

Felix Bölter
Hanno Müller

GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

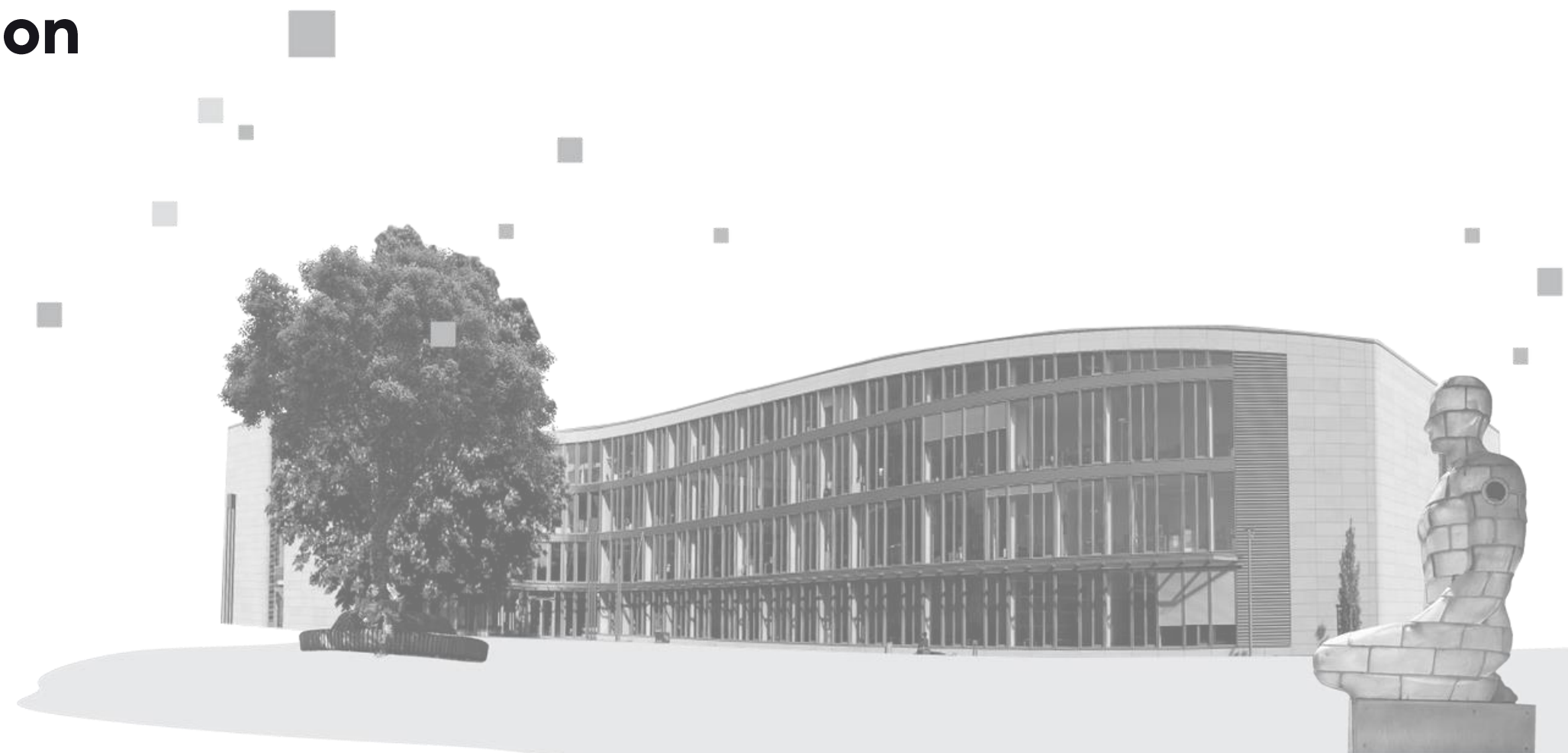
Effizientes Abfragen von Informationen aus Dokumenten.

Lokales

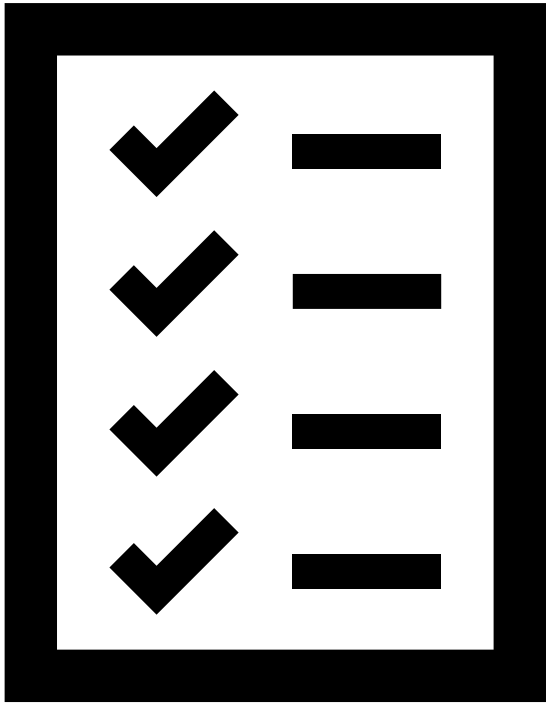
Retrieval Augmented Generation
System

Design IT.
Create Knowledge.

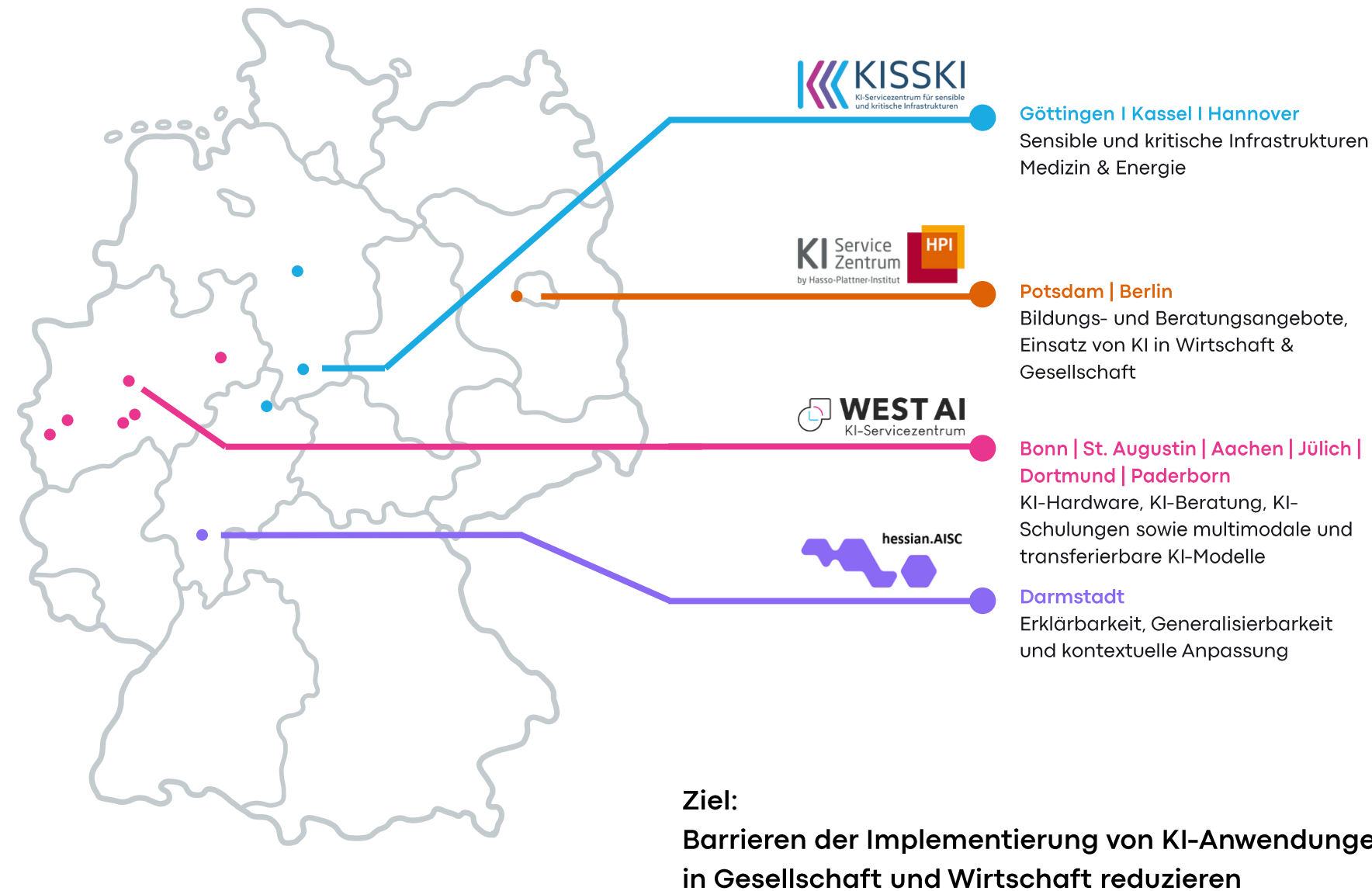
www.hpi.de



Attendance List



KI-Servicezentrum Berlin-Brandenburg



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



GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Talks



tele-task.de/series/1463

- Gastvorträge zu Forschung und Innovation

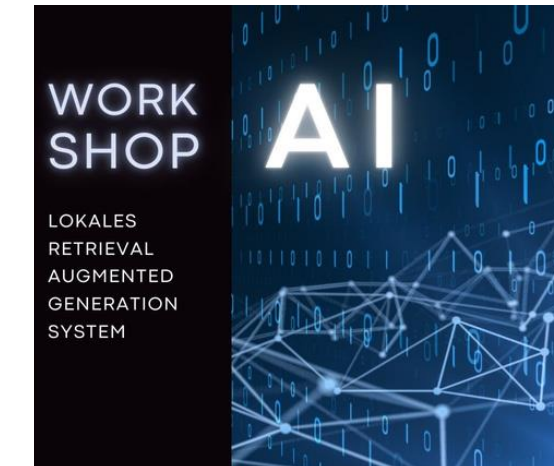


Workshops



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





Sprechstunde buchen

KI-Sprechstunde

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

KI-Pilotprojekte

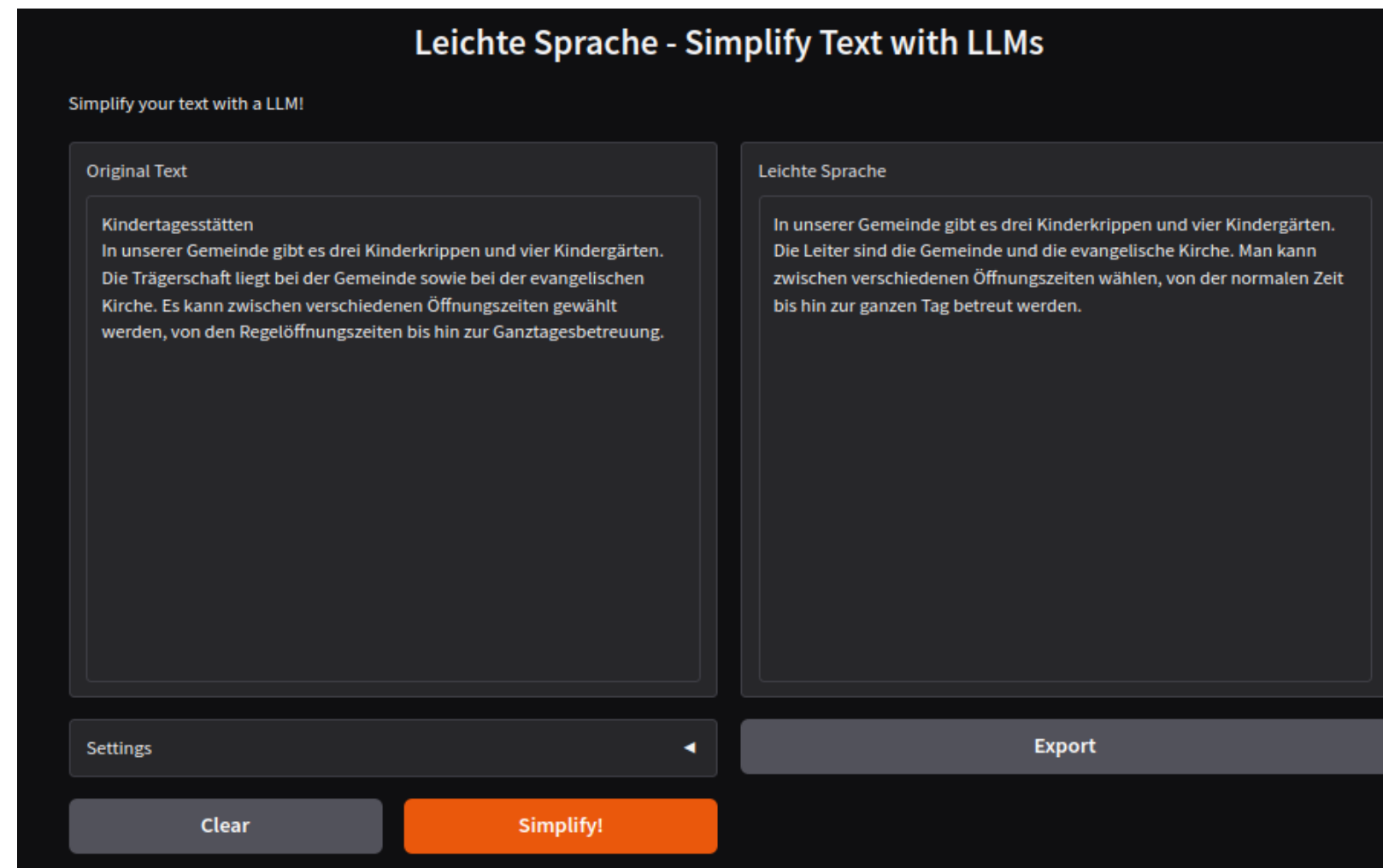
- Co-Entwicklung eines Prototyps
- Bewerbung alle drei Monate
- Auswahlkriterien z.B. KI-Reife, Gemeinwohl
- Veröffentlichung der Ergebnisse

Kooperationen

- Gemeinsam organisierte Netzwerktreffen



BERATUNG



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



- 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

Edge

- ARMv8 CPU
- NVIDIA Jetson AGX Module

Inferenz

- 40 NVIDIA A30 GPU

Neuromorph

- 288 SpiNNaker2 Chips

ARM Server

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

Speicher

- 1.5 PB NVRAM

GPU Server

- AMD Epyc CPU
- 8 x NVIDIA L40S GPU

Netzwerk

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

GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

FORSCHUNG

- **KI-Betriebsforschung**

GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

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

Sedighch Eslami, Christoph Meinel, Gerard de Melo

Hasso Plattner Institute / University of Potsdam

{sedigheh.eslami, christoph.meinel, gerard.demelo}@hpi.de

Abstract

Contrastive Language-Image Pre-training (CLIP) has shown remarkable success in learning with cross-modal supervision from extensive amounts of image-text pairs collected online. Thus far, the effectiveness of CLIP has been investigated primarily in general-domain multimodal problems. In this work, we evaluate the effectiveness of CLIP for the task of Medical Visual Question Answering (MedVQA). We present PubMedCLIP, a fine-tuned version of CLIP for the medical domain based on PubMed articles. Our experiments conducted on two MedVQA benchmark datasets illustrate that PubMedCLIP achieves superior results improving the overall accuracy up to 3% in comparison to the state-of-the-art Model-Agnostic Meta-Learning (MAML) networks pre-trained only on visual data. The PubMedCLIP model with different back-ends, the source code for pre-training them and reproducing our MedVQA pipeline is publicly available at <https://github.com/sarahelsil/PubMedCLIP>.

1 Introduction

Medical visual question answering (MedVQA) seeks answers to natural language questions about a given medical image. The development of MedVQA has considerable potential to benefit healthcare systems, as it may aid clinicians in interpreting medical images and obtaining more accurate diagnoses by consulting a second opinion. Thus, it has become a very active area of research, with competitive benchmarks and yearly competitions (Abacha et al., 2021). Yet, visual question answering in the medical domain in particular remains non-trivial as we suffer from a general lack of large balanced training data, in part due to privacy concerns. To solve the multimodal task of MedVQA, a system must understand both medical images and textual questions and infer the associations between them sufficiently well to produce a correct answer (An

Findings of the Association for Computational Linguistics
May 2-6, 2023 ©2023 Association for Computational Linguistics

Exploring Paracrawl for Document-level Neural Machine Translation

Yusser Al Ghussin^{1,2}, Jingyi Zhang³ and Josef van Genabith^{1,2}

¹German Research Center for Artificial Intelligence (DFKI),
Saarland Informatics Campus, Saarbrücken, Germany

²Department of Language Science and Technology, Saarland University, Germany

³Hasso-Plattner-Institut (HPI), Potsdam, Germany

yusser.al_ghussin/Josef.Van_Genabith@dfki.de, Jingyi.Zhang@hpi.de

Abstract

Document-level neural machine translation (NMT) has outperformed sentence-level NMT on a number of datasets. However, document-level NMT is still not widely adopted in real-world translation systems mainly due to the lack of large-scale general-domain training data for document-level NMT. We examine the effectiveness of using Paracrawl for learning document-level translation. Paracrawl is a large-scale parallel corpus crawled from the Internet and contains data from various domains. The official Paracrawl corpus was released as parallel sentences (extracted from parallel webpages) and therefore previous works only used Paracrawl for learning sentence-level translation. In this work, we extract parallel paragraphs from Paracrawl parallel webpages using automatic sentence alignments and we use the extracted parallel paragraphs as training elements for training document-level translation models. We show that document-level NMT models trained with only parallel paragraphs from Paracrawl can be used to translate real outputs from TED, News and Europarl, outperforming sentence-level NMT models. We also perform a targeted pronoun evaluation and show that document-level models trained with Paracrawl data can help context-aware pronoun translation. We release our data and code here¹.

1 Introduction

The Transformer translation model (Vaswani et al., 2017), which performs sentence-level translation based on attention networks, has achieved great success and significantly improved the state-of-the-art in machine translation. Compared to sentence-level translation, document-level translation (Xu et al., 2021; Bao et al., 2021; Jauregi Unanue et al., 2020; Ma et al., 2020; Maruf et al., 2019; Tu et al., 2018; Maruf and Haffari, 2018) performs translation at document-level and can potentially fur-

¹<https://github.com/Yusser96/Exploring-Paracrawl-for-Documents-level-Neural-Machine-Translation>

13
 Proceedings of the 17th Conference of the European Chapter of
 May 2-6, 2023 ©2023 Association

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
e 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 layers. 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.

1 Introduction

The number of parameters and computational resources spent on training deep neural networks is growing rapidly [13, 14]. The largest models consisting of hundreds of billions of parameters do not even fit onto a single hardware accelerator. Thus, training these models requires various ways of reducing the memory requirements, such as ZeRO [16], activation checkpointing [2], and 3D-parallel (data, tensor, and pipeline parallel) training [13]. 3D parallelism, in particular, has been demonstrated to be effective for the training of Transformer-based large language models (LLMs) with hundreds of billions of parameters [13].

However, training these models efficiently with 3D parallelism requires significant domain expertise and extensive manual effort to determine the ideal configurations. These configurations not only need to combine data, model, and pipeline parallelism most efficiently, but also consider complex interactions with other memory and compute optimizations. FLASHATTENTION [5] in particular has had a notable impact since its release, enabling us to train models at previously impossible degrees of training efficiency. In light of these developments, we conduct a systematic study via a large-scale training efficiency sweep of these interactions. We consider up to 256 GPUs and LLAMA models with up to 65 billion parameters.

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

Motivation

Retrieval Augmented Generation (RAG)

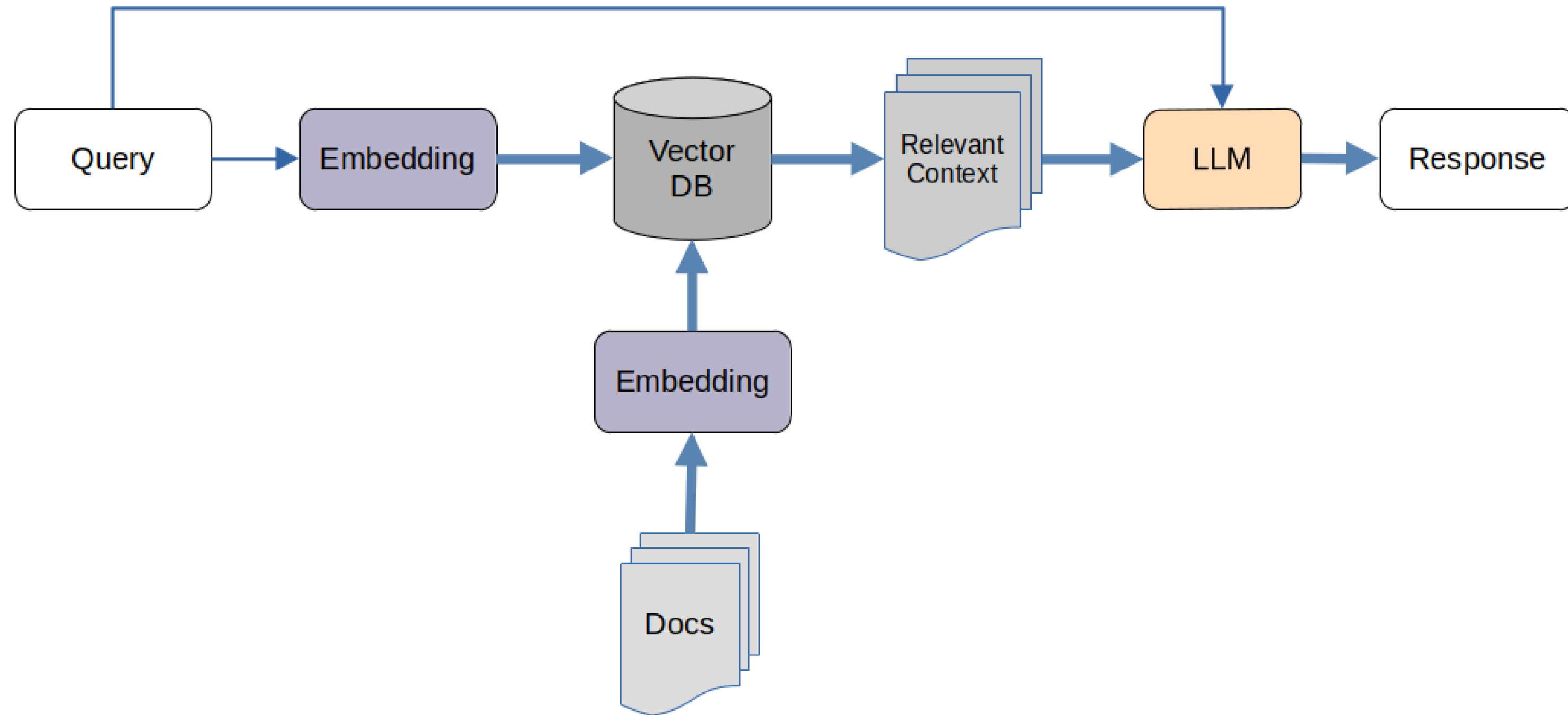
- LLMs werden mit (größtenteils) öffentlichen Daten trainiert
- LLMs haben kein Wissen über ‚private‘ Daten
 - Interne Dokumente über Betriebsabläufe
 - Kostenpflichtige Dokumente wie wissenschaftliche Publikationen
 - ...
- LLMs vermischen Daten
 - Gesetzestexte zu Parkverboten werden mit Gesetzestexten zu Halteverboten in Verbindung gebracht
 - Wissenschaftliche Texte werden mit Populärwissenschaftlichen Texten in Verbindung gebracht
 - Wenn keine konkreten 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)

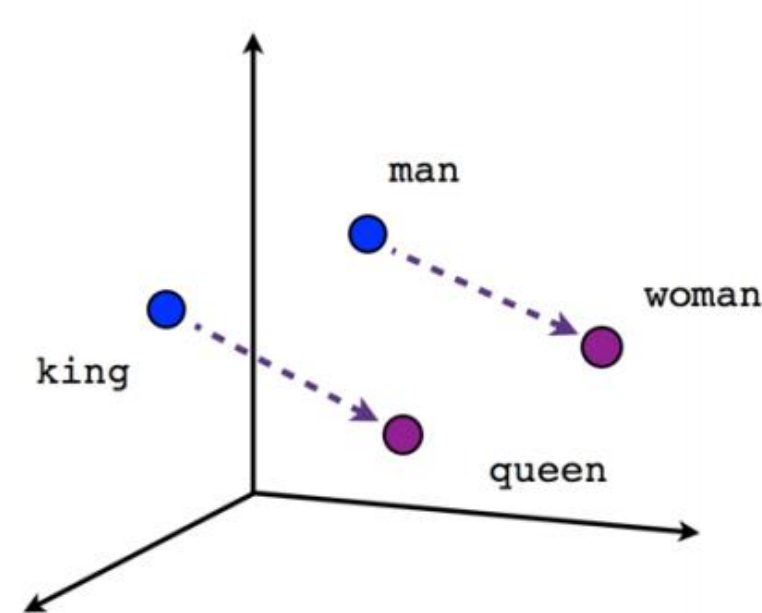


GEFÖRDERT VOM

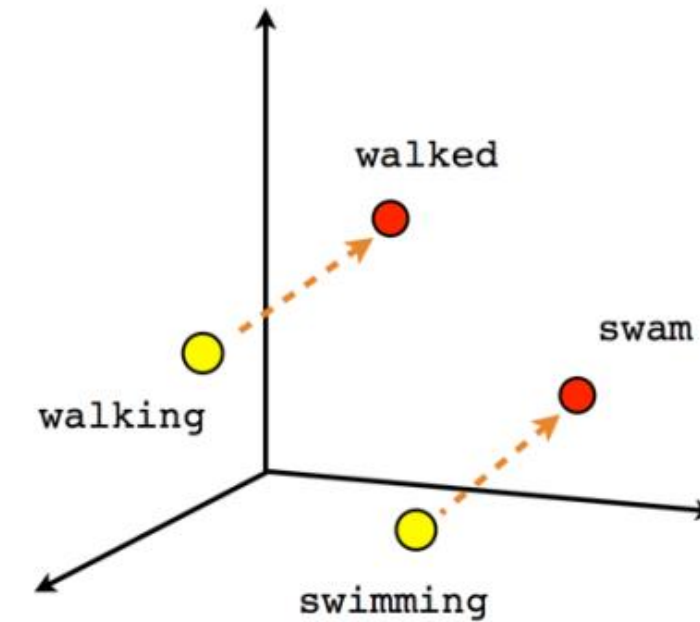


Bundesministerium
für Bildung
und Forschung

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

Von Sätzen zu Embeddings

GEFÖRDERT VOM

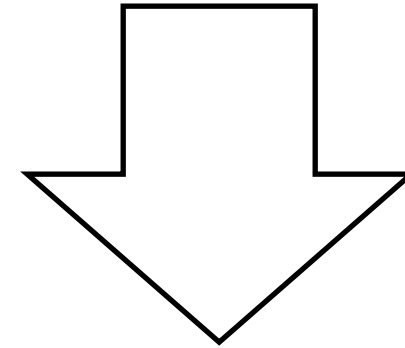


Bundesministerium
für Bildung
und Forschung

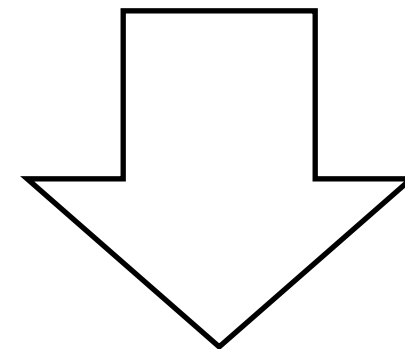
Tokenisation

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

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

Zufällig initialisierte
Zeilen-Vektoren

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.

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

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

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

Ihre Meinung ist
uns wichtig!



QR-Code zum Feedback-Formular

