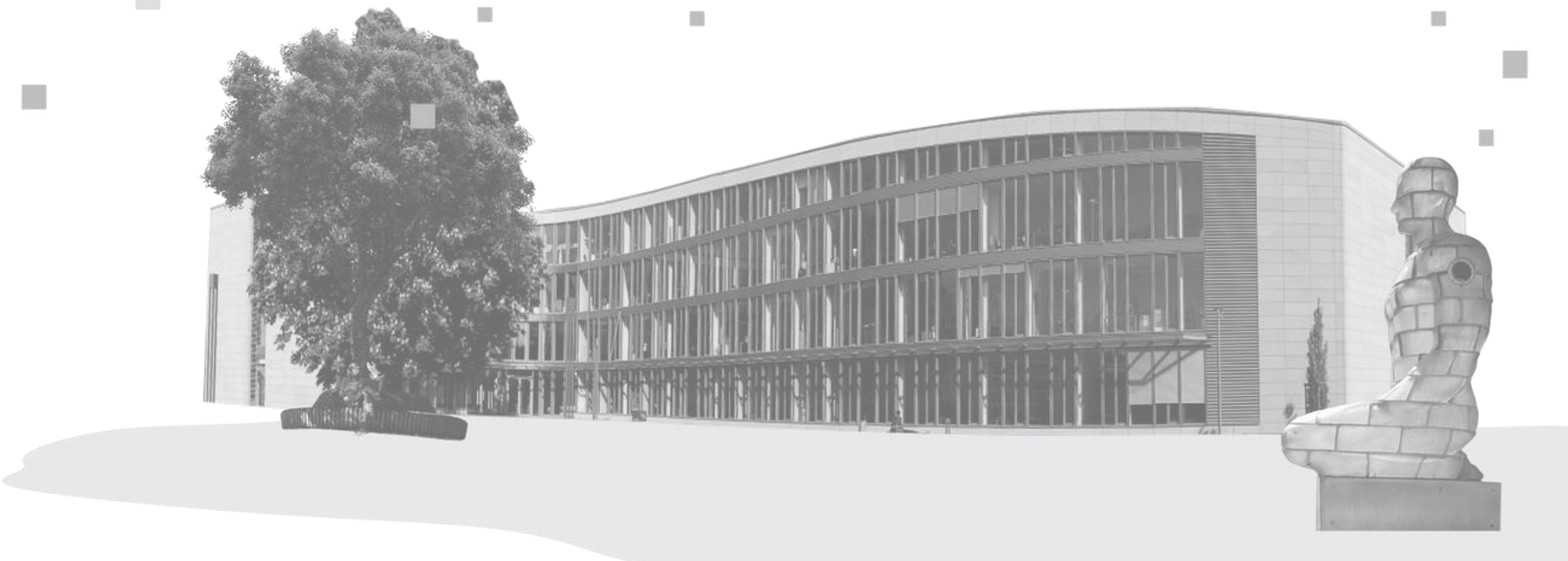
Felix Bölter
Hanno Müller

KI Service Zentrum
by Hasso-Plattner-Institut
HPI

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# Effizientes Abfragen von Informationen aus Dokumenten.

## Lokales Retrieval Augmented Generation System



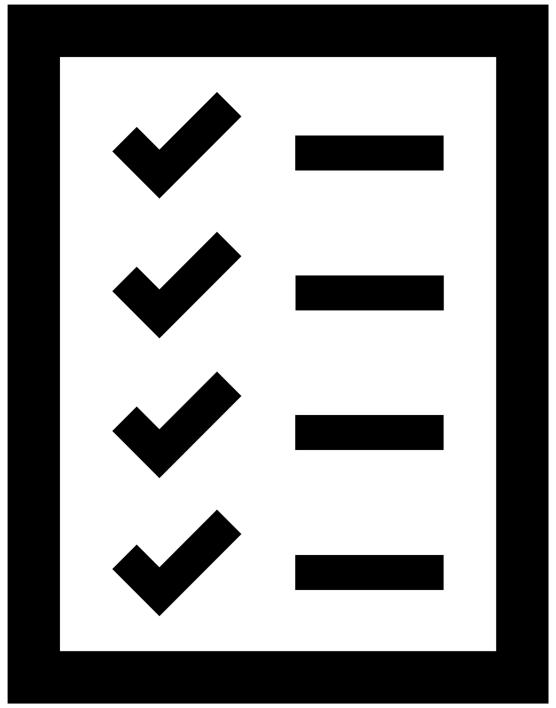
**Design IT.**
**Create Knowledge.**

www.hpi.de

# Attendance List

KI Service Zentrum
by Hasso-Plattner-Institut

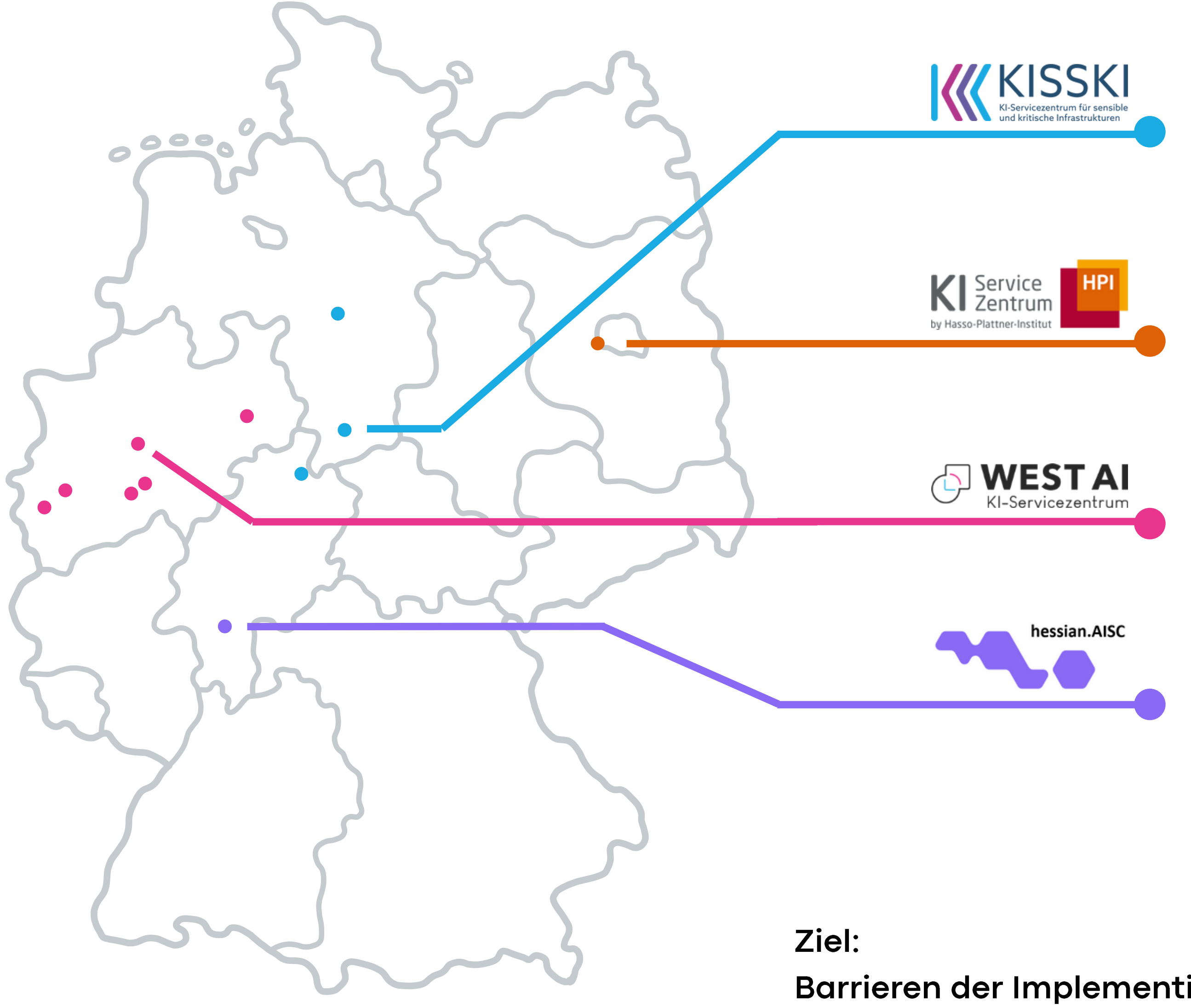
kisz@hpi.de
hpi.de/kisz



GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

# KI-Servicezentrum Berlin-Brandenburg

kisz@hpi.de
hpi.de/kisz



KISSKI

**Göttingen I Kassel I Hannover**
Sensible und kritische Infrastrukturen
Medizin & Energie

**Potsdam | Berlin**
Bildungs- und Beratungsangebote, Einsatz von KI in Wirtschaft & Gesellschaft

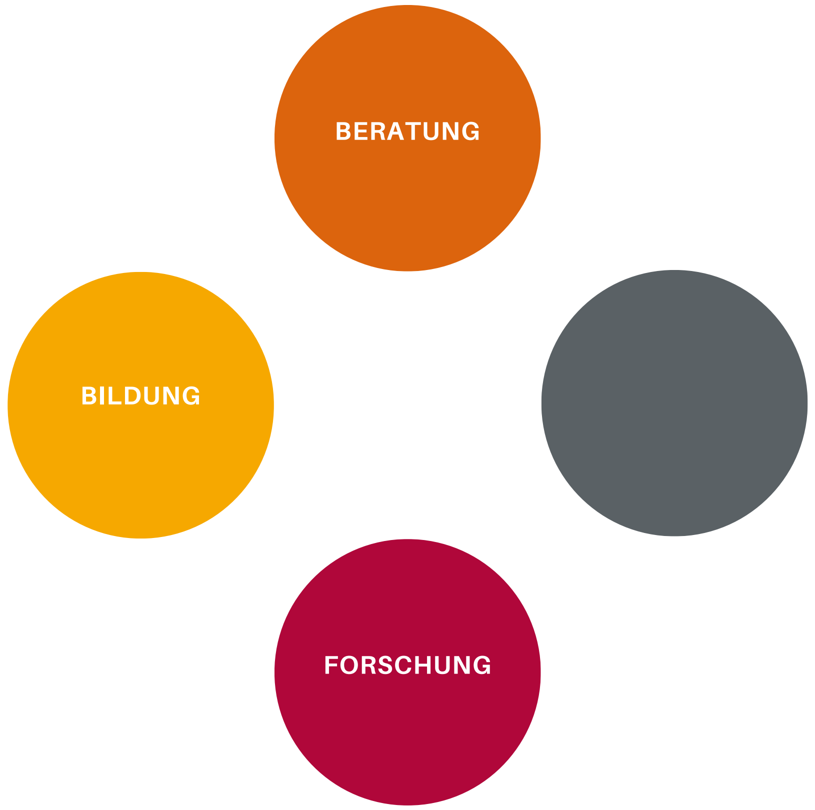
WEST AI KI-Servicezentrum

**Bonn | St. Augustin | Aachen | Jülich | Dortmund | Paderborn**
KI-Hardware, KI-Beratung, KI-Schulungen sowie multimodale und transferierbare KI-Modelle

hessian.AISC

**Darmstadt**
Erklärbarkeit, Generalisierbarkeit und kontextuelle Anpassung

**Ziel:**
**Barrieren der Implementierung von KI-Anwendungen in Gesellschaft und Wirtschaft reduzieren**




**Hasso-Plattner-Institut gGmbH**

- Bildet mit **Universität Potsdam** die **Digital Engineering Fakultät**
- Vereint **Forschung, Lehre** mit den Vorteilen einer **privat finanzierten, gebührenfreien Institution**
- Besteht aus Einrichtungen wie der **E-School**, der **D-School**,
- dem **Mittelstandsdigitalzentrum** und **dem KI-Servicezentrum**

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Newsletter

# BILDUNG

## Talks

tele-task.de/series/1463

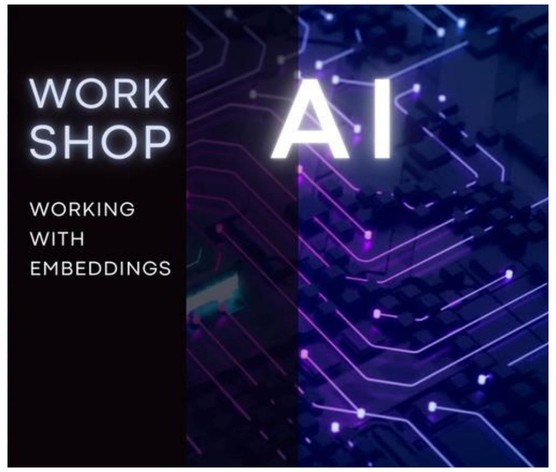
- Gastvorträge zu Forschung und Innovation




## Work shops

aimaker.community

- Praxisnahe Themen
- Beispielthemen: Speech2summary, Docker für ML, semantische Suche

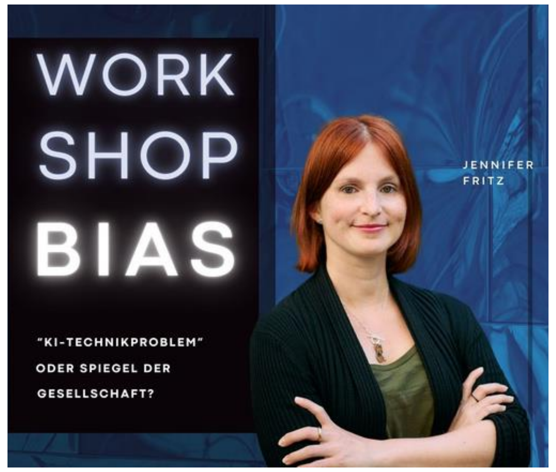


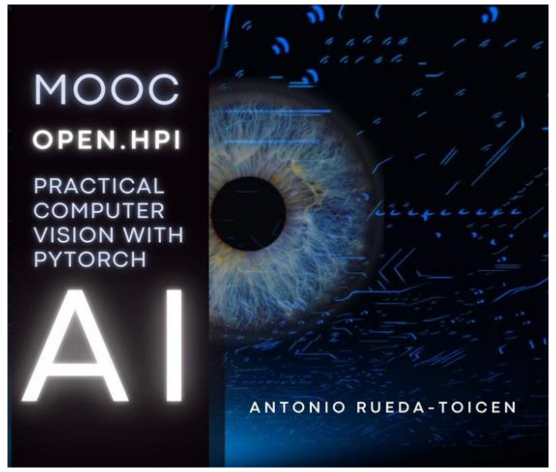



## MOOCs

open.hpi.de/channels/ai-service-center

- ChatGPT: Was bedeutet generative KI für unsere Gesellschaft?
- Profitable KI
- KI Biases verstehen und vermeiden





GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Sprechstunde buchen

# BERATUNG

## KI-Sprechstunde

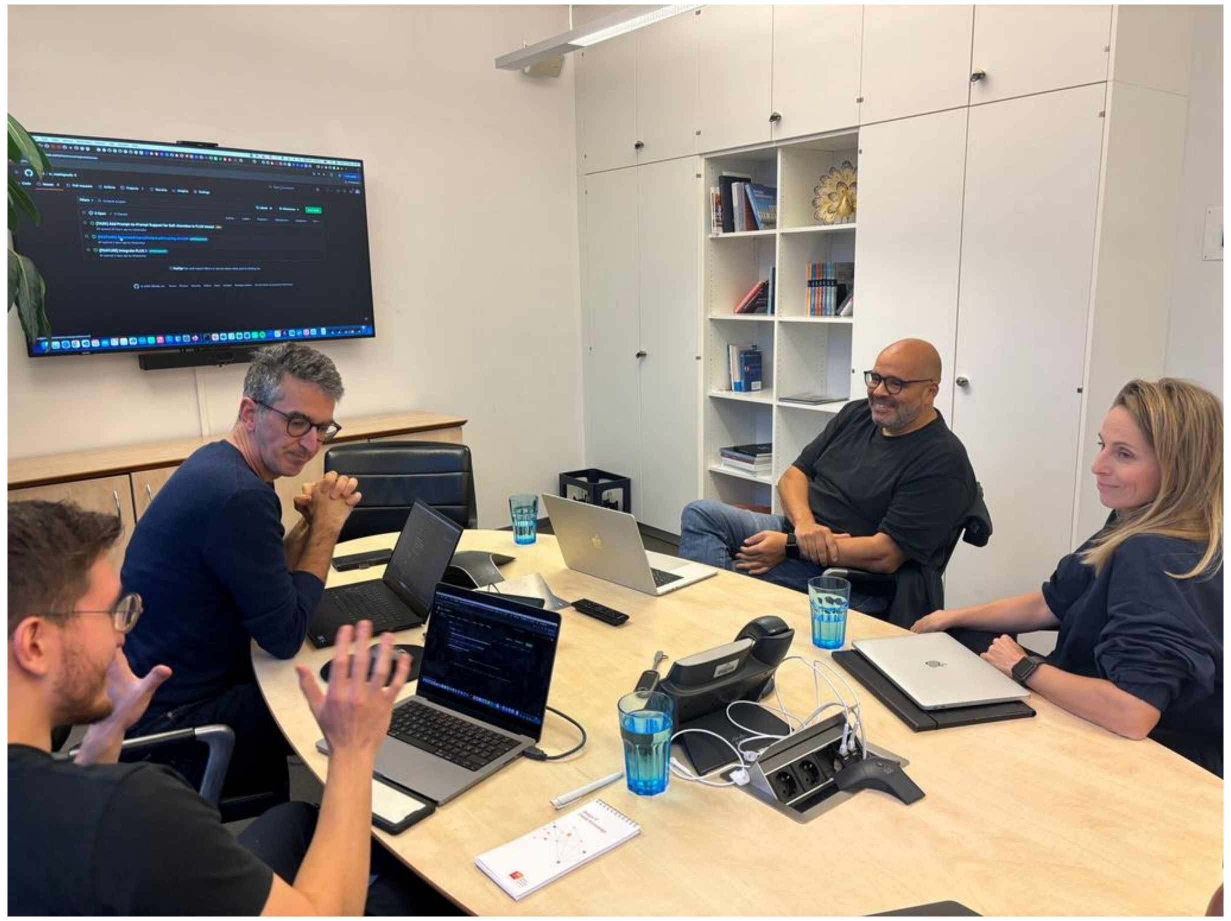
- Beantwortung von Fragen:
  - zu KI-Infrastruktur
  - zu KI-Modellen & Frameworks

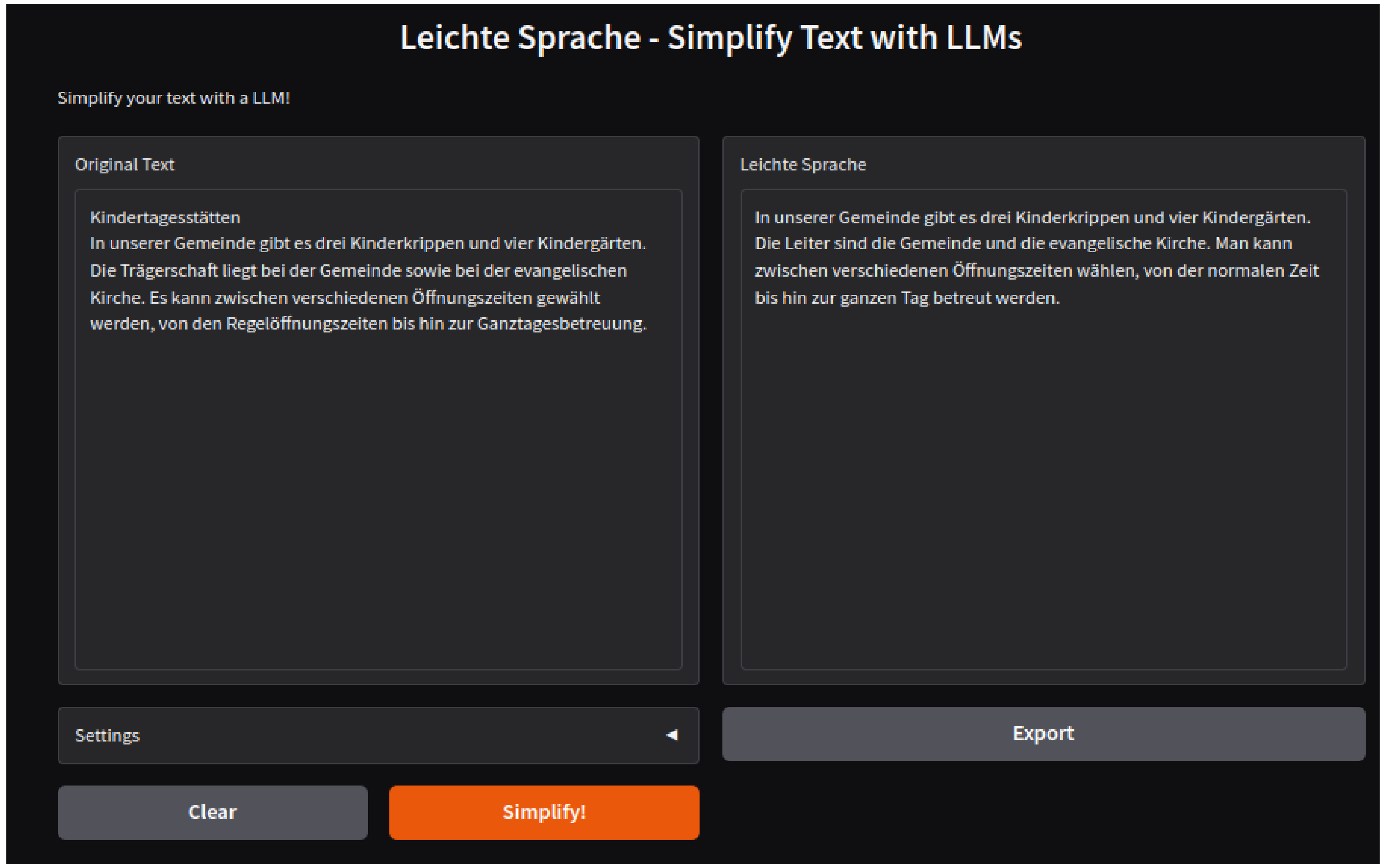
## KI-Pilotprojekte

- Co-Entwicklung eines Prototyps
- Bewerbung alle drei Monate
- Auswahlkriterien z.B. KI-Reife, Gemeinwohl
- Veröffentlichung der Ergebnisse

## Kooperationen

- Gemeinsam organisierte Netzwerktreffen






github.com/aihpi/leichte-sprache

**Bisherige KI-Pilotprojekte**

- Generierung Mathematik-Problemen
- Leichte Sprache
- Generierung von Upcycling Vorschlägen
- Reduzierung von Food Waste
- Datierung mittels Handschrift

Jetzt bewerben!

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Zugangsanfrage
aisc.hpi.de

# INFRASTRUKTUR


- Zugang **kostenfrei**
- kein Produktionsbetrieb
  - Daten sollten **anonymisiert** oder **synthetisiert** sein
  - kein **Hosting** von Produkten
- **Reporting** & **Veröffentlichung** durch Nutzende
- **Altrechte** bleiben bei Nutzenden
- **Neurechte** bleiben bei Nutzenden
  - Einräumen von Nutzungsrechten für Forschung und Lehre


**Training**
- 64 NVIDIA H100 GPU

**Inferenz**
- 40 NVIDIA A30 GPU

**ARM Server**
- Ampere Altra Max M128-30 CPU
- 2 x NVIDIA L40 GPUs

**GPU Server**
- AMD Epyc CPU
- 8 x NVIDIA L40S GPU

**Edge**
- ARMv8 CPU
- NVIDIA Jetson AGX Module

**Neuromorph**
- 288 SpiNNaker2 Chips

**Speicher**
- 1.5 PB NVRAM

**Netwerk**
- 400 Gb/s Infiniband
- 200 Gb/s Ethernet

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# FORSCHUNG

- **KI-Methoden Forschung**
- **KI-Betriebsforschung**

**PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain?**

Sedigheh Eslami, Christoph Meinel, Gerard de Melo
Hasso Plattner Institute / University of Potsdam
{sedigheh.eslami, christoph.meinel, gerard.demelo}@hpi.de

**Exploring Paracrawl for Document-level Neural Machine Translation**

Yusser Al Ghussin[1,2], Jingyi Zhang[3] and Josef van Genabith[1,2]
[1]German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus, Saarbrücken, Germany
[2]Department of Language Science and Technology, Saarland University, Germany
[3]Hasso-Plattner-Institut (HPI), Potsdam, Germany
yusser.al_ghussin/Josef.Van_Genabith@dfki.de,Jingyi.Zhang@hpi.de

**Efficient Parallelization Layouts for Large-Scale Distributed Model Training**

Johannes Hagemann, Aleph Alpha / Hasso Plattner Institute, johannes.hagemann@student.hpi.de
Samuel Weinbach, Aleph Alpha, samuel.weinbach@aleph-alpha.com
Konstantin Dobler, Hasso Plattner Institute, konstantin.dobler@hpi.de
Maximilian Schall, Hasso Plattner Institute, maximilian.schall@hpi.de
Gerard de Melo, Hasso Plattner Institute, gerard.demelo@hpi.de

Abstract

Efficiently training large language models requires parallelizing across hundreds of hardware accelerators and invoking various compute and memory optimizations. When combined, many of these strategies have complex interactions regarding the final training efficiency. Prior work tackling this problem did not have access to the latest set of optimizations, such as FLASHATTENTION or sequence parallelism. In this work, we conduct a comprehensive ablation study of possible training configurations for large language models. We distill this large study into several key recommendations for the most efficient training. For instance, we find that using a micro-batch size of 1 usually enables the most efficient training layouts. Larger micro-batch sizes necessitate activation checkpointing or higher degrees of model parallelism and also lead to larger pipeline bubbles. Our most efficient configurations enable us to achieve state-of-the-art training efficiency results over a range of model sizes, most notably a Model FLOPs utilization of 70.5% when training a LLAMA 13B model.

Workshop on Advancing Neural Network Training at 37th Conference on Neural Information Processing Systems (WANT@NeurIPS 2023).

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

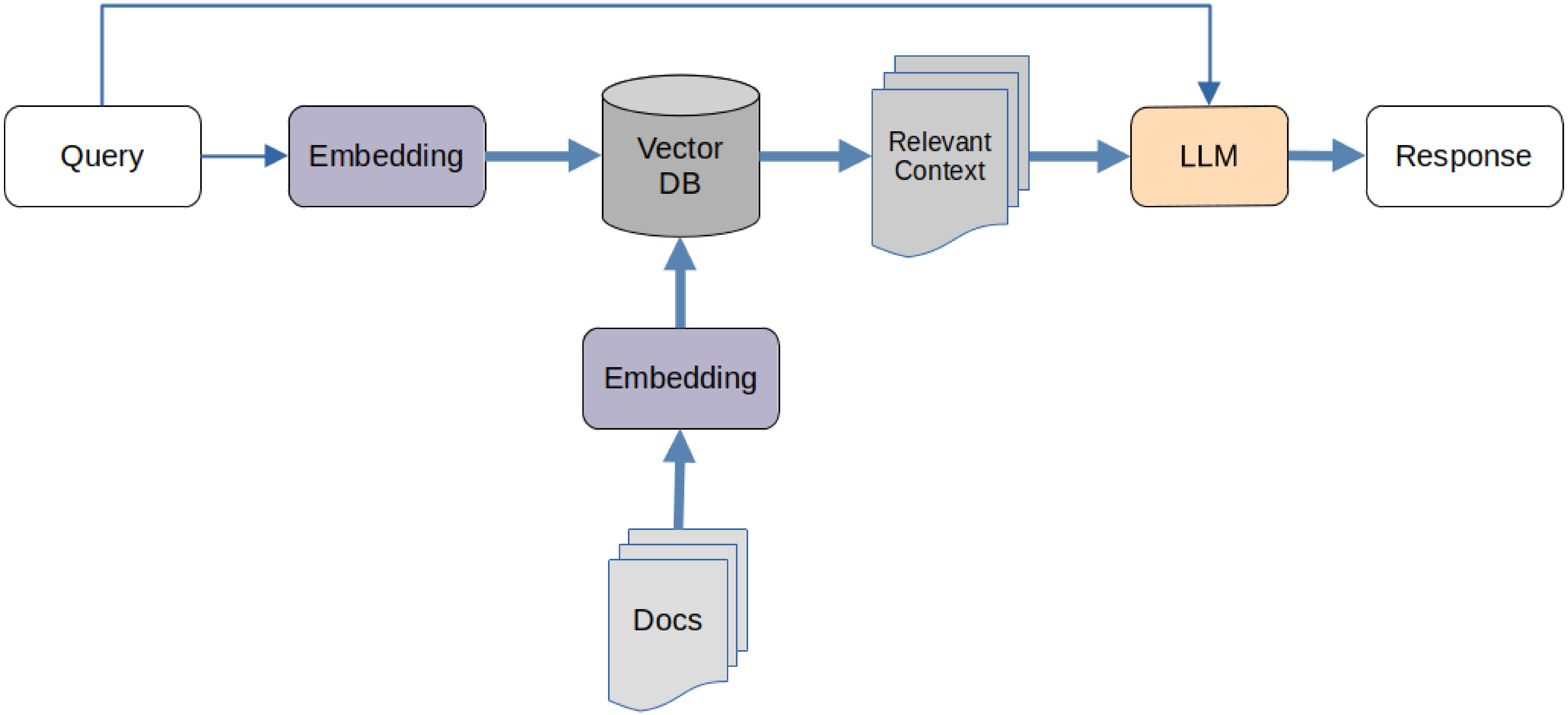
# Motivation

## Retrieval Augmented Generation (RAG)

- LLMs werden mit (größtenteils) öffentlichen Daten trainiert
- LLMs haben kein Wissen über ‚private' Daten
  - Interne Dokumente über Betriebsabläufe
  - Kostenpflichtige Dokumente wie wissenschaftliche Publikationen
  - …
- LLMs vermischen Daten
  - Gesetzestexte zu Parkverboten werden mit Gesetzestexten zu Halteverboten in Verbindung gebracht
  - Wissenscahftliche Texte werden mit Populärwissenscahftlichen Texten in Verbindung gebracht
  - Wenn keine konkrete Informationen zur Beantwortung einer Frage vorliegen, werden ähnliche Informationen benutzt
- RAG erlaubt es, vordefinierte Dokumente zur Fragenbeantwortung zugrunde zu legen!

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# Retrieval Augmented Generation (RAG)

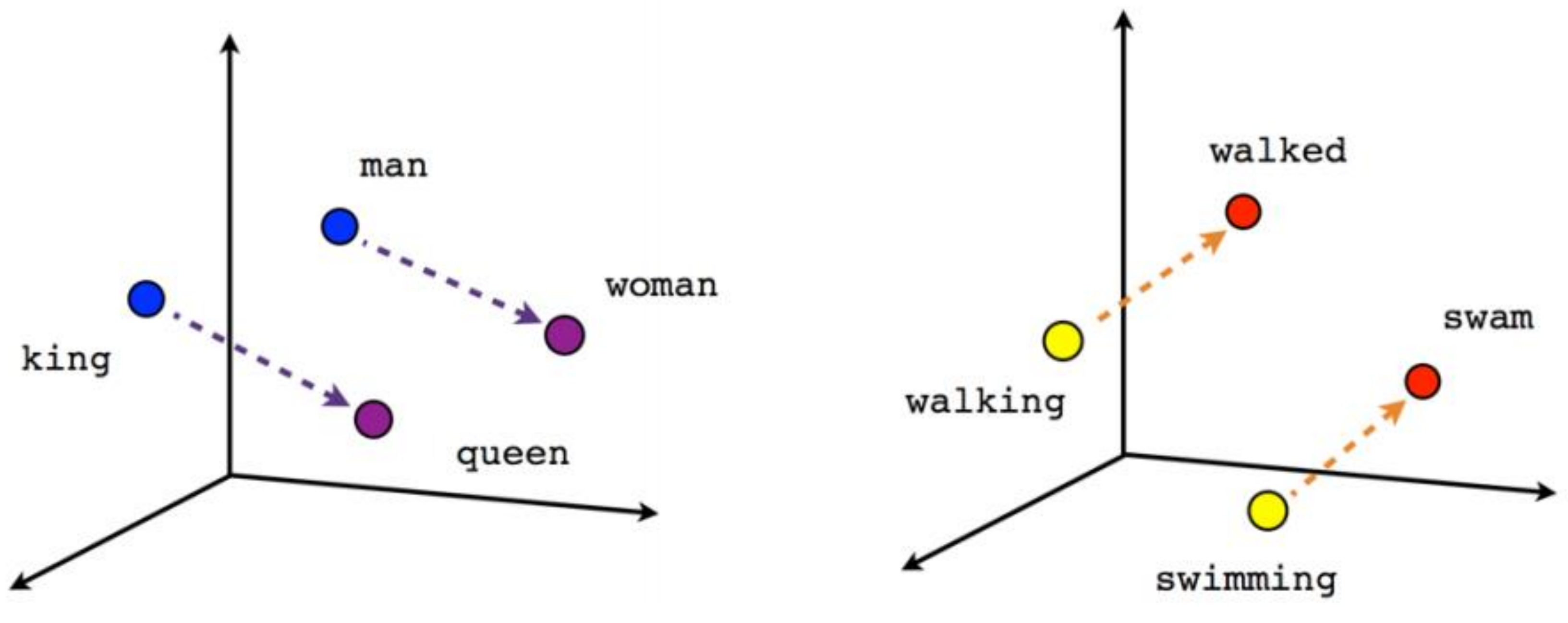



GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

# Von Sätzen zu Embeddings

Embeddings sind multidimensionale Vektoren, die Wörter oder Sätze und deren Bedeutung repräsentieren.




Male-Female

Verb tense

Die Berechnung von Embeddings erfordert die folgenden Schritten:

1. Tokenisation

2. Vectorisation

3. Embedding

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# Tokenisation

z.B. [tiktokenizer]

Morphologisch komplexe Wörter bestehen aus Sub-Wort-Einheiten.

Morphologisch komplexe Wörter bestehen aus Sub-Wort-Einheiten.

44, 16751, 1640, 16438, 84869, 8536, 468, 9603, 466, 1888, 41797, 9608, 3804, 13299, 371, 13737, 258, 90349, 13

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

# Vectorisation

Look-Up Table

Zufällig initialisierte Zeilen-Vektoren

| Token | ID | $d_1$ | $d_2$ | ... | $d_n$ |
|---|---|---|---|---|---|
| M | 44 | 0.433 | 0.012 | ... | 0.124 |
| orph | 16751 | 0.234 | 0.432 | ... | 0.191 |
| olog | 1640 | -0.123 | 0.002 | ... | -0.191 |
| isch | 16438 | -0.543 | -0.042 | ... | -0.200 |
| komple | 84869 | 0.723 | -0.431 | ... | -0.102 |
| xe | 8536 | 0.413 | -0.984 | ... | -0.099 |
| W | 468 | -0.043 | -0.012 | ... | 0.911 |
| ör | 9603 | -0.043 | 0.010 | ... | -0.812 |
| ter | 466 | -0.133 | -0.151 | ... | -0.221 |

Die Größe der Matrix *E* wird festgelegt durch die Anzahl der Tokens im Vokabular *V* und die Anzahl der Dimensionen *D*. Beide Parameter sind Hyperparameter.

$$\mathrm{E} \in \mathbb{R}^{V \times D}$$

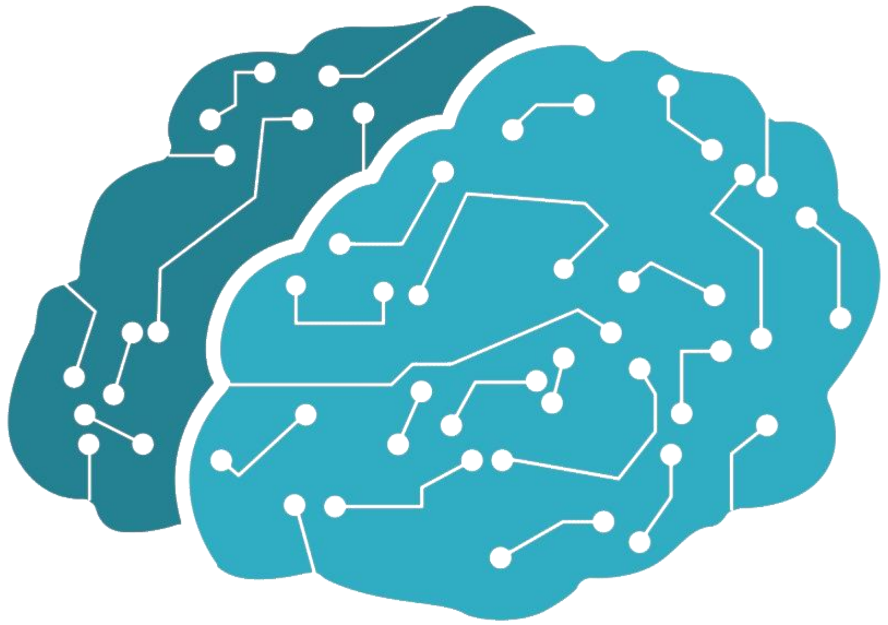
Die Arbeit mit zu großen Matrizen benötigt viele Rechenressourcen. Subwort-Einheiten sind die optimale Balance zwischen Buchstaben- und Wort-basierten Vektoren.

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# Embedding

| | |
|---|---|
| M | 44 |
| orph | 16751 |
| olog | 1640 |
| isch | 16438 |
| komple | 84869 |
| xe | 8536 |
| W | 468 |
| ör | 9603 |
| ter | 466 |
| best | 1888 |
| ehen | 41797 |
| aus | 9608 |
| Sub | 3804 |
| -W | 13299 |
| ort | 371 |
| -E | 13737 |
| in | 258 |
| heiten | 90349 |

## 1. Forward Pass



| Vorhergesagt | Beobachtet |
|---|---|
| in 258 | . 13 |

## 2. Backpropagation

Verlustfunktion (*Loss function*) zur Berechnung des Fehlers

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# Hands-On

https://github.com/aihpi/kisz-local-rag/blob/main/workflow.ipynb

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

kisz@hpi.de

hpi.de/kisz

KI Service Zentrum
HPI
by Hasso-Plattner-Institut

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

# Ihre Meinung ist uns wichtig!

QR-Code zum Feedback-Formular