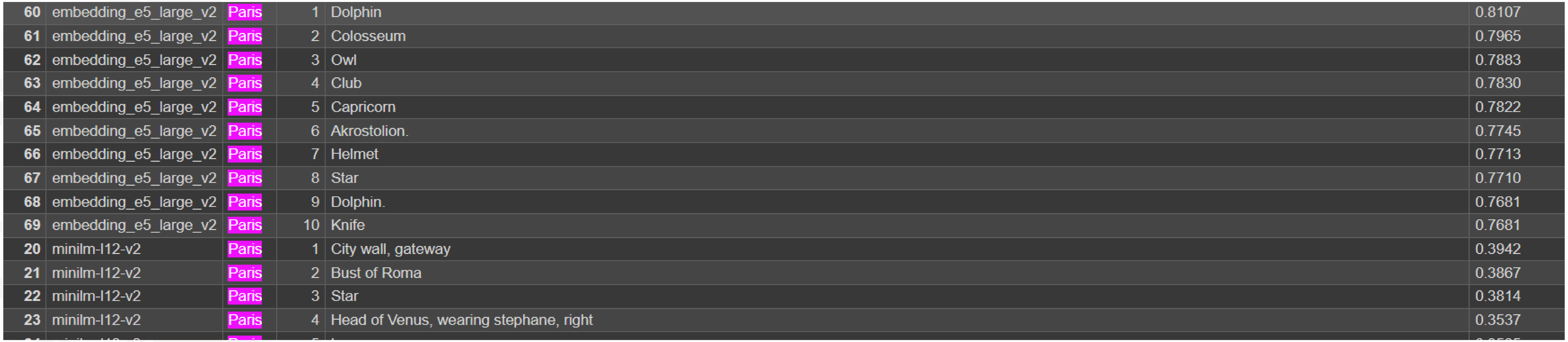

Modelle mit seltenen Antworten
Anzahl seltener Antworten
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
baai_bge_large_v1.5
embedding_e5_large_v2
minilm-l12-v2
mistral-embed
mpnet_base_v2
openai_3_large
openai_3_small
Modell

Auswertung

Beispiele

Prompt: Paris

Paris kommt nur einmal im Datensatz vor.



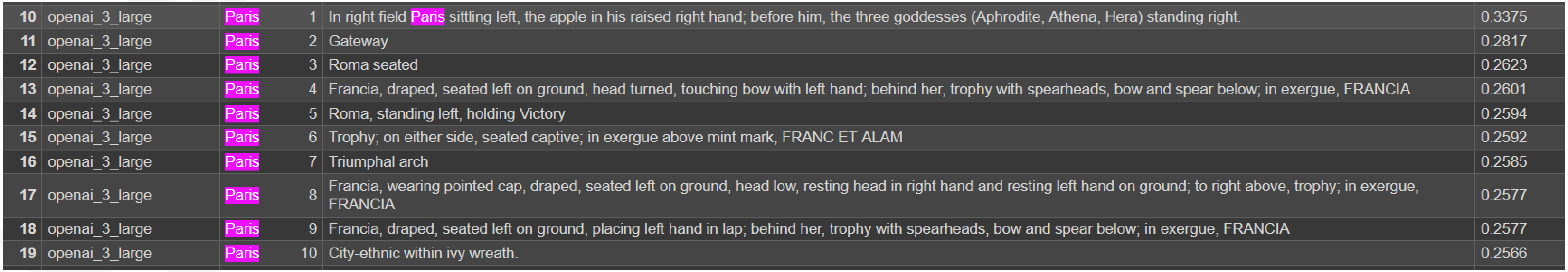
60	embedding_e5_large_v2	Paris	1	Dolphin	0.8107
61	embedding_e5_large_v2	Paris	2	Colosseum	0.7965
62	embedding_e5_large_v2	Paris	3	Owl	0.7883
63	embedding_e5_large_v2	Paris	4	Club	0.7830
64	embedding_e5_large_v2	Paris	5	Capricorn	0.7822
65	embedding_e5_large_v2	Paris	6	Akrostolion.	0.7745
66	embedding_e5_large_v2	Paris	7	Helmet	0.7713
67	embedding_e5_large_v2	Paris	8	Star	0.7710
68	embedding_e5_large_v2	Paris	9	Dolphin.	0.7681
69	embedding_e5_large_v2	Paris	10	Knife	0.7681
20	minilm-l12-v2	Paris	1	City wall, gateway	0.3942
21	minilm-l12-v2	Paris	2	Bust of Roma	0.3867
22	minilm-l12-v2	Paris	3	Star	0.3814
23	minilm-l12-v2	Paris	4	Head of Venus, wearing stephane, right	0.3537

Auswertung

Beispiele

Prompt: Paris

Paris kommt nur einmal im Datensatz vor.



10	openai_3_large	Paris	1	In right field Paris sittling left, the apple in his raised right hand; before him, the three goddesses (Aphrodite, Athena, Hera) standing right.	0.3375
11	openai_3_large	Paris	2	Gateway	0.2817
12	openai_3_large	Paris	3	Roma seated	0.2623
13	openai_3_large	Paris	4	Francia, draped, seated left on ground, head turned, touching bow with left hand; behind her, trophy with spearheads, bow and spear below; in exergue, FRANCIA	0.2601
14	openai_3_large	Paris	5	Roma, standing left, holding Victory	0.2594
15	openai_3_large	Paris	6	Trophy; on either side, seated captive; in exergue above mint mark, FRANC ET ALAM	0.2592
16	openai_3_large	Paris	7	Triumphal arch	0.2585
17	openai_3_large	Paris	8	Francia, wearing pointed cap, draped, seated left on ground, head low, resting head in right hand and resting left hand on ground; to right above, trophy; in exergue, FRANCIA	0.2577
18	openai_3_large	Paris	9	Francia, draped, seated left on ground, placing left hand in lap; behind her, trophy with spearheads, bow and spear below; in exergue, FRANCIA	0.2577
19	openai_3_large	Paris	10	City-ethnic within ivy wreath.	0.2566

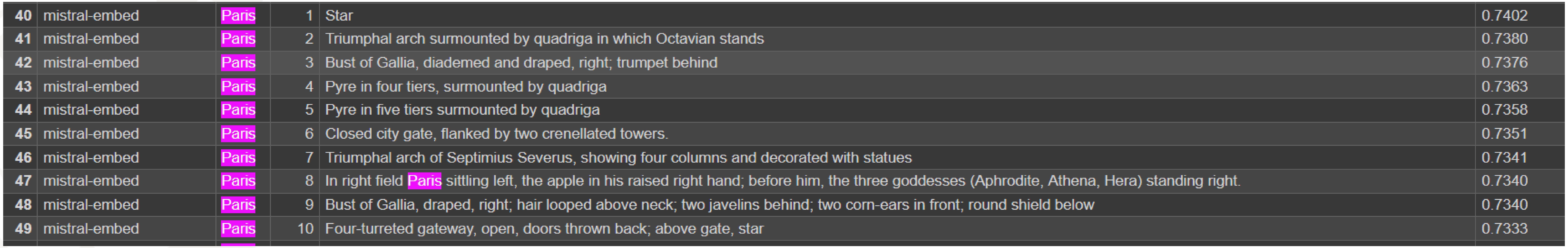
Auswertung

Beispiele

Prompt: Paris

Paris kommt nur einmal im Datensatz vor.

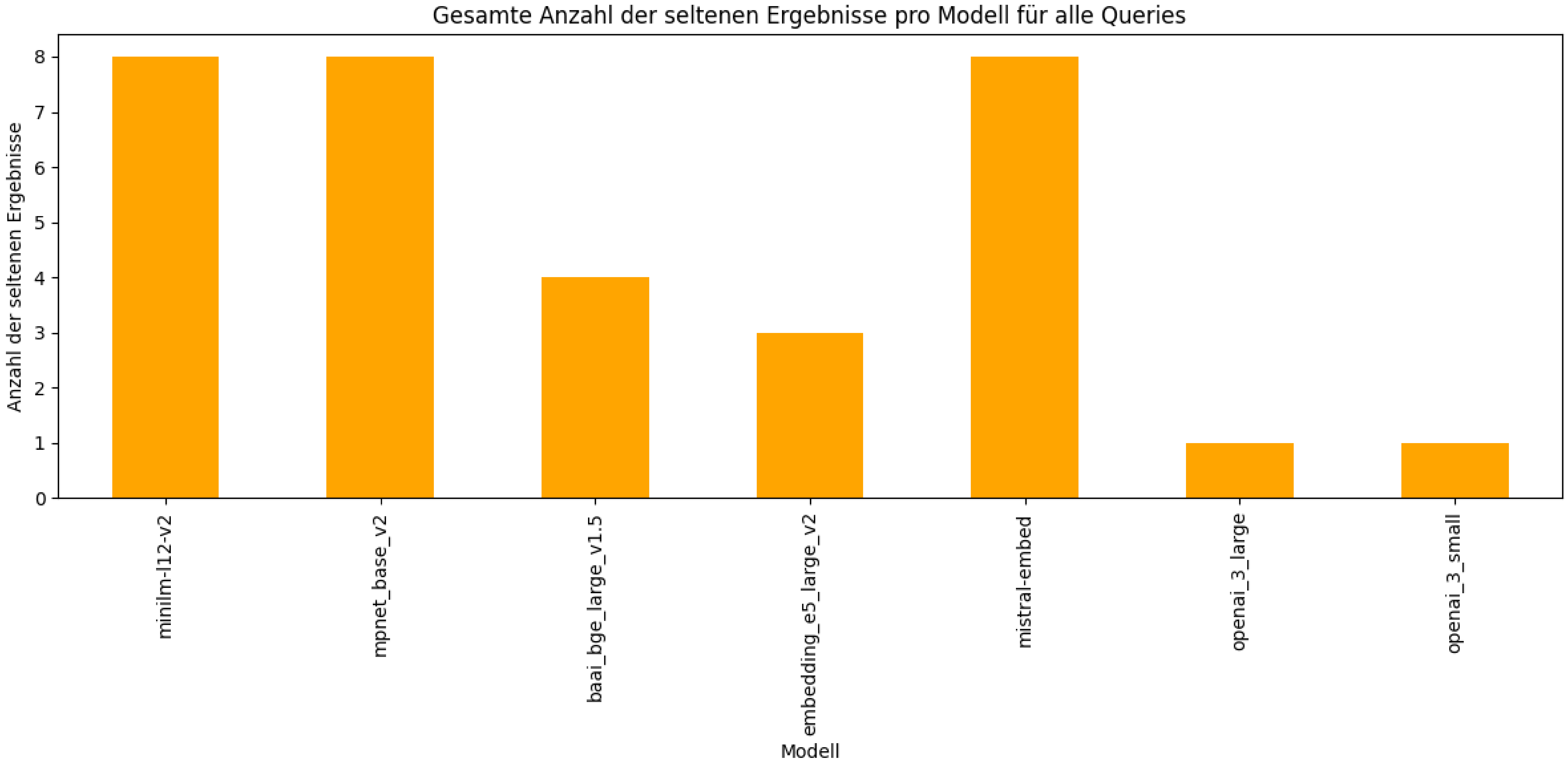
Score: Top-1 richtige Antwort: OpenAI large
Top-10 richtige Antwort: OpenAI small, Mistral



40	mistral-embed	Paris	1	Star	0.7402
41	mistral-embed	Paris	2	Triumphal arch surmounted by quadriga in which Octavian stands	0.7380
42	mistral-embed	Paris	3	Bust of Gallia, diademed and draped, right; trumpet behind	0.7376
43	mistral-embed	Paris	4	Pyre in four tiers, surmounted by quadriga	0.7363
44	mistral-embed	Paris	5	Pyre in five tiers surmounted by quadriga	0.7358
45	mistral-embed	Paris	6	Closed city gate, flanked by two crenellated towers.	0.7351
46	mistral-embed	Paris	7	Triumphal arch of Septimius Severus, showing four columns and decorated with statues	0.7341
47	mistral-embed	Paris	8	In right field Paris sittling left, the apple in his raised right hand; before him, the three goddesses (Aphrodite, Athena, Hera) standing right.	0.7340
48	mistral-embed	Paris	9	Bust of Gallia, draped, right; hair looped above neck; two javelins behind; two corn-ears in front; round shield below	0.7340
49	mistral-embed	Paris	10	Four-turreted gateway, open, doors thrown back; above gate, star	0.7333

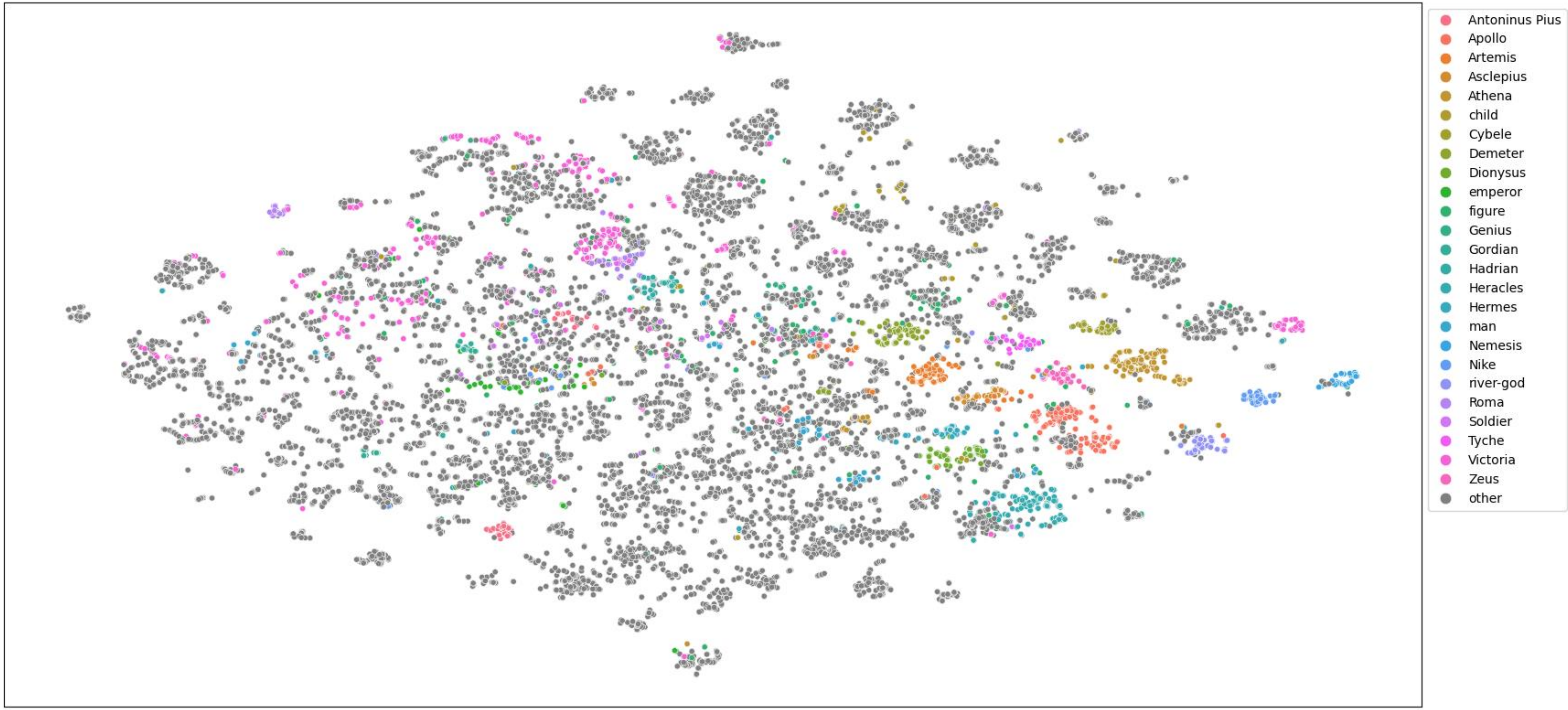
Auswertung

Zwischenstand – nach 20 Queries



Auswertung

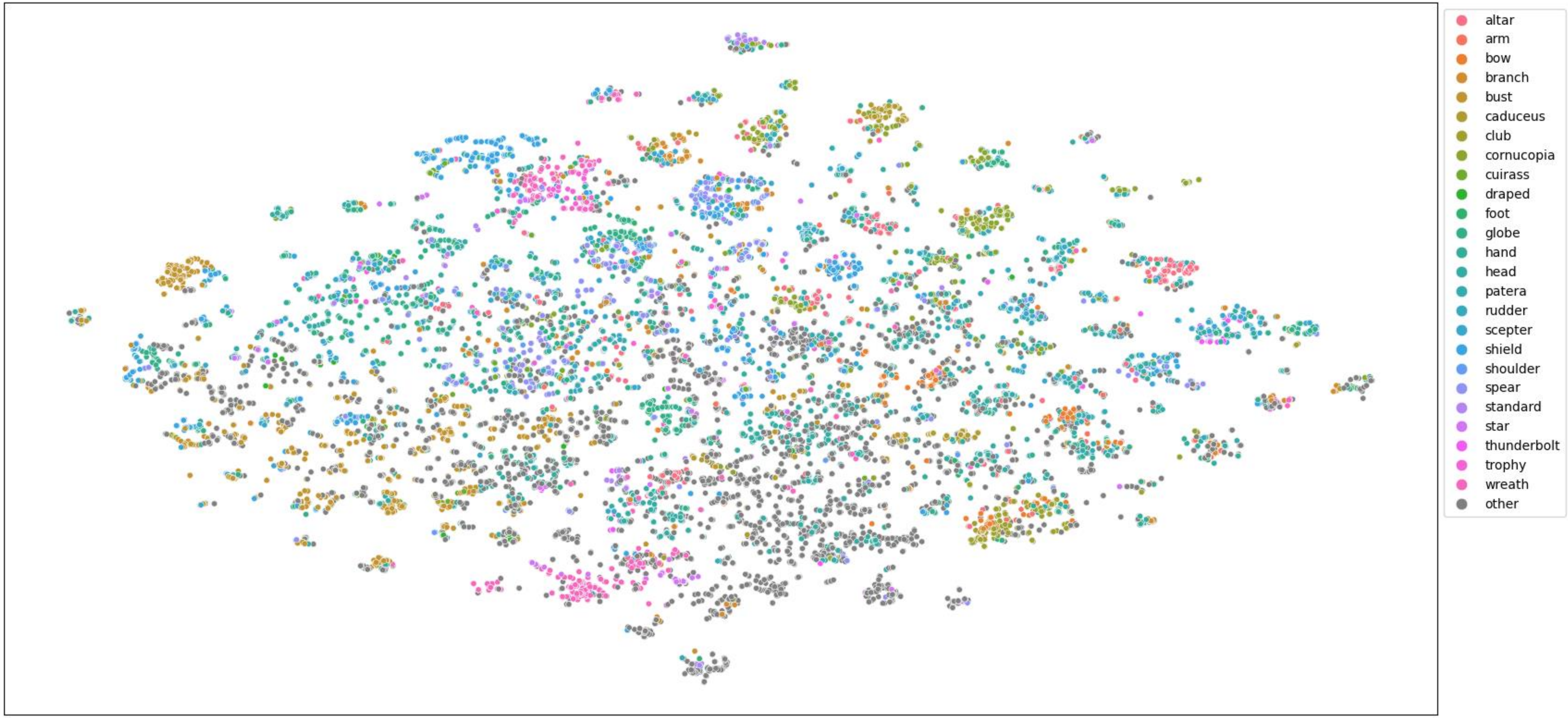
Visualisierungen



OpenAI 3 Large – 25 meist erwähnten Personen

Auswertung

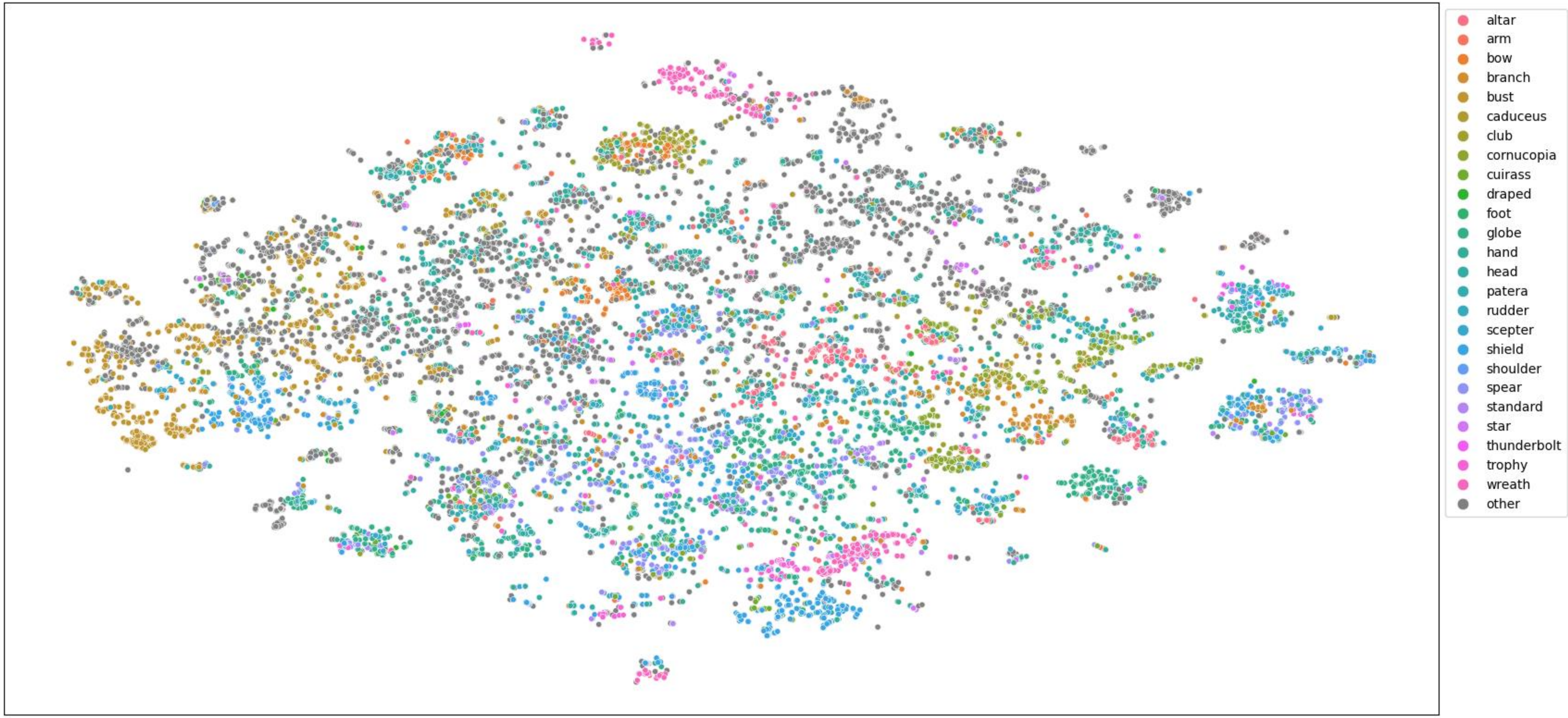
Visualisierungen



OpenAI 3 Large – 25 meist erwähnten Objekte

Auswertung

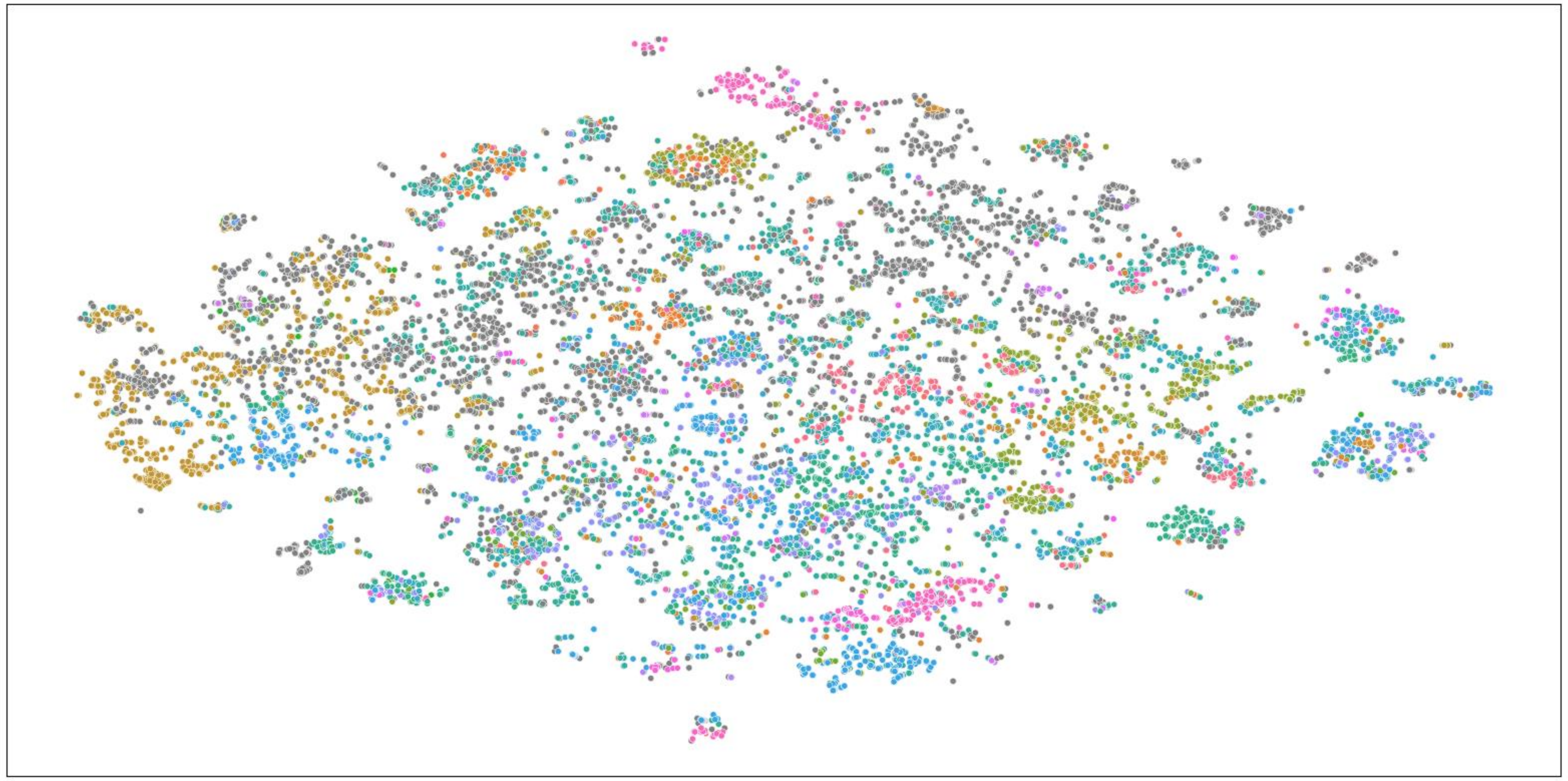
Visualisierungen



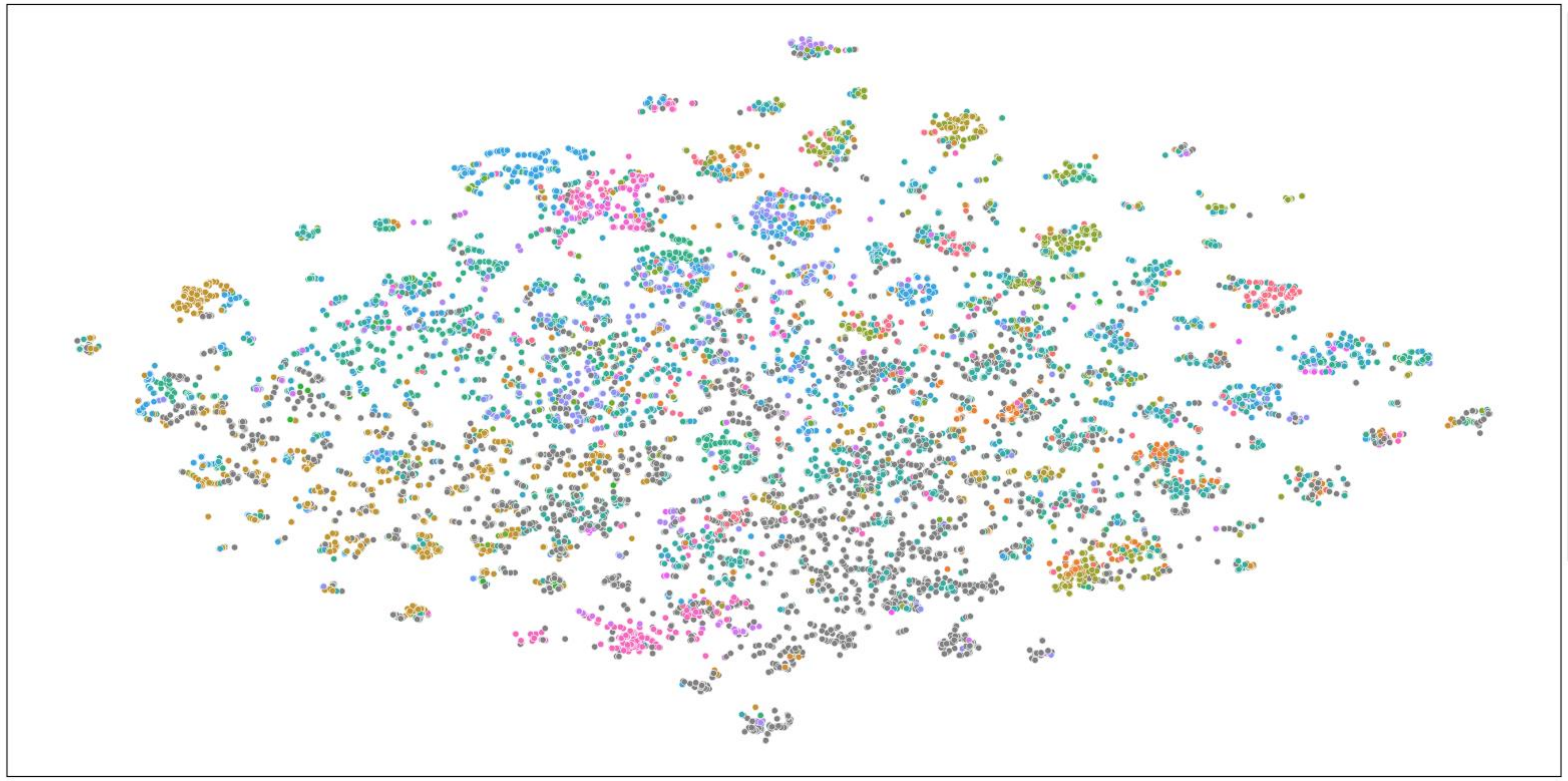
MPnet Base v2 – 25 meist erwähnten Objekte

Auswertung

Visualisierungen



MPnet Base v2 – 25 meist erwähnten Objekte



OpenAI 3 Large – 25 meist erwähnten Objekte

Demo

Live Demonstration – hat jemand Vorschläge zu Suchanfragen?

Fazit

Was lief gut? Was lief schlecht?

Fazit

Was lief gut? Was lief schlecht?

- **Semantische Suche** funktioniert gut, Synonyme geben auch Ergebnisse
- Extrahierung von Kombinationen **mehrerer, seltener Schlagworte** eher schlecht
- **OpenAI 3 Large** bietet das für uns beste Ergebnis
  - Weitere Arbeit zu großen Embedding-Modellen möglich / nötig
- Integration relativ einfach möglich
- Kosten für Berechnung des gesamten Datensatzes:
  - **OpenAI 3 Large:** 14 cent
  - **Microsoft e5-large-v2:** 14 cent
  - **Mistral-Embed:** 17 cent