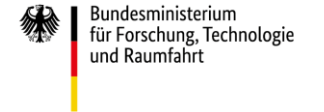


Gefördert durch:

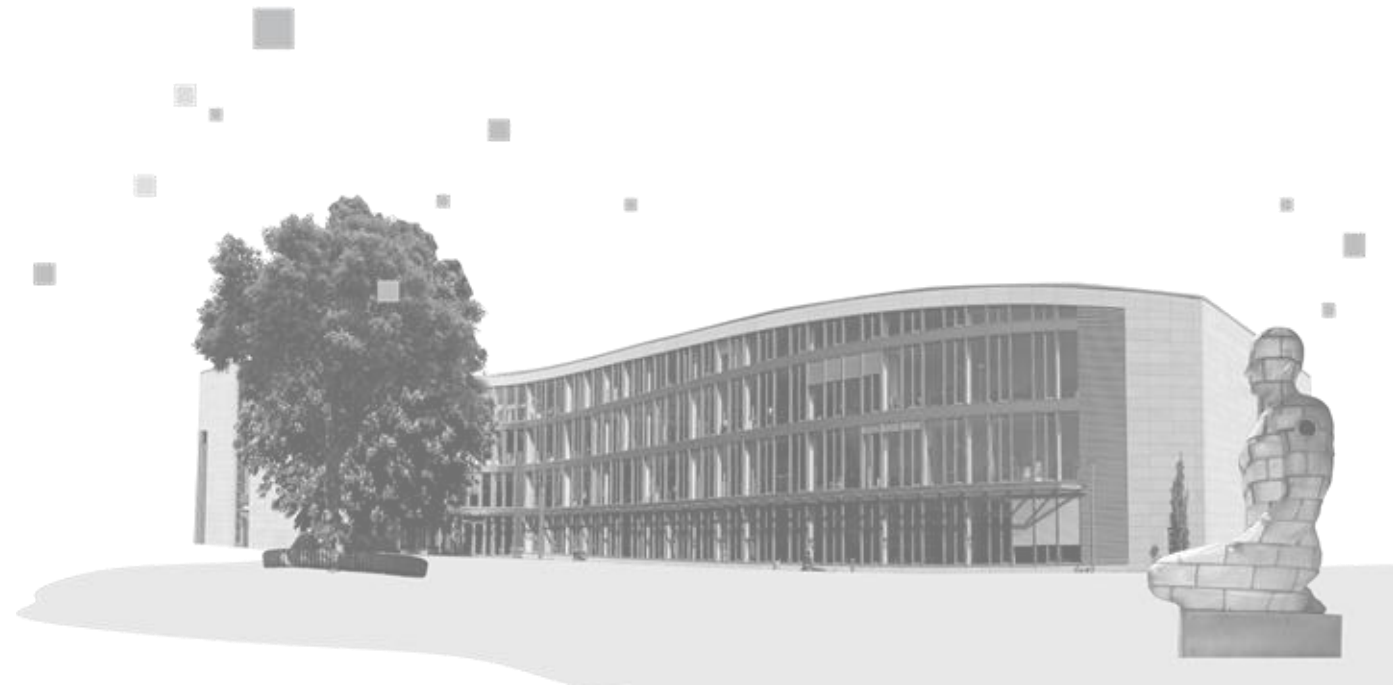


# Erste Schritte zur eigenen KI- Anwendung

Einrichtung der wichtigsten Tools für lokale KI-Entwicklung

**Design IT.  
Create Knowledge.**

[www.hpi.de](http://www.hpi.de)



# KI-Entwicklung in der Praxis

- Ich möchte Code entwickeln, versionieren und gemeinsam pflegen.
- Ich möchte bestehenden Code und externe Bibliotheken zuverlässig integrieren.
- Ich möchte heterogene Datentypen (Text, Tabellen, Bilder, Audio, Video) verarbeiten.
- Ich möchte, dass Code überall gleich funktioniert (Laptop, Server, Rechenzentrum).
- Ich möchte Anwendungen intern bereitstellen und betreiben.
- Ich möchte KI-Modelle souverän und datenschutzkonform ausführen.
- Ich möchte skalierbare Rechenressourcen für Training und Inferenz nutzen.
- Ich möchte große Dokumentbestände semantisch durchsuchen.

Gefördert durch:

# Agenda

## 1. Installation

- a) VS Code
- b) GitHub
- c) Python
- d) UV
- e) Jupyter
- f) Docker Desktop
- g) Ollama
- h) vLLM
- i) Qdrant

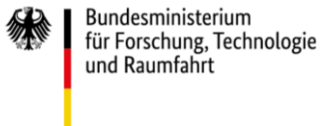
## 2. Outlook

- Workshops des KI-Servicezentrums

Gefördert durch:

# Überblick der installierten Software

Gefördert durch:



# Visual Studio Code (VS Code)

- Code-Editor:
  - Leichtgewichtig
  - Plattformübergreifend
  - Ausbaubar mithilfe von Extensions
- Unterstützt verschiedene Werkzeuge:
  - Python, Jupyter Notebook, Git, Docker, SSH
- Integration von Coding Assistenten (z.B. Claude)

Gefördert durch:

# GitHub & Git

- Plattform für Versionsverwaltung von Code
  - Entwicklung, Versionierung, Überprüfung, Dokumentation
- Reproduktion von KI-Projekten
  - Nützlich für eigene Weiterbildung
  - Verfügbarkeit von Lösungen für eigene Probleme

Gefördert durch:

# Python

- Universelle Programmiersprache mit zentralen Bibliotheken für KI & Data Science
  - NumPy: numerische Berechnungen
  - PyTorch: neuronale Netzwerke
  - Librosa: audio Verarbeitung
  - ...
- Viele Tutorials verfügbar
- Geeignet für Datenverarbeitung, Modelltraining, I/O-Prozesse

Gefördert durch:

# UV

- Schneller Python Paket- & Projektmanager
  - Managt Installation von Paketen
  - Erzeugt reproduzierbare Entwicklungsumgebungen (d.h. gleiche Paket-Versionen)
- Stellt sicher, dass Code auf verschiedenen Maschinen fehlerfrei läuft
  - Windows 10, 11, XP
  - Linux Mint, Ubuntu, Debian
  - MacOS

Gefördert durch:

# Jupyter

- Interaktive Notebook-Umgebung
  - Verarbeitung von Code, Text, Formeln (Markdown), Visualisierungen
  - Ideal für explorative Datenanalyse, Prototyping & Unterricht
- Einfaches Debugging
  - Code-Blöcke können unabhängig voneinander ausgeführt werden

Gefördert durch:

# Docker

- Container-Plattform, die Laufzeitumgebungen reproduzierbar verpackt
  - Ermöglicht einfaches Ausführen von Anwendungen auf verschiedenen Maschinen, Servern & High Performance Clustern (HPC)
- Ideal für komplexe Software Stacks

Gefördert durch:

# Ollama & vLLM

- Tools zum lokalen Ausführen & Verwalten von großen Sprachmodellen
- Schnelles lokales Testen von KI-Anwendungen
- Betrieb von KI-Anwendungen, die sensible Daten verarbeiten
  - Alternative zu Cloud-Abhängigkeiten
- Ollama unterstützt CPUs & GPUs; relative langsam
- vLLM unterstützt CPUs (Windows, Linux, MacOS) & GPUs (Windows, Linux); MacOS GPU-Unterstützung in Entwicklung

Gefördert durch:

# Qdrant

- Vektor-Datenbank für das Speichern & Durchsuchen von *Embeddings* (semantische Repräsentationen von Texten / Bildern)
- Ermöglicht schnellen Zugriff auf Wissensdatenbanken
- Relevant für RAG-Systeme

Gefördert durch:

# Überblick KI-Servicezentrum

Gefördert durch:



Bundesministerium  
für Forschung, Technologie  
und Raumfahrt

# KI-Servicezentrum Berlin Brandenburg

Struktur



Gefördert durch:



Bundesministerium  
für Forschung, Technologie  
und Raumfahrt



Newsletter

## Bildung

### Talks



[tele-task.de/series/1463](https://tele-task.de/series/1463)

- Gastvorträge über Forschung und Innovation

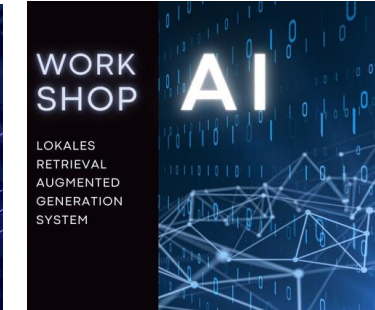
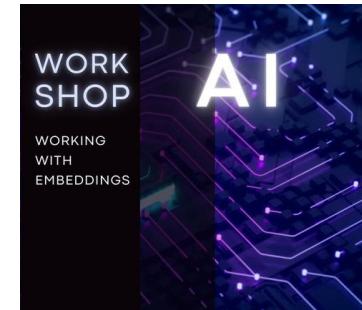


### Workshops



[aimaker.community](https://aimaker.community)

- Praktische Themen
- Beispielthemen: Speech2summary, Docker für ML, Semantische Suche



### MOOCs



[open.hpi.de/channels/ai-service-center](https://open.hpi.de/channels/ai-service-center)

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



Gefördert durch:



Bundesministerium  
für Forschung, Technologie  
und Raumfahrt



### Erste Schritte zu Ihrer eigenen KI-Anwendung

morgen um 10:00

Hasso -Plattner-Institut

Kostenlos



### Guest Talk: Regulatorische Aspekte von KI

Do., 22. Jan., 17:00

Hasso -Plattner-Institut

Kostenlos

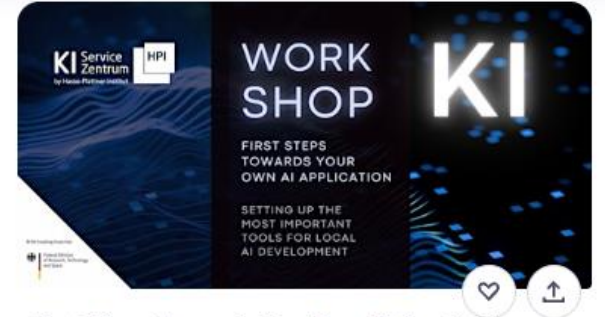


### Effizientes Abfragen von Informationen - Teil I

Di., 27. Jan., 15:00

Hasso -Plattner-Institut

Kostenlos



### First Steps Towards Your Own AI Application

Do., 12. Feb., 15:00

Hasso -Plattner-Institut

Kostenlos



### Effizientes Abfragen von Informationen - Teil II

Mi., 25. Feb., 13:00

Hasso -Plattner-Institut

Kostenlos



### Erste Schritte zu Ihrer eigenen KI-Anwendung

Fr., 13. März, 13:00

Hasso -Plattner-Institut

Kostenlos

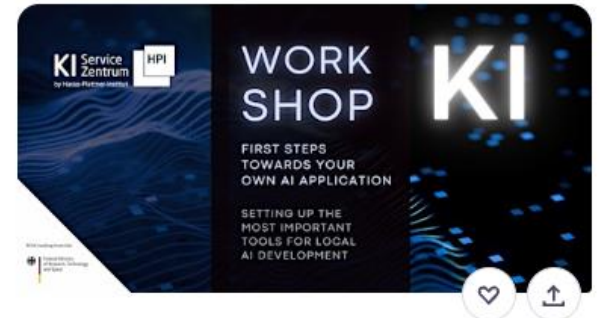


### Effizientes Abfragen von Informationen - Teil III

Do., 26. März, 14:00

Hasso -Plattner-Institut

Kostenlos



### First Steps Towards Your Own AI Application

Mo., 13. Apr., 10:00

Hasso -Plattner-Institut

Kostenlos



### Effizientes Abfragen von Informationen - Teil IV

Di., 28. Apr., 15:00

Hasso -Plattner-Institut

Kostenlos



Sprechstunde buchen

# Beratung

## KI-Sprechstunden

Beantwortung von Fragen

- zu KI-Infrastrukturen
- zu KI-Modellen & -Frameworks
- zu KI-Anwendungsfällen

## KI-Pilotprojekte

Gemeinsame Entwicklung eines Prototyps

- Bewerbung alle drei Monate
- Auswahlkriterien, z. B. KI-Reifegrad, gesellschaftlicher Nutzen
- Veröffentlichung der Ergebnisse

## Kooperation

Gemeinsam organisierte Netzwerktreffen



## Frühere KI-Pilotprojekte

- Transkription Französischer Sprache
- Übersetzung Behördensprache zu Leichte Sprache
- Reduktion von Lebensmittelverschwendung
- Dokumentdatierung anhand von Handschrift



Jetzt bewerben!

KI-Prototyp Leichte-Sprache-Assistent (LeISA)

Vereinfache Texte mit LLMs! - [Zum Projekt auf GitHub](#)

Originaltext

Das KI-Servicezentrum Berlin-Brandenburg ist ein Projekt vom Hasso-Plattner-Institut. Unser Ziel ist, den Einsatz von Künstlicher Intelligenz in Wirtschaft und Gesellschaft zu fördern.

Unser Schwerpunkt liegt auf der Unterstützung von Start-ups, kleinen und mittelständischen Unternehmen, öffentlichen Einrichtungen, Studierenden und der Forschung. Wir bieten kostenfrei die notwendigen Beratungsleistungen, Bildungsangebote und Infrastruktur, um KI-Anwendungen zu entwickeln, zu betreiben und erfolgreich einzusetzen.

Leichte Sprache

\*\*\*Unser Ziel\*\*\*  
Wir möchten, dass viele Menschen und Unternehmen KI-Anwendungen nutzen können.

\*\*\*Wer kann sich an uns wenden?\*\*\*  
Studierende  
Kleine und mittelständische Unternehmen  
Start-ups  
Öffentliche Einrichtungen

\*\*\*Was bieten wir an?\*\*\*  
• Kostenlose Beratung  
• Bildungsangebote  
• Infrastruktur, um KI-Anwendungen zu entwickeln

\*\*\*Beispiele für unsere Arbeit\*\*\*  
Ein Unternehmen möchte eine App entwickeln, die Menschen bei der Einkaufsliste hilft.  
Eine Forschungsgruppe möchte ein System bauen, das Menschen bei der Krankenvorsorge unterstützt.  
Wir helfen ihnen gerne!

Hinweis zur Übersetzung: Dieser Text wurde mit einer KI erstellt.  
Bitte verarbeiten Sie keine personenbezogenen Daten mit dem Tool. Falls der Prototyp einmal nicht verfügbar sein sollte, kontaktieren Sie uns gern, wir kümmern uns darum.

Vereinfachen! Löschen

Der KI-Prototyp LeISA ist ein gemeinsames Projekt der Digitalagentur Brandenburg GmbH und des KI-Servicezentrums (KSZ) vom Hasso-Plattner-Institut.  
\*Fördert durch das Ministerium der Justiz und für Digitalisierung des Landes Brandenburg  
\*Fördert vom Bundesministerium für Bildung und Forschung

[github.com/aihpi/leichte-sprache](https://github.com/aihpi/leichte-sprache)



Zugangsanfrage  
[aisc.hpi.de](https://aisc.hpi.de)

## Infrastruktur

Gefördert durch:



Bundesministerium  
für Forschung, Technologie  
und Raumfahrt



- **Kostenloser Zugang**
- Kein Produktionsbetrieb
  - Daten sollten **anonymisiert** oder **synthetisiert** werden
  - Kein **hosting** von Produkten
- **Berichterstattung & Veröffentlichung** durch Nutzer:innen
- **Alte Rechte** verbleiben bei Nutzer:innen
- **Neue Rechte** verbleiben bei Nutzer:innen
  - Gewährung 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

### Storage

- 1.5 PB NVMe

### Network

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

[kisz@hpi.de](mailto:kisz@hpi.de)

[hpi.de/kisz](https://hpi.de/kisz)

Deine Meinung  
ist wichtig!



QR Code zum  
Feedback-Formular



Gefördert durch: