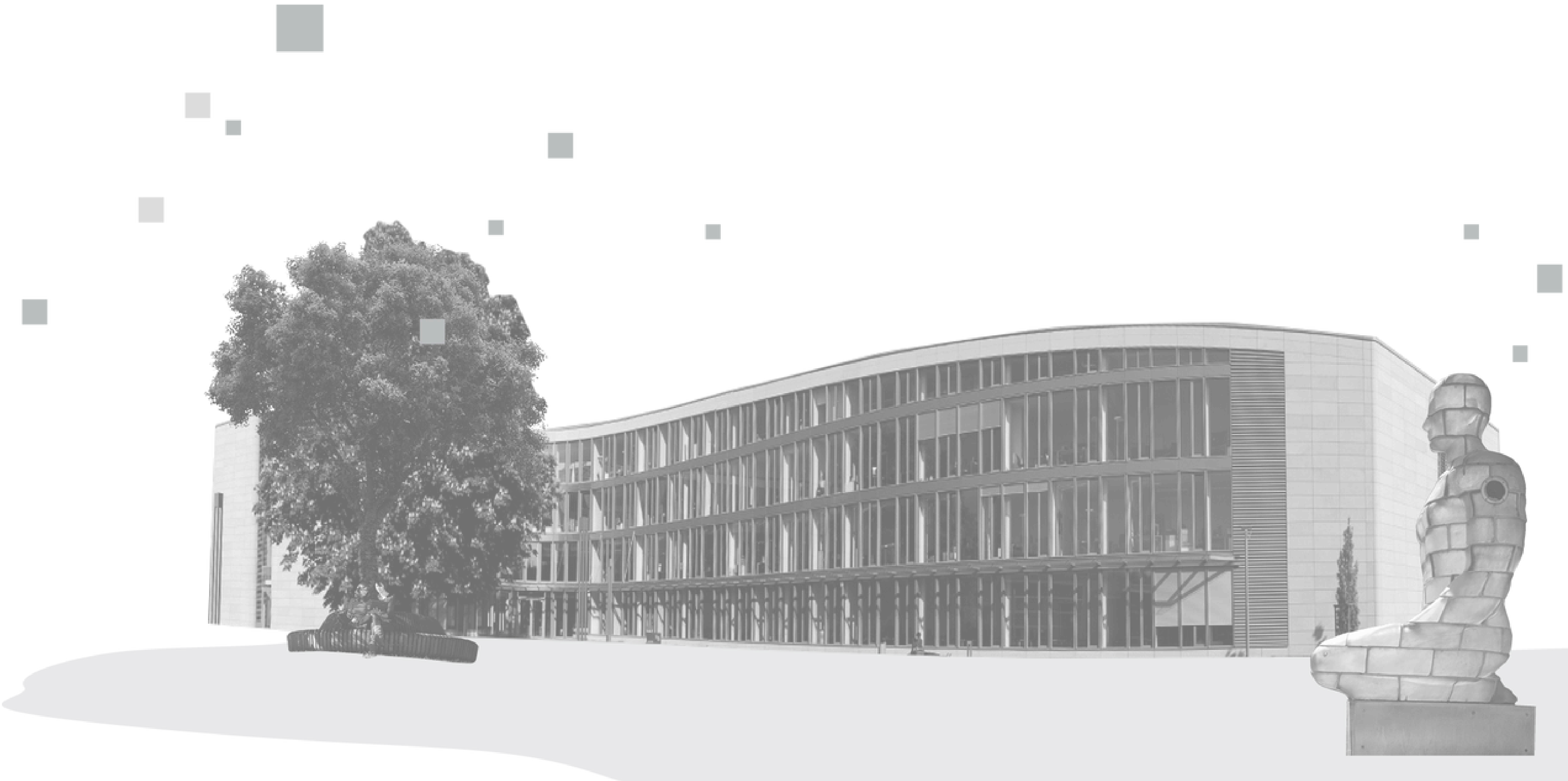
KI Service Zentrum by Hasso-Plattner-Institut HPI

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

KI-Servicezentrum Berlin-Brandenburg

Prof. Dr. Holger Karl
Lasse Kohlmeyer



Design IT.
Create Knowledge.

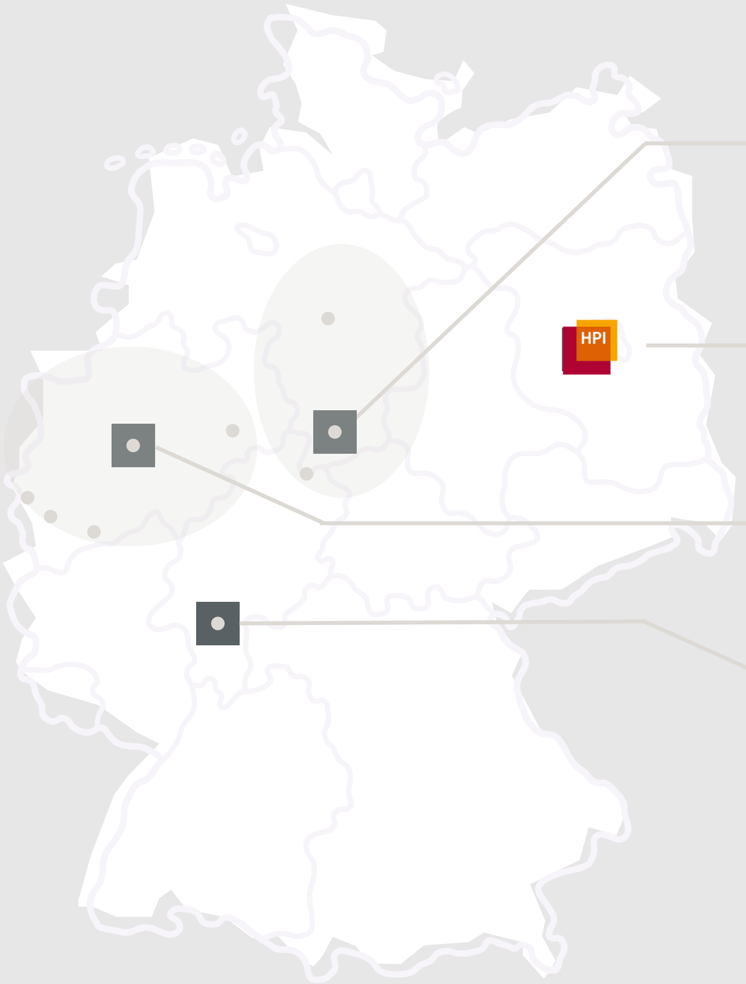
www.hpi.de

KI-Servicezentrum Berlin-Brandenburg

KI Service Zentrum HPI
by Hasso-Plattner-Institut

kisz@hpi.de

hpi.de/kisz



KISSKI

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

KI-Services.HPI

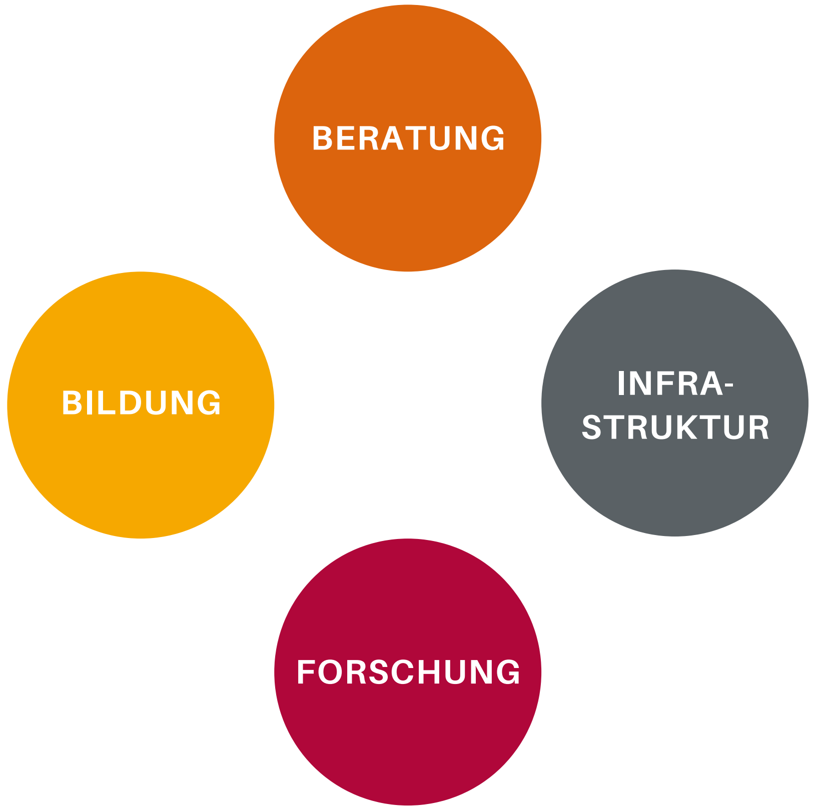
Berlin I Brandenburg
Bildungs- und Beratungsangebote,
Einsatz von KI in Wirtschaft & Gesellschaft

WestAI

Bonn | Fraunhofer | Aachen | Jülich | Dortmund | Paderborn
Multimodale sowie große und transferierbare KI-Modelle

hessian.AISC

Darmstadt
Erklärbarkeit, Generalisierbarkeit
und kontextuelle Anpassung



BERATUNG

BILDUNG

INFRA-STRUKTUR

FORSCHUNG

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

WEITERBILDUNG

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Talks

Forschung, innovationsorientierte Gastvorträge

Work shops

- Verschiedene praxisnahe Themen
- Begleitet von Einführungsvideos
- Beispielthemen: Speech2summary, Docker für ML, semantische Suche

MOOCs

- zukünftig: Ausbau des MOOC-Angebots
- Beispielthema: „ChatGPT: Was bedeutet generative KI für unsere Gesellschaft?"

KI Service Zentrum by Hasso-Plattner-Institut HPI

kisz@hpi.de

hpi.de/kisz

tele-TASK HPI by Hasso-Plattner-Institut

OPEN HPI

EDUCATION & TRAINING

Talks

Democratizing and Productionizing LLMs

Martin Grund , Matei Zaharia

Date: May 17, 2023
Language: English
Duration: 01:19:45

Federated Learning

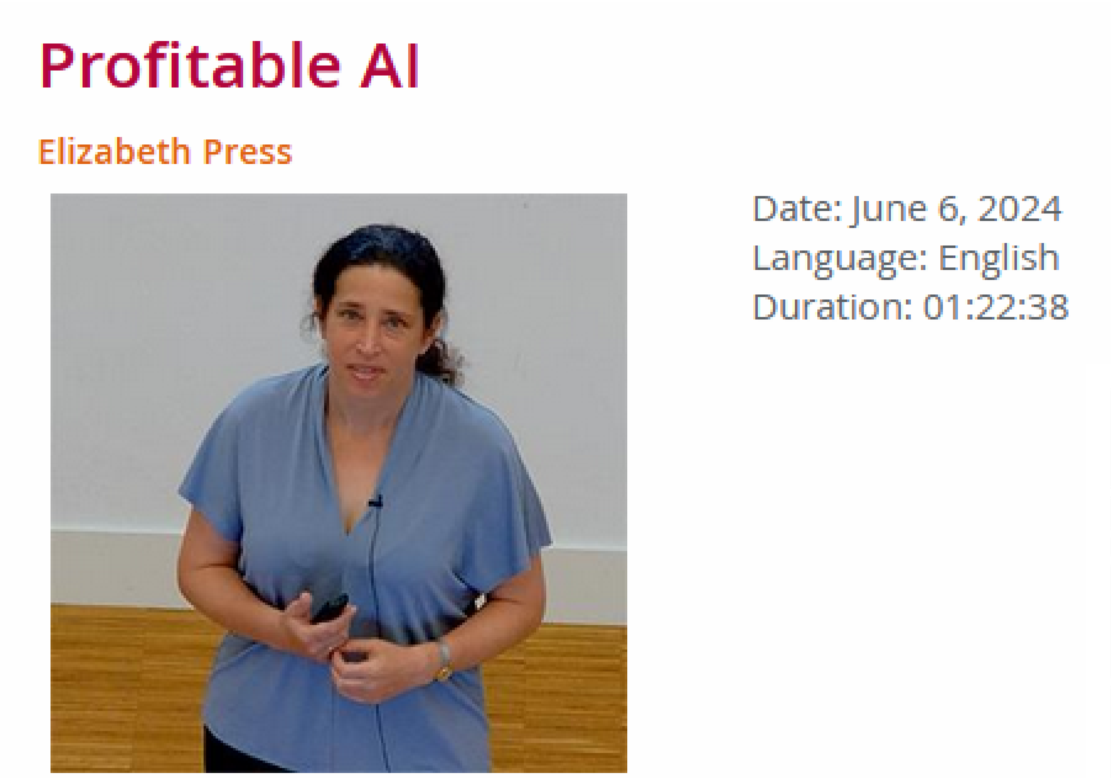
Ramin Khalili

Date: July 27, 2023
Language: English
Duration: 01:25:11

Generative AI in Banking

Max Sommerfeld

Date: April 22, 2024
Language: English
Duration: 01:44:49



Profitable AI

Elizabeth Press

Date: June 6, 2024
Language: English
Duration: 01:22:38

KI Service Zentrum
by Hasso-Plattner-Institut
HPI

SPONSORED BY THE
Federal Ministry
of Education
and Research

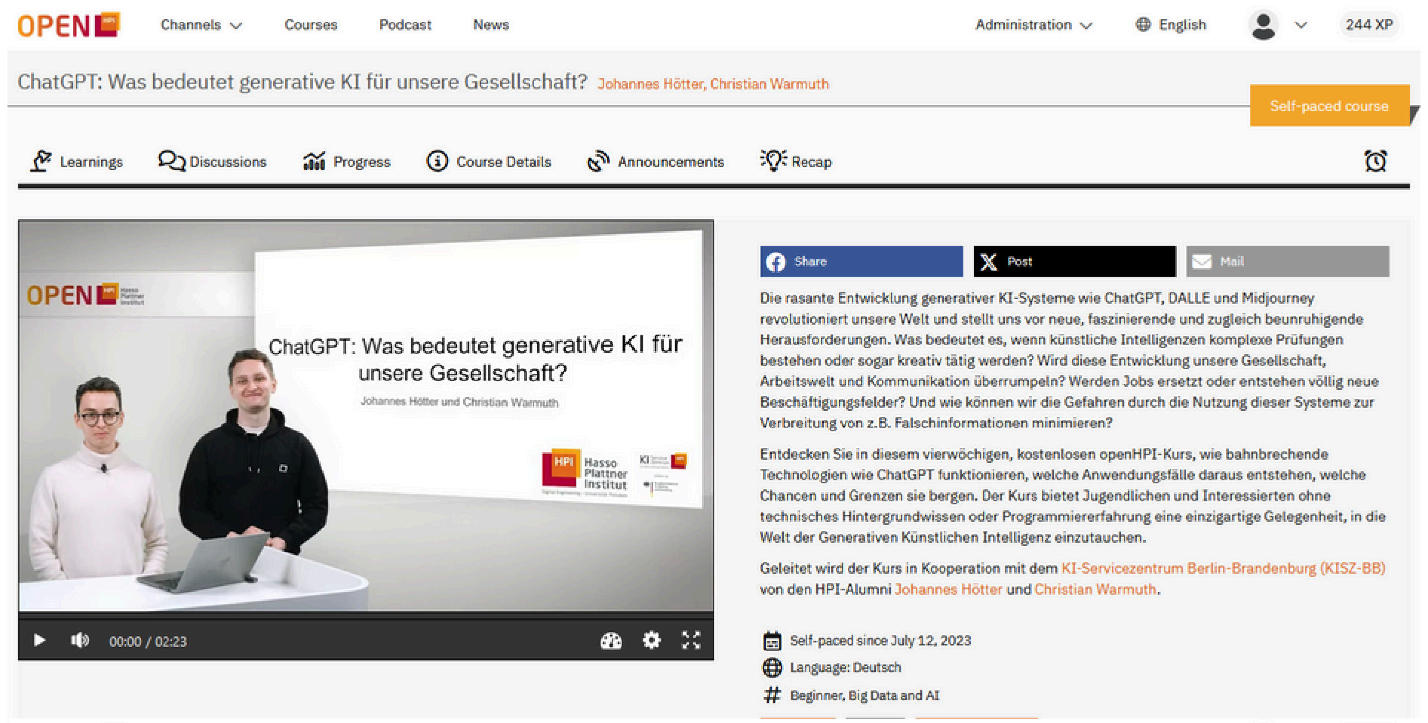
kisz@hpi.de
hpi.de/kisz

tele-task.de/series/1463

Design IT. Create Knowledge.

EDUCATION & TRAINING

MOOCs



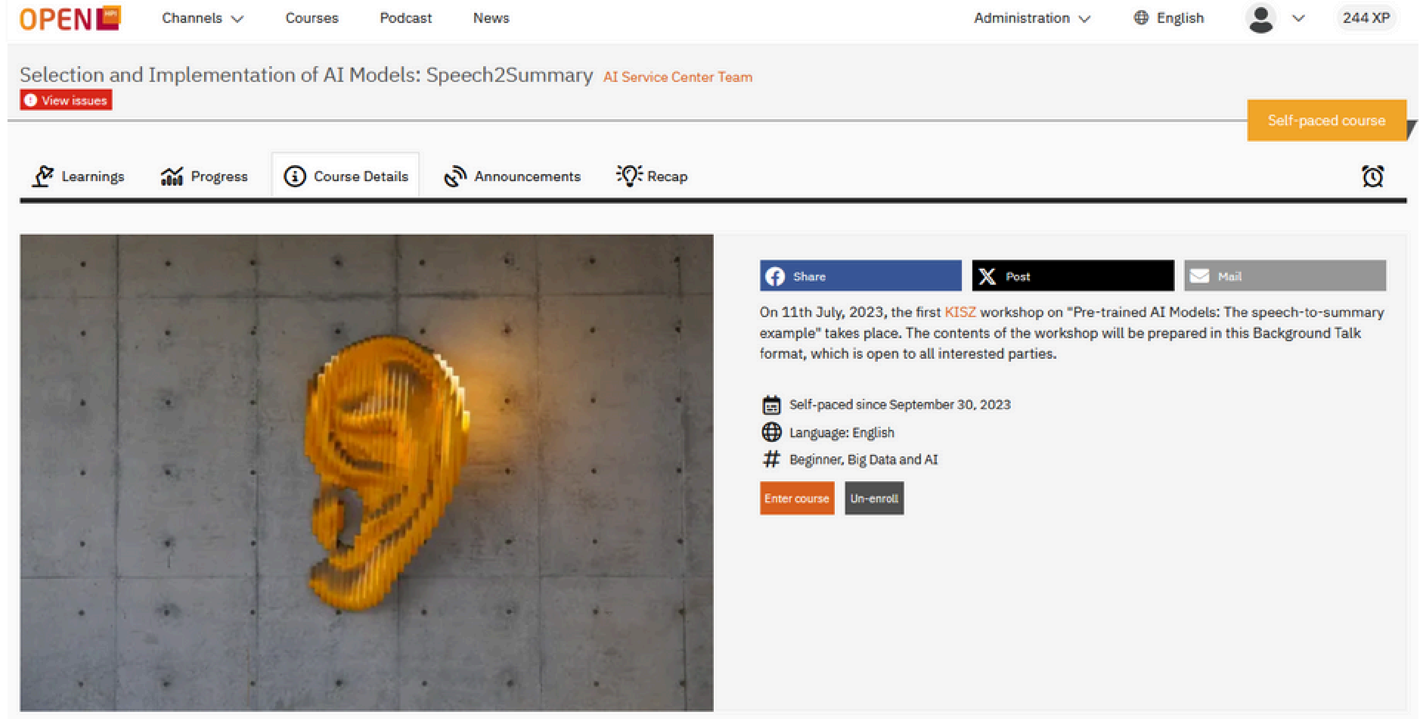
ChatGPT: Was bedeutet generative KI für unsere Gesellschaft? Johannes Hötter, Christian Warmuth

Die rasante Entwicklung generativer KI-Systeme wie ChatGPT, DALLE und Midjourney revolutioniert unsere Welt und stellt uns vor neue, faszinierende und zugleich beunruhigende Herausforderungen. Was bedeutet es, wenn künstliche Intelligenzen komplexe Prüfungen bestehen oder sogar kreativ tätig werden? Wird diese Entwicklung unsere Gesellschaft, Arbeitswelt und Kommunikation überrumpeln? Werden Jobs ersetzt oder entstehen völlig neue Beschäftigungsfelder? Und wie können wir die Gefahren durch die Nutzung dieser Systeme zur Verbreitung von z.B. Falschinformationen minimieren?

Entdecken Sie in diesem vierwöchigen, kostenlosen openHPI-Kurs, wie bahnbrechende Technologien wie ChatGPT funktionieren, welche Anwendungsfälle daraus entstehen, welche Chancen und Grenzen sie bergen. Der Kurs bietet Jugendlichen und Interessierten ohne technisches Hintergrundwissen oder Programmiererfahrung eine einzigartige Gelegenheit, in die Welt der Generativen Künstlichen Intelligenz einzutauchen.

Geleitet wird der Kurs in Kooperation mit dem KI-Servicezentrum Berlin-Brandenburg (KISZ-BB) von den HPI-Alumni Johannes Hötter und Christian Warmuth.

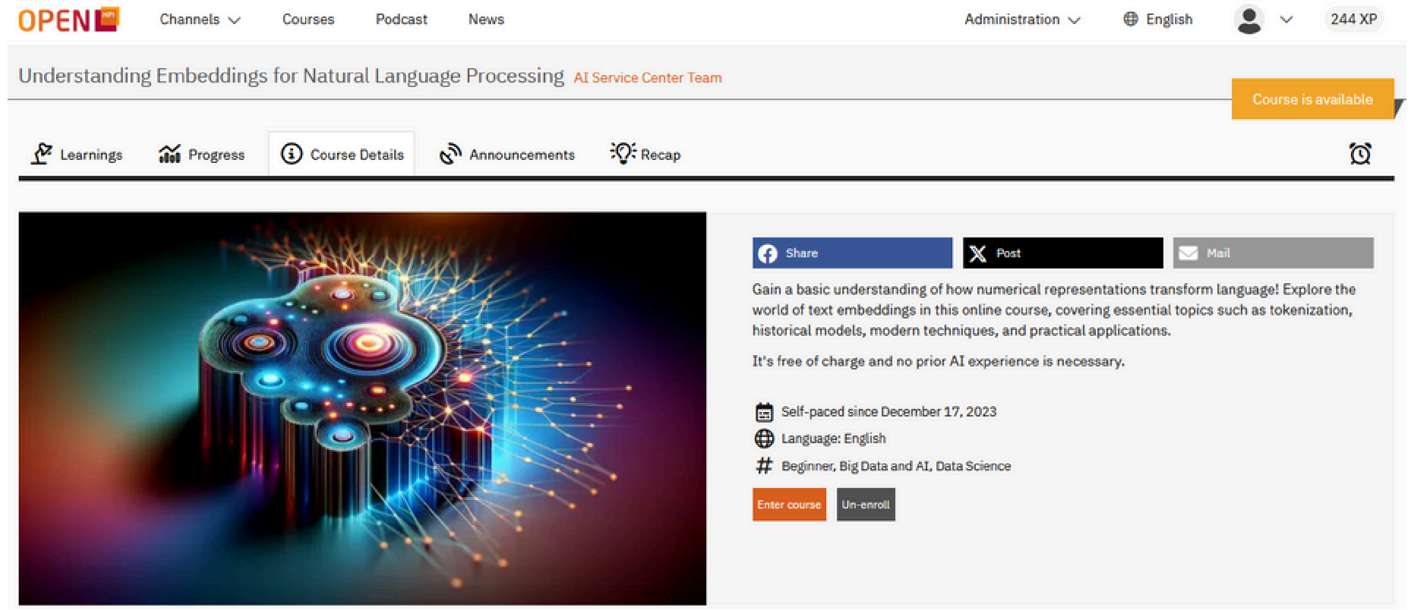
Self-paced since July 12, 2023
Language: Deutsch
Beginner, Big Data and AI



Selection and Implementation of AI Models: Speech2Summary AI Service Center Team

On 11th July, 2023, the first KISZ workshop on "Pre-trained AI Models: The speech-to-summary example" takes place. The contents of the workshop will be prepared in this Background Talk format, which is open to all interested parties.

Self-paced since September 30, 2023
Language: English
Beginner, Big Data and AI



Understanding Embeddings for Natural Language Processing AI Service Center Team

Gain a basic understanding of how numerical representations transform language! Explore the world of text embeddings in this online course, covering essential topics such as tokenization, historical models, modern techniques, and practical applications.

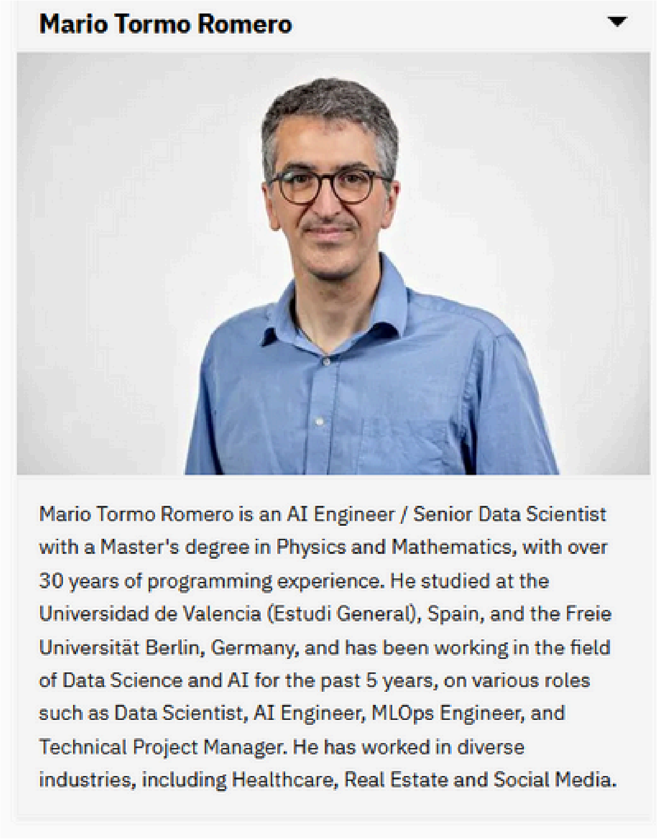
It's free of charge and no prior AI experience is necessary.

Self-paced since December 17, 2023
Language: English
Beginner, Big Data and AI, Data Science

KI Service Zentrum by Hasso-Plattner-Institut HPI

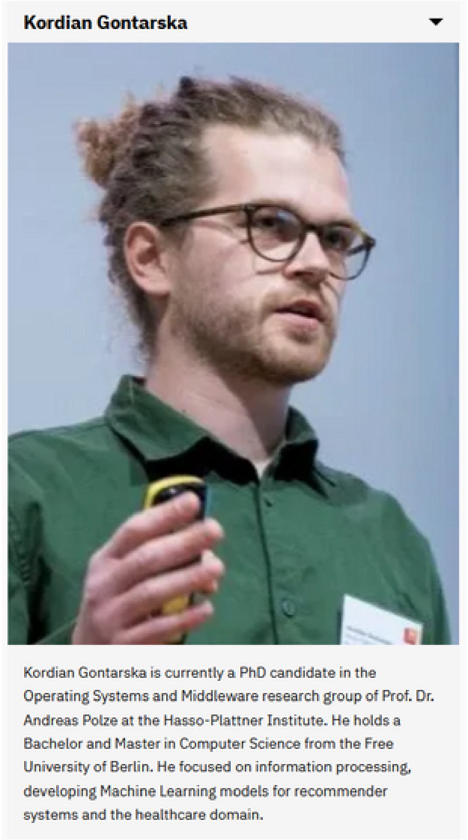
SPONSORED BY THE

Federal Ministry of Education and Research



Mario Tormo Romero

Mario Tormo Romero is an AI Engineer / Senior Data Scientist with a Master's degree in Physics and Mathematics, with over 30 years of programming experience. He studied at the Universidad de Valencia (Estudi General), Spain, and the Freie Universität Berlin, Germany, and has been working in the field of Data Science and AI for the past 5 years, on various roles such as Data Scientist, AI Engineer, MLOps Engineer, and Technical Project Manager. He has worked in diverse industries, including Healthcare, Real Estate and Social Media.



Kordian Gontarska

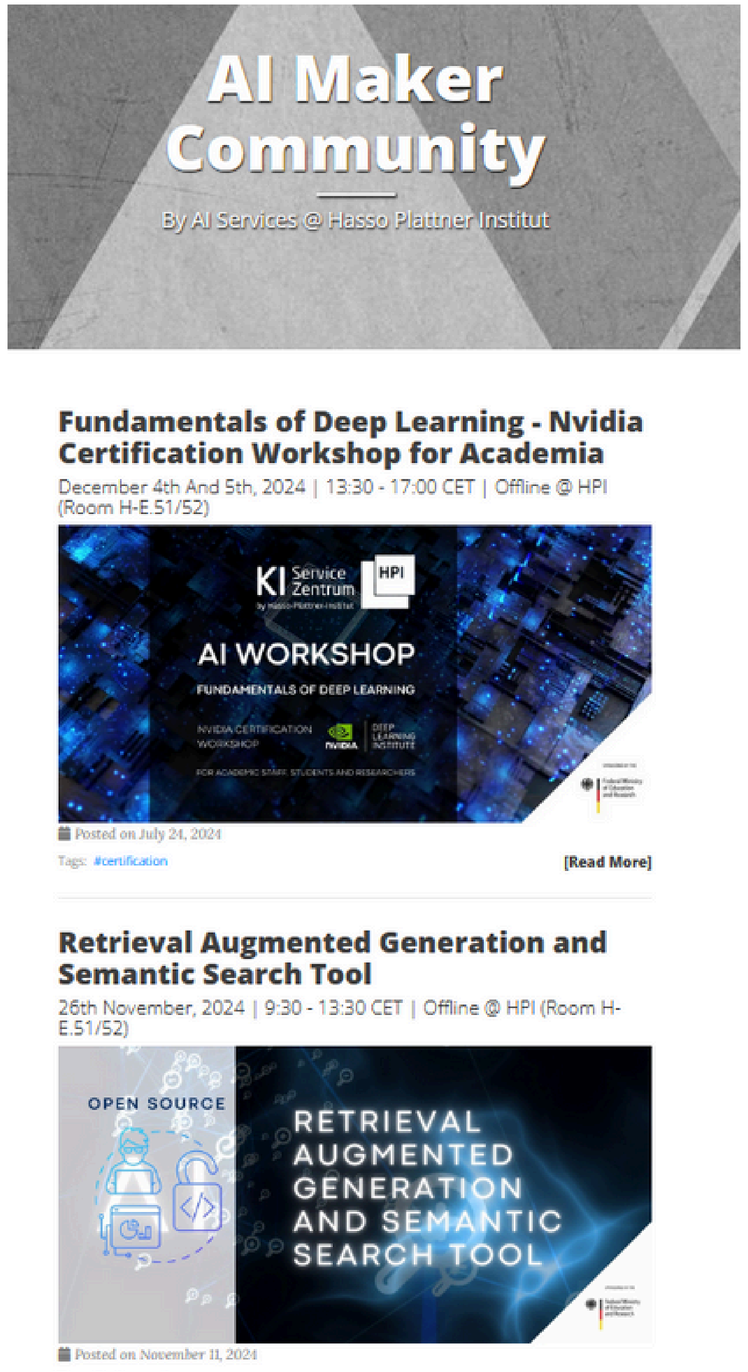
Kordian Gontarska is currently a PhD candidate in the Operating Systems and Middleware research group of Prof. Dr. Andreas Polze at the Hasso-Plattner Institute. He holds a Bachelor and Master in Computer Science from the Free University of Berlin. He focused on information processing, developing Machine Learning models for recommender systems and the healthcare domain.

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

Design IT. Create Knowledge.

EDUCATION & TRAINING

Work shops


AI Maker Community
By AI Services @ Hasso Plattner Institut

Fundamentals of Deep Learning - Nvidia Certification Workshop for Academia

December 4th And 5th, 2024 | 13:30 - 17:00 CET | Offline @ HPI (Room H-E.51/52)

KI Service Zentrum
HPI
AI WORKSHOP
FUNDAMENTALS OF DEEP LEARNING
NVIDIA CERTIFICATION WORKSHOP
NVIDIA
DEEP LEARNING INSTITUTE
FOR ACADEMIC STAFF, STUDENTS AND RESEARCHERS

Posted on July 24, 2024

Tags: #certification

[Read More]

Retrieval Augmented Generation and Semantic Search Tool

26th November, 2024 | 9:30 - 13:30 CET | Offline @ HPI (Room H-E.51/52)

OPEN SOURCE
RETRIEVAL AUGMENTED GENERATION AND SEMANTIC SEARCH TOOL

Posted on November 11, 2024

KI Service Zentrum by Hasso-Plattner-Institut HPI

SPONSORED BY THE

Federal Ministry of Education and Research

kisz@hpi.de

hpi.de/kisz

https://aimaker.community

https://join.aimaker.community

BEISPIEL RAG

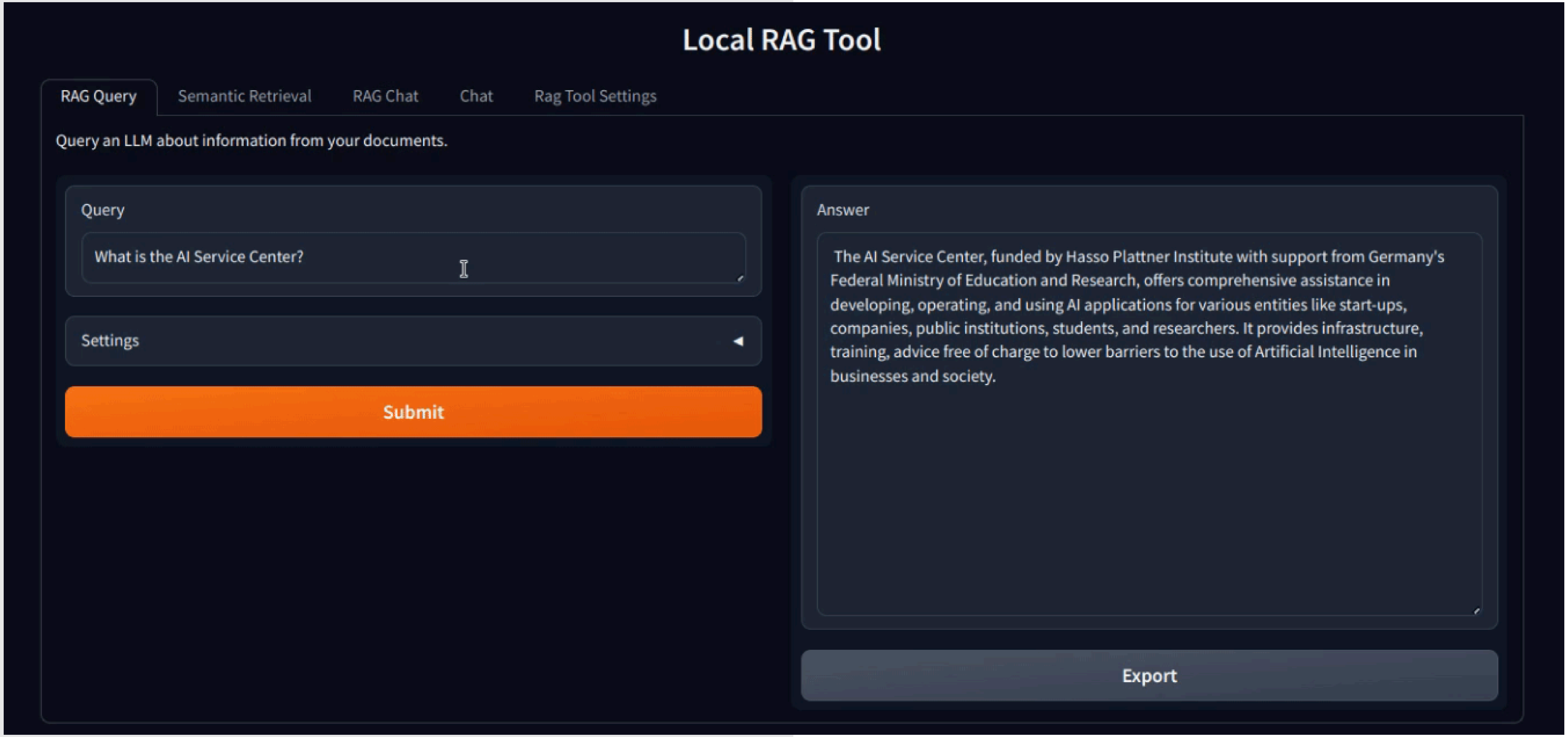
KI Service Zentrum by Hasso-Plattner-Institut HPI

SPONSORED BY THE

Federal Ministry of Education and Research

kisz@hpi.de

hpi.de/kisz



Local RAG Tool

Query an LLM about information from your documents.

Query

What is the AI Service Center?

Answer

The AI Service Center, funded by Hasso Plattner Institute with support from Germany's Federal Ministry of Education and Research, offers comprehensive assistance in developing, operating, and using AI applications for various entities like start-ups, companies, public institutions, students, and researchers. It provides infrastructure, training, advice free of charge to lower barriers to the use of Artificial Intelligence in businesses and society.

https://github.com/aihpi/ragsst

BERATUNG & MENTORING

KI-Sprechstunde

- Beantwortung von Fragen:
    - Allgemein zu KI
    - zu Rechenressourcen und Infrastruktur
    - zu Modellen, Architekturen und Frameworks

KI-Pilotprojekte

- Bewerbungsphase alle drei Monate
- Auswahl anhand von Kriterien (z. B. KI-Reife, Gemeinwohl)
- Co-Entwicklung einer prototypischen Lösung
- Gemeinsame Veröffentlichung der Ergebnisse

Kooperationen

- Gemeinsam organisierte Netzwerktreffen

Design IT. Create Knowledge.

KI Service Zentrum by Hasso-Plattner-Institut HPI

SPONSORED BY THE

Federal Ministry
of Education
and Research

kisz@hpi.de

hpi.de/kisz

HPI ECOSYSTEM

School of Design Thinking Hasso-Plattner-Institut Universität Potsdam HPI

School of Entrepreneurship Hasso-Plattner-Institut HPI

HPI ACADEMY HPI

Mittelstand-Digital
Zentrum
Berlin

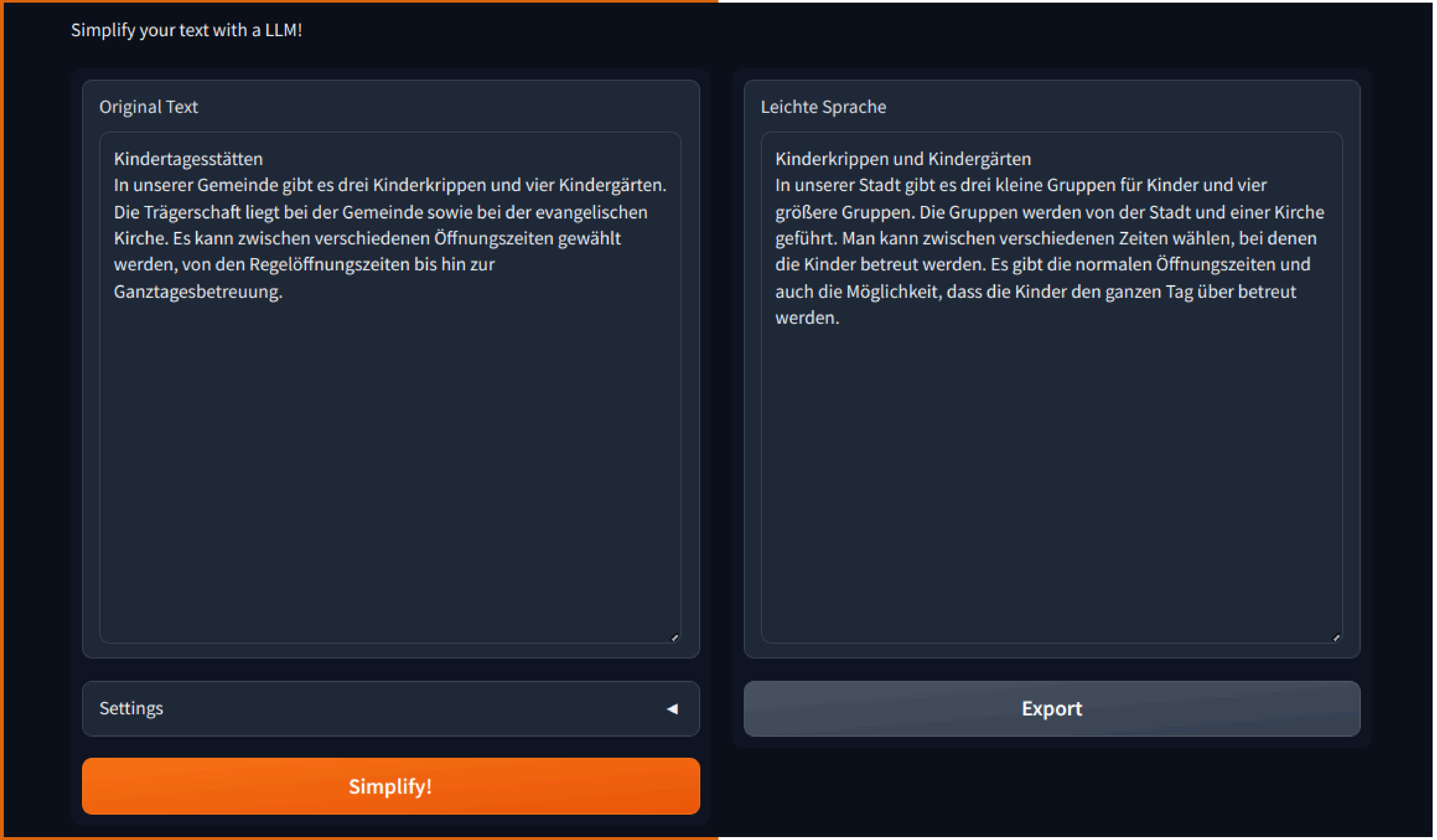
Online Sprechstunde

ÜBERSETZUNG VERWALTUNGSDEUTSCH IN LEICHTE SPRACHE

KI Service Zentrum by Hasso-Plattner-Institut HPI

SPONSORED BY THE Federal Ministry of Education and Research

kisz@hpi.de
hpi.de/kisz



Simplify your text with a LLM!

Original Text

Kindertagesstätten
In unserer Gemeinde gibt es drei Kinderkrippen und vier Kindergärten. Die Trägerschaft liegt bei der Gemeinde sowie bei der evangelischen Kirche. Es kann zwischen verschiedenen Öffnungszeiten gewählt werden, von den Regelöffnungszeiten bis hin zur Ganztagesbetreuung.

Leichte Sprache

Kinderkrippen und Kindergärten
In unserer Stadt gibt es drei kleine Gruppen für Kinder und vier größere Gruppen. Die Gruppen werden von der Stadt und einer Kirche geführt. Man kann zwischen verschiedenen Zeiten wählen, bei denen die Kinder betreut werden. Es gibt die normalen Öffnungszeiten und auch die Möglichkeit, dass die Kinder den ganzen Tag über betreut werden.

https://github.com/aihpi/leichte-sprache

GENERIERUNG VON MATHEMATIK TEXTAUFGABEN

KI Service Zentrum by Hasso-Plattner-Institut

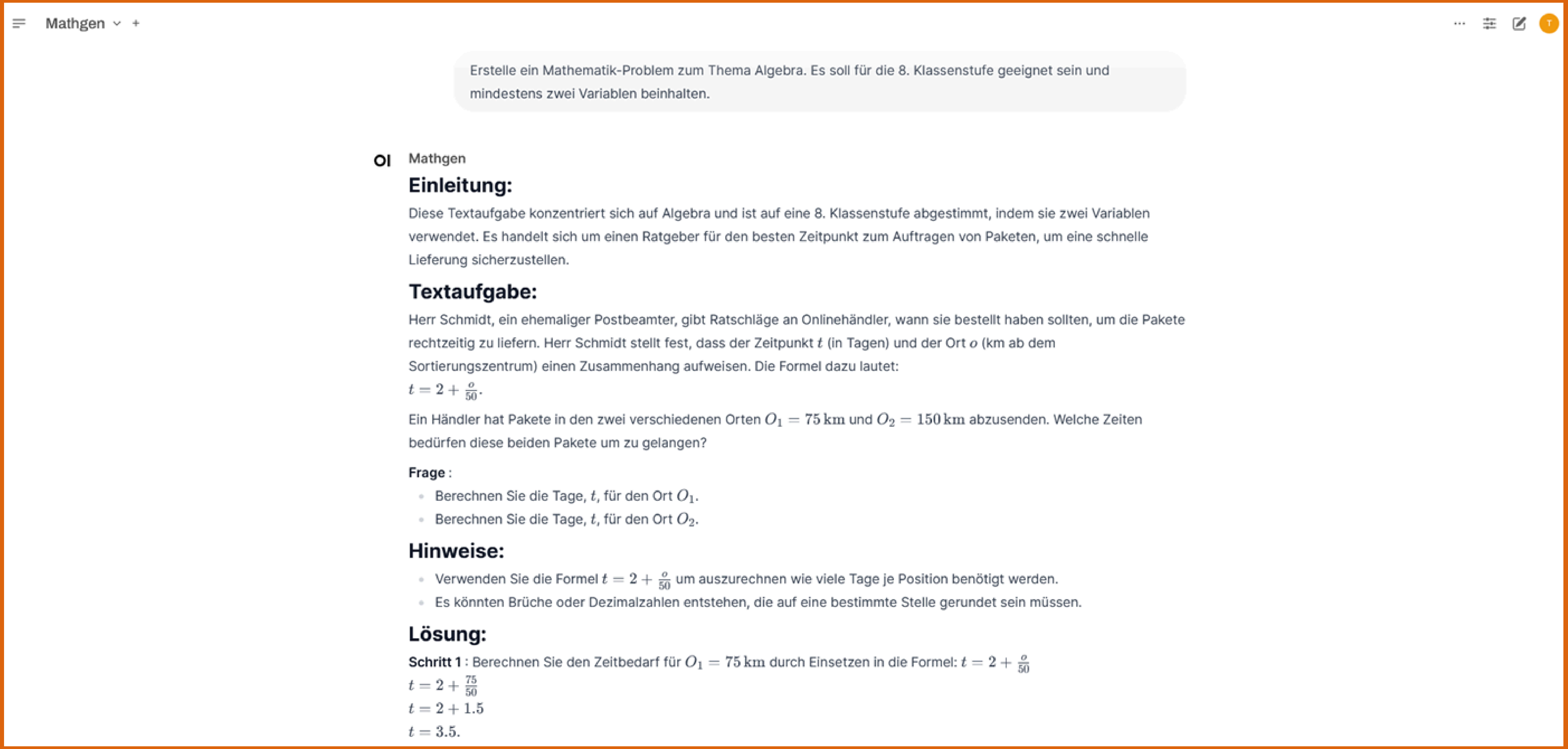
HPI

SPONSORED BY THE

Federal Ministry of Education and Research

kisz@hpi.de

hpi.de/kisz



Erstelle ein Mathematik-Problem zum Thema Algebra. Es soll für die 8. Klassenstufe geeignet sein und mindestens zwei Variablen beinhalten.

Mathgen

Einleitung:

Diese Textaufgabe konzentriert sich auf Algebra und ist auf eine 8. Klassenstufe abgestimmt, indem sie zwei Variablen verwendet. Es handelt sich um einen Ratgeber für den besten Zeitpunkt zum Auftragen von Paketen, um eine schnelle Lieferung sicherzustellen.

Textaufgabe:

Herr Schmidt, ein ehemaliger Postbeamter, gibt Ratschläge an Onlinehändler, wann sie bestellt haben sollten, um die Pakete rechtzeitig zu liefern. Herr Schmidt stellt fest, dass der Zeitpunkt t (in Tagen) und der Ort o (km ab dem Sortierungszentrum) einen Zusammenhang aufweisen. Die Formel dazu lautet:

$t = 2 + \frac{o}{50}$.

Ein Händler hat Pakete in den zwei verschiedenen Orten $O_1 = 75\,\mathrm{km}$ und $O_2 = 150\,\mathrm{km}$ abzusenden. Welche Zeiten bedürfen diese beiden Pakete um zu gelangen?

Frage :

- Berechnen Sie die Tage, t, für den Ort O_1.
- Berechnen Sie die Tage, t, für den Ort O_2.

Hinweise:

- Verwenden Sie die Formel $t = 2 + \frac{o}{50}$ um auszurechnen wie viele Tage je Position benötigt werden.
- Es könnten Brüche oder Dezimalzahlen entstehen, die auf eine bestimmte Stelle gerundet sein müssen.

Lösung:

Schritt 1 : Berechnen Sie den Zeitbedarf für $O_1 = 75\,\mathrm{km}$ durch Einsetzen in die Formel: $t = 2 + \frac{o}{50}$

$t = 2 + \frac{75}{50}$

$t = 2 + 1.5$

$t = 3.5.$

https://mathgen.hpi-sci.de/

UNSERE RESSOURCEN

KI Service Zentrum by Hasso-Plattner-Institut HPI

SPONSORED BY THE

Federal Ministry of Education and Research

kisz@hpi.de

hpi.de/kisz

Training
- 64 NVIDIA H100 GPU

Inference
- 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

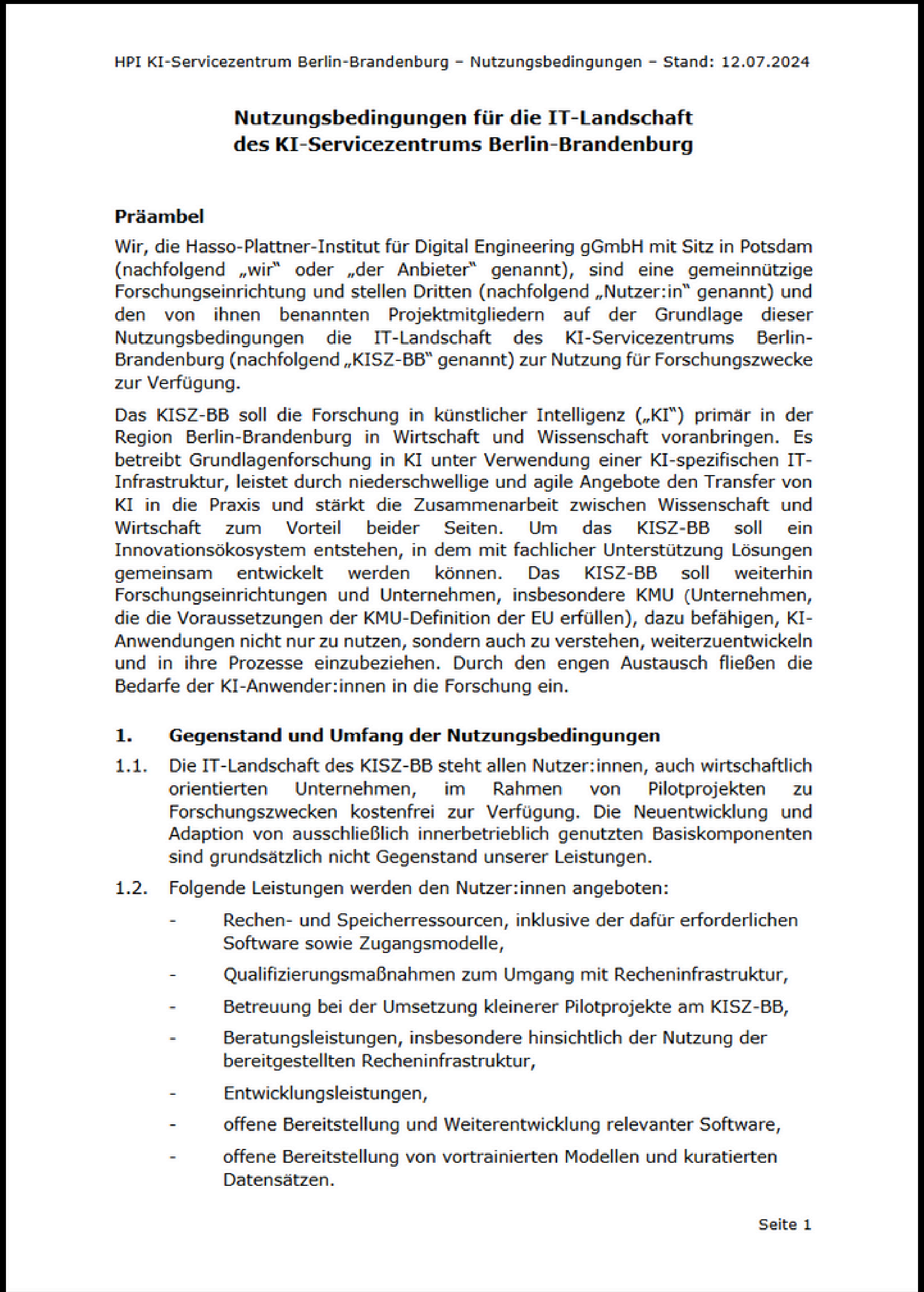
Edge
- ARMv8 CPU
- NVIDIA Jetson AGX Module

Neuromorphic (soon)
- 288 SpiNNaker2 Chips

Storage
- 1.5 PB NVRAM

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

NUTZUNGSBEDINGUNGEN



HPI KI-Servicezentrum Berlin-Brandenburg – Nutzungsbedingungen – Stand: 12.07.2024

Nutzungsbedingungen für die IT-Landschaft
des KI-Servicezentrums Berlin-Brandenburg

Präambel

Wir, die Hasso-Plattner-Institut für Digital Engineering gGmbH mit Sitz in Potsdam (nachfolgend „wir" oder „der Anbieter" genannt), sind eine gemeinnützige Forschungseinrichtung und stellen Dritten (nachfolgend „Nutzer:in" genannt) und den von ihnen benannten Projektmitgliedern auf der Grundlage dieser Nutzungsbedingungen die IT-Landschaft des KI-Servicezentrums Berlin-Brandenburg (nachfolgend „KISZ-BB" genannt) zur Nutzung für Forschungszwecke zur Verfügung.

Das KISZ-BB soll die Forschung in künstlicher Intelligenz („KI") primär in der Region Berlin-Brandenburg in Wirtschaft und Wissenschaft voranbringen. Es betreibt Grundlagenforschung in KI unter Verwendung einer KI-spezifischen IT-Infrastruktur, leistet durch niederschwellige und agile Angebote den Transfer von KI in die Praxis und stärkt die Zusammenarbeit zwischen Wissenschaft und Wirtschaft zum Vorteil beider Seiten. Um das KISZ-BB soll ein Innovationsökosystem entstehen, in dem mit fachlicher Unterstützung Lösungen gemeinsam entwickelt werden können. Das KISZ-BB soll weiterhin Forschungseinrichtungen und Unternehmen, insbesondere KMU (Unternehmen, die die Voraussetzungen der KMU-Definition der EU erfüllen), dazu befähigen, KI-Anwendungen nicht nur zu nutzen, sondern auch zu verstehen, weiterzuentwickeln und in ihre Prozesse einzubeziehen. Durch den engen Austausch fließen die Bedarfe der KI-Anwender:innen in die Forschung ein.

1. Gegenstand und Umfang der Nutzungsbedingungen

1.1. Die IT-Landschaft des KISZ-BB steht allen Nutzer:innen, auch wirtschaftlich orientierten Unternehmen, im Rahmen von Pilotprojekten zu Forschungszwecken kostenfrei zur Verfügung. Die Neuentwicklung und Adaption von ausschließlich innerbetrieblich genutzten Basiskomponenten sind grundsätzlich nicht Gegenstand unserer Leistungen.

1.2. Folgende Leistungen werden den Nutzer:innen angeboten:

- Rechen- und Speicherressourcen, inklusive der dafür erforderlichen Software sowie Zugangsmodelle,
- Qualifizierungsmaßnahmen zum Umgang mit Recheninfrastruktur,
- Betreuung bei der Umsetzung kleinerer Pilotprojekte am KISZ-BB,
- Beratungsleistungen, insbesondere hinsichtlich der Nutzung der bereitgestellten Recheninfrastruktur,
- Entwicklungsleistungen,
- offene Bereitstellung und Weiterentwicklung relevanter Software,
- offene Bereitstellung von vortrainierten Modellen und kuratierten Datensätzen.

Seite 1

- **Allgemein**
  - Zugang **kostenfrei**
  - für **KI-Pilotprojekte**
- **Einschränkungen**
  - kein Produktionsbetrieb
    - Daten sollten **anonymisiert** oder **synthetisiert** sein
    - kein **Hosting** von Produkten
  - **Schadensersatz** ausgeschlossen
  - Verfügbarkeit
- **Verpflichtungen & Rechte**
  - **Reporting** & **Veröffentlichung** durch Nutzende
  - **Altrechte** bleiben bei Nutzenden
  - **Neurechte** bleiben bei Nutzenden
    - --> Einräumen von Nutzungsrechten für Forschung und Lehre

KI Service Zentrum by Hasso-Plattner-Institut HPI

SPONSORED BY THE
Federal Ministry
of Education
and Research

kisz@hpi.de
hpi.de/kisz

Design IT. Create Knowledge.

FORSCHUNG

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

KI Service Zentrum
by Hasso-Plattner-Institut

HPI

SPONSORED BY THE

Federal Ministry
of Education
and Research

kisz@hpi.de

hpi.de/kisz

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

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/sarahESL/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 Computati
May 2-6, 2023 ©2023 Associ

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, Saarbrucken, 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

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 parallel documents 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 documents 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 here[1].

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-Document-level-Neural-Machine-Translation

13

Proceedings of the 17th Conference of the European Chapter of t
May 2-6, 2023 ©2023 Associatio

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.

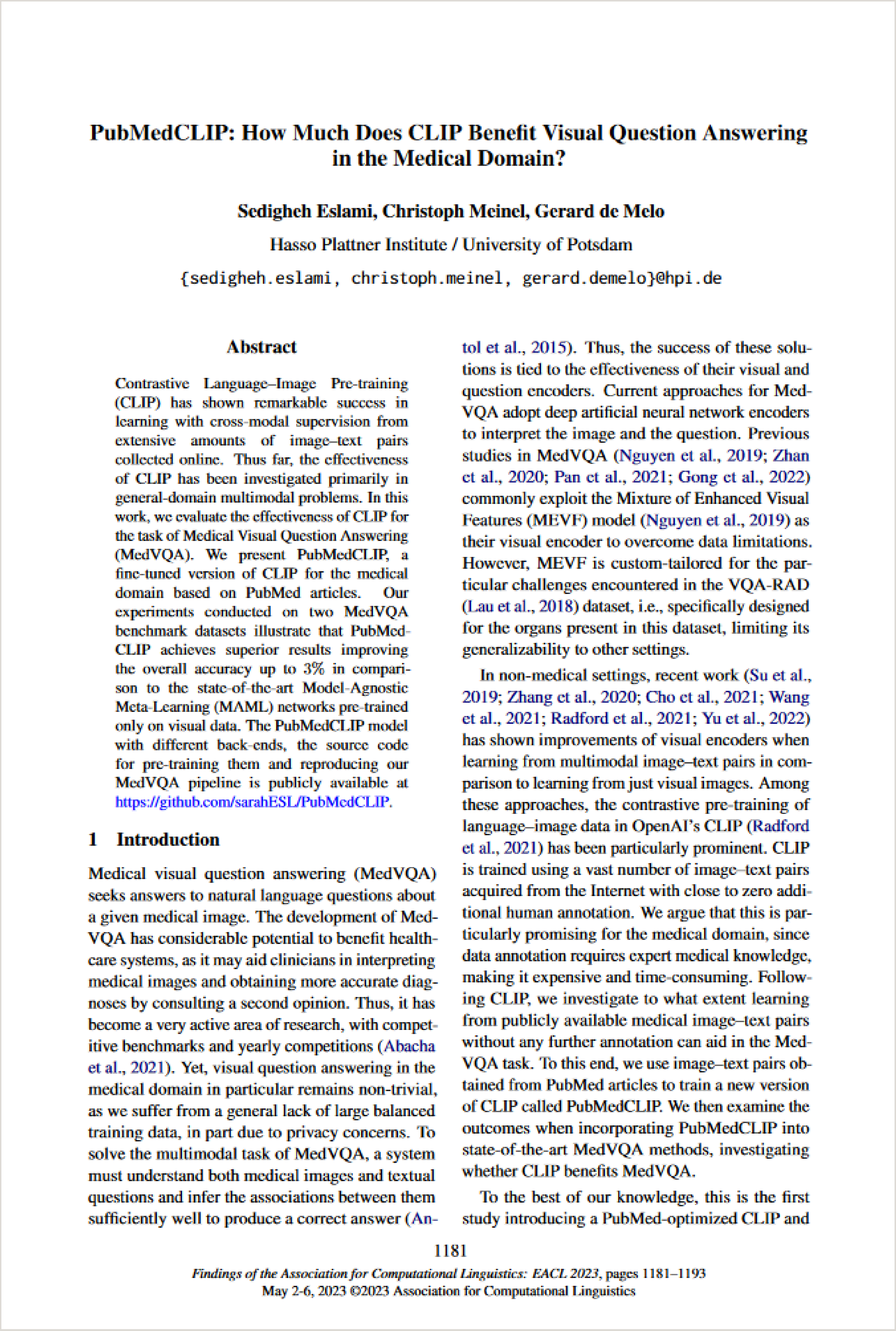
1 Introduction

The number of parameters and computational resources spent on training deep neural networks is growing rapidly [1, 3, 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].

Design IT. Create Knowledge.

RESEARCH

- **AI Methods Research**
  - ○ Visual Question Answering
  - ○ Medical Domain
  - ○ EACL 2023



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

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/sarahESL/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 (Antol et al., 2015). Thus, the success of these solutions is tied to the effectiveness of their visual and question encoders. Current approaches for MedVQA adopt deep artificial neural network encoders to interpret the image and the question. Previous studies in MedVQA (Nguyen et al., 2019; Zhan et al., 2020; Pan et al., 2021; Gong et al., 2022) commonly exploit the Mixture of Enhanced Visual Features (MEVF) model (Nguyen et al., 2019) as their visual encoder to overcome data limitations. However, MEVF is custom-tailored for the particular challenges encountered in the VQA-RAD (Lau et al., 2018) dataset, i.e., specifically designed for the organs present in this dataset, limiting its generalizability to other settings.

In non-medical settings, recent work (Su et al., 2019; Zhang et al., 2020; Cho et al., 2021; Wang et al., 2021; Radford et al., 2021; Yu et al., 2022) has shown improvements of visual encoders when learning from multimodal image–text pairs in comparison to learning from just visual images. Among these approaches, the contrastive pre-training of language–image data in OpenAI's CLIP (Radford et al., 2021) has been particularly prominent. CLIP is trained using a vast number of image–text pairs acquired from the Internet with close to zero additional human annotation. We argue that this is particularly promising for the medical domain, since data annotation requires expert medical knowledge, making it expensive and time-consuming. Following CLIP, we investigate to what extent learning from publicly available medical image–text pairs without any further annotation can aid in the MedVQA task. To this end, we use image–text pairs obtained from PubMed articles to train a new version of CLIP called PubMedCLIP. We then examine the outcomes when incorporating PubMedCLIP into state-of-the-art MedVQA methods, investigating whether CLIP benefits MedVQA.

To the best of our knowledge, this is the first study introducing a PubMed-optimized CLIP and

1181

Findings of the Association for Computational Linguistics: EACL 2023, pages 1181–1193
May 2-6, 2023 ©2023 Association for Computational Linguistics

KI Service Zentrum by Hasso-Plattner-Institut

SPONSORED BY THE

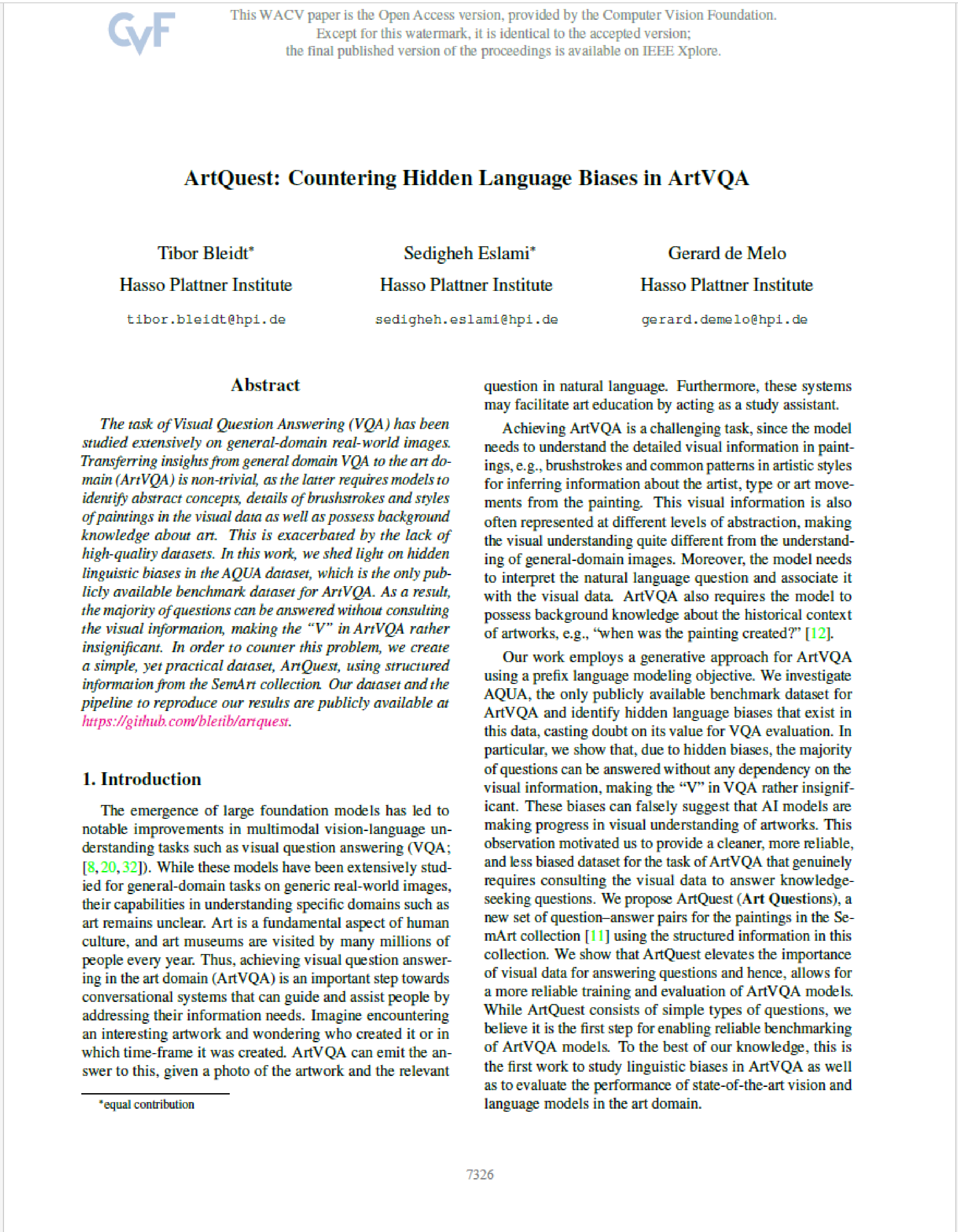
Federal Ministry of Education and Research

kisz@hpi.de

hpi.de/kisz

RESEARCH

- **AI Methods Research**
  - ○ Visual Question Answering
  - ○ Arts Domain
  - ○ WACV 2024



This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

ArtQuest: Countering Hidden Language Biases in ArtVQA

Tibor Bleidt* — Hasso Plattner Institute — tibor.bleidt@hpi.de
Sedigheh Eslami* — Hasso Plattner Institute — sedigheh.eslami@hpi.de
Gerard de Melo — Hasso Plattner Institute — gerard.demelo@hpi.de

Abstract

The task of Visual Question Answering (VQA) has been studied extensively on general-domain real-world images. Transferring insights from general domain VQA to the art domain (ArtVQA) is non-trivial, as the latter requires models to identify abstract concepts, details of brushstrokes and styles of paintings in the visual data as well as possess background knowledge about art. This is exacerbated by the lack of high-quality datasets. In this work, we shed light on hidden linguistic biases in the AQUA dataset, which is the only publicly available benchmark dataset for ArtVQA. As a result, the majority of questions can be answered without consulting the visual information, making the "V" in ArtVQA rather insignificant. In order to counter this problem, we create a simple, yet practical dataset, ArtQuest, using structured information from the SemArt collection. Our dataset and the pipeline to reproduce our results are publicly available at https://github.com/bletib/artquest.

1. Introduction

The emergence of large foundation models has led to notable improvements in multimodal vision-language understanding tasks such as visual question answering (VQA; [8, 20, 32]). While these models have been extensively studied for general-domain tasks on generic real-world images, their capabilities in understanding specific domains such as art remains unclear. Art is a fundamental aspect of human culture, and art museums are visited by many millions of people every year. Thus, achieving visual question answering in the art domain (ArtVQA) is an important step towards conversational systems that can guide and assist people by addressing their information needs. Imagine encountering an interesting artwork and wondering who created it or in which time-frame it was created. ArtVQA can emit the answer to this, given a photo of the artwork and the relevant question in natural language. Furthermore, these systems may facilitate art education by acting as a study assistant.

Achieving ArtVQA is a challenging task, since the model needs to understand the detailed visual information in paintings, e.g., brushstrokes and common patterns in artistic styles for inferring information about the artist, type or art movements from the painting. This visual information is also often represented at different levels of abstraction, making the visual understanding quite different from the understanding of general-domain images. Moreover, the model needs to interpret the natural language question and associate it with the visual data. ArtVQA also requires the model to possess background knowledge about the historical context of artworks, e.g., "when was the painting created?" [12].

Our work employs a generative approach for ArtVQA using a prefix language modeling objective. We investigate AQUA, the only publicly available benchmark dataset for ArtVQA and identify hidden language biases that exist in this data, casting doubt on its value for VQA evaluation. In particular, we show that, due to hidden biases, the majority of questions can be answered without any dependency on the visual information, making the "V" in VQA rather insignificant. These biases can falsely suggest that AI models are making progress in visual understanding of artworks. This observation motivated us to provide a cleaner, more reliable, and less biased dataset for the task of ArtVQA that genuinely requires consulting the visual data to answer knowledge-seeking questions. We propose ArtQuest (**Art Quest**ions), a new set of question–answer pairs for the paintings in the SemArt collection [11] using the structured information in this collection. We show that ArtQuest elevates the importance of visual data for answering questions and hence, allows for a more reliable training and evaluation of ArtVQA models. While ArtQuest consists of simple types of questions, we believe it is the first step for enabling reliable benchmarking of ArtVQA models. To the best of our knowledge, this is the first work to study linguistic biases in ArtVQA as well as to evaluate the performance of state-of-the-art vision and language models in the art domain.

*equal contribution

7326

KI Service Zentrum by Hasso-Plattner-Institut HPI

SPONSORED BY THE

Federal Ministry of Education and Research

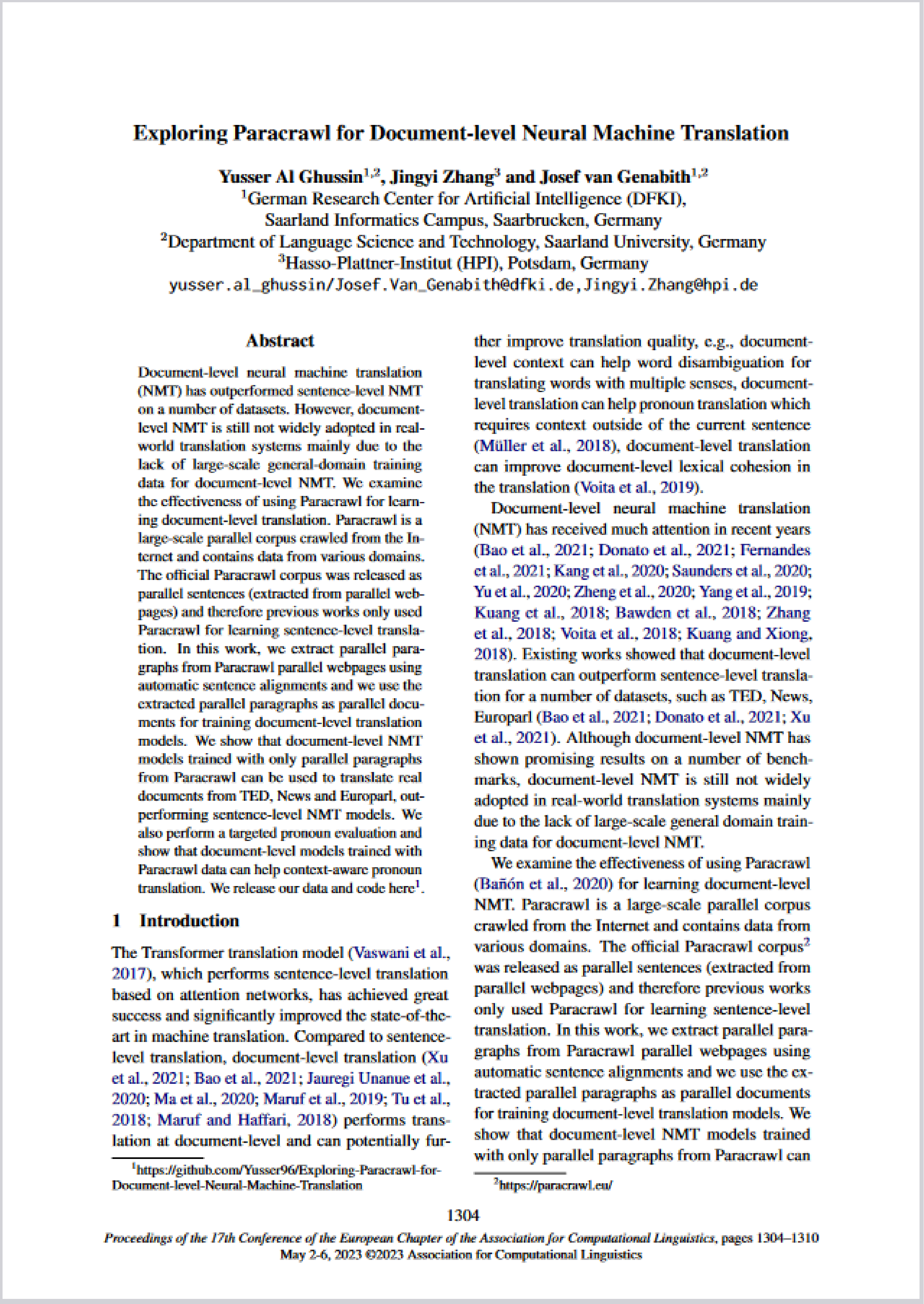
kisz@hpi.de

hpi.de/kisz

Design IT. Create Knowledge.

RESEARCH

- **AI Methods Research**
  - Machine Translation
  - Parallel Paragraph Extraction
  - EACL 2023



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, Saarbrucken, 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

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 parallel documents 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 documents 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 here[1].

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 further improve translation quality, e.g., document-level context can help word disambiguation for translating words with multiple senses, document-level translation can help pronoun translation which requires context outside of the current sentence (Müller et al., 2018), document-level translation can improve document-level lexical cohesion in the translation (Voita et al., 2019).

Document-level neural machine translation (NMT) has received much attention in recent years (Bao et al., 2021; Donato et al., 2021; Fernandes et al., 2021; Kang et al., 2020; Saunders et al., 2020; Yu et al., 2020; Zheng et al., 2020; Yang et al., 2019; Kuang et al., 2018; Bawden et al., 2018; Zhang et al., 2018; Voita et al., 2018; Kuang and Xiong, 2018). Existing works showed that document-level translation can outperform sentence-level translation for a number of datasets, such as TED, News, Europarl (Bao et al., 2021; Donato et al., 2021; Xu et al., 2021). Although document-level NMT has shown promising results on a number of benchmarks, 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 (Bañón et al., 2020) for learning document-level NMT. Paracrawl is a large-scale parallel corpus crawled from the Internet and contains data from various domains. The official Paracrawl corpus[2] 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 parallel documents for training document-level translation models. We show that document-level NMT models trained with only parallel paragraphs from Paracrawl can

[1] https://github.com/Yusser96/Exploring-Paracrawl-for-Document-level-Neural-Machine-Translation

[2] https://paracrawl.eu/

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1304–1310
May 2-6, 2023 ©2023 Association for Computational Linguistics

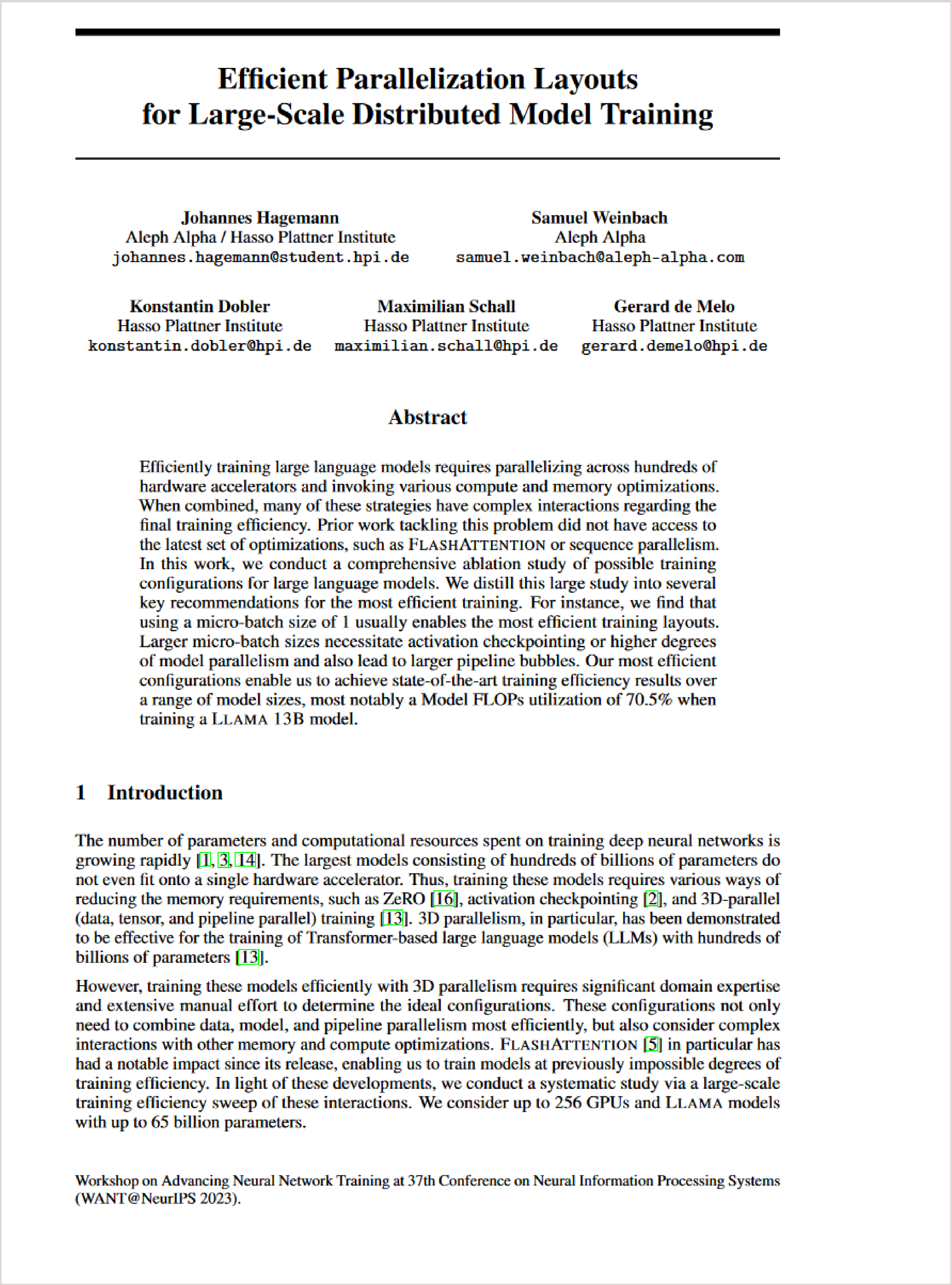
KI Service Zentrum
by Hasso-Plattner-Institut
HPI

SPONSORED BY THE
Federal Ministry of Education and Research

kisz@hpi.de
hpi.de/kisz

RESEARCH

- **AI Methods Research**
  - LLM research
  - Increasing Training Efficiency
  - Best Pager Award NeurIPS 2023



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.

1 Introduction

The number of parameters and computational resources spent on training deep neural networks is growing rapidly [1, 3, 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).

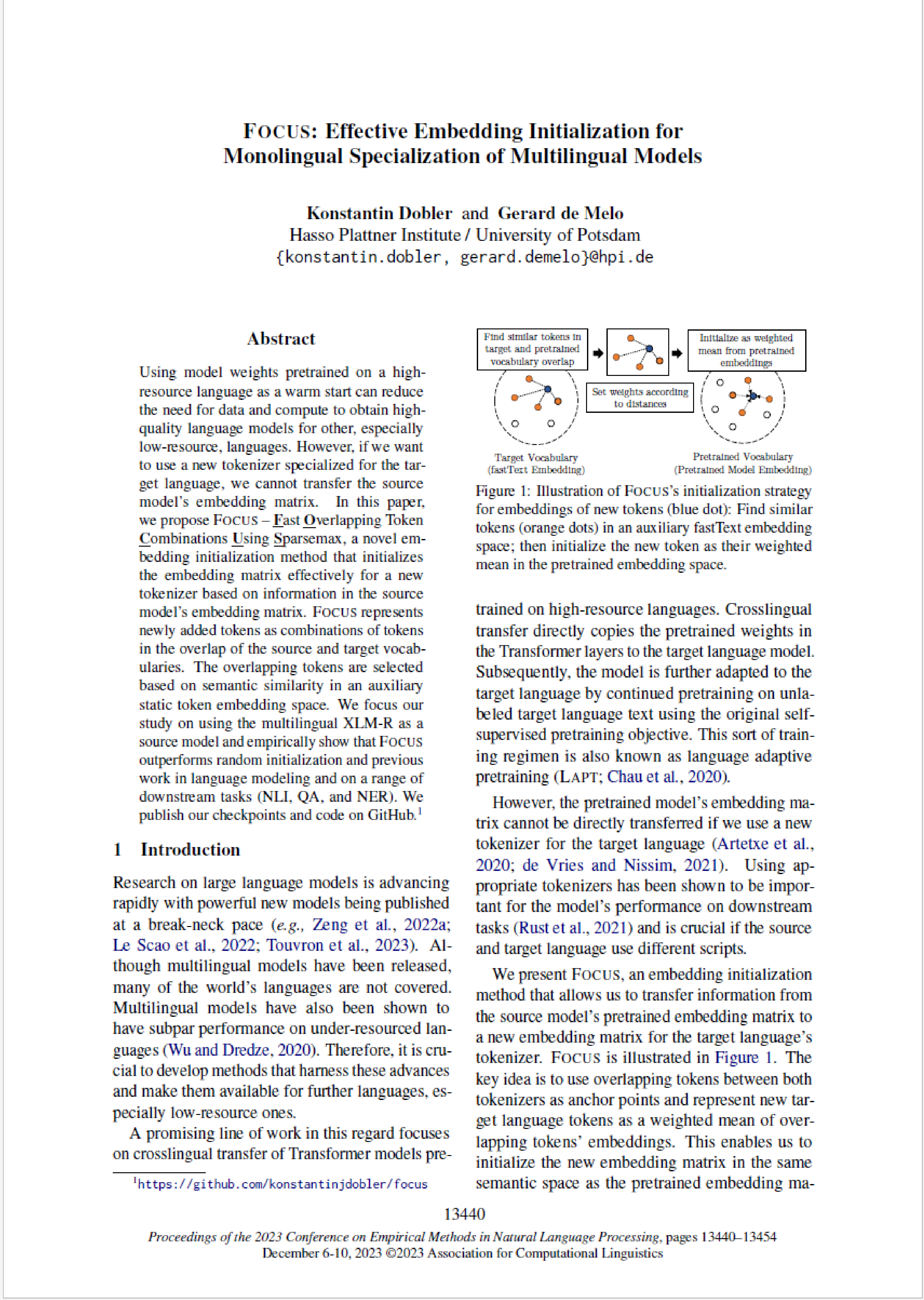
KI Service Zentrum
by Hasso-Plattner-Institut
HPI

SPONSORED BY THE
Federal Ministry of Education and Research

kisz@hpi.de
hpi.de/kisz

RESEARCH

- **AI Methods Research**
  - Machine Translation
  - Adaptation to languages with sparse training material
  - EMNLP 2023



FOCUS: Effective Embedding Initialization for Monolingual Specialization of Multilingual Models

Konstantin Dobler and **Gerard de Melo**
Hasso Plattner Institute / University of Potsdam
{konstantin.dobler, gerard.demelo}@hpi.de

Abstract

Using model weights pretrained on a high-resource language as a warm start can reduce the need for data and compute to obtain high-quality language models for other, especially low-resource, languages. However, if we want to use a new tokenizer specialized for the target language, we cannot transfer the source model's embedding matrix. In this paper, we propose FOCUS – Fast Overlapping Token Combinations Using Sparsemax, a novel embedding initialization method that initializes the embedding matrix effectively for a new tokenizer based on information in the source model's embedding matrix. FOCUS represents newly added tokens as combinations of tokens in the overlap of the source and target vocabularies. The overlapping tokens are selected based on semantic similarity in an auxiliary static token embedding space. We focus our study on using the multilingual XLM-R as a source model and empirically show that FOCUS outperforms random initialization and previous work in language modeling and on a range of downstream tasks (NLI, QA, and NER). We publish our checkpoints and code on GitHub.[1]

1 Introduction

Research on large language models is advancing rapidly with powerful new models being published at a break-neck pace (*e.g.,* Zeng et al., 2022a; Le Scao et al., 2022; Touvron et al., 2023). Although multilingual models have been released, many of the world's languages are not covered. Multilingual models have also been shown to have subpar performance on under-resourced languages (Wu and Dredze, 2020). Therefore, it is crucial to develop methods that harness these advances and make them available for further languages, especially low-resource ones.

Find similar tokens in target and pretrained vocabulary overlap
Set weights according to distances
Initialize as weighted mean from pretrained embeddings
Target Vocabulary (fastText Embedding)
Pretrained Vocabulary (Pretrained Model Embedding)

Figure 1: Illustration of FOCUS's initialization strategy for embeddings of new tokens (blue dot): Find similar tokens (orange dots) in an auxiliary fastText embedding space; then initialize the new token as their weighted mean in the pretrained embedding space.

A promising line of work in this regard focuses on crosslingual transfer of Transformer models pretrained on high-resource languages. Crosslingual transfer directly copies the pretrained weights in the Transformer layers to the target language model. Subsequently, the model is further adapted to the target language by continued pretraining on unlabeled target language text using the original self-supervised pretraining objective. This sort of training regimen is also known as language adaptive pretraining (LAPT; Chau et al., 2020).

However, the pretrained model's embedding matrix cannot be directly transferred if we use a new tokenizer for the target language (Artetxe et al., 2020; de Vries and Nissim, 2021). Using appropriate tokenizers has been shown to be important for the model's performance on downstream tasks (Rust et al., 2021) and is crucial if the source and target language use different scripts.

We present FOCUS, an embedding initialization method that allows us to transfer information from the source model's pretrained embedding matrix to a new embedding matrix for the target language's tokenizer. FOCUS is illustrated in Figure 1. The key idea is to use overlapping tokens between both tokenizers as anchor points and represent new target language tokens as a weighted mean of overlapping tokens' embeddings. This enables us to initialize the new embedding matrix in the same semantic space as the pretrained embedding ma-

[1] https://github.com/konstantinjdobler/focus

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 13440–13454
December 6-10, 2023 ©2023 Association for Computational Linguistics

KI Service Zentrum
by Hasso-Plattner-Institut
HPI

SPONSORED BY THE
Federal Ministry of Education and Research

kisz@hpi.de
hpi.de/kisz

Design IT. Create Knowledge.

KI Service Zentrum HPI
by Hasso-Plattner-Institut

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

HPI

in

kisz@hpi.de
hpi.de/kisz

Design IT.
Create Knowledge.

www.hpi.de