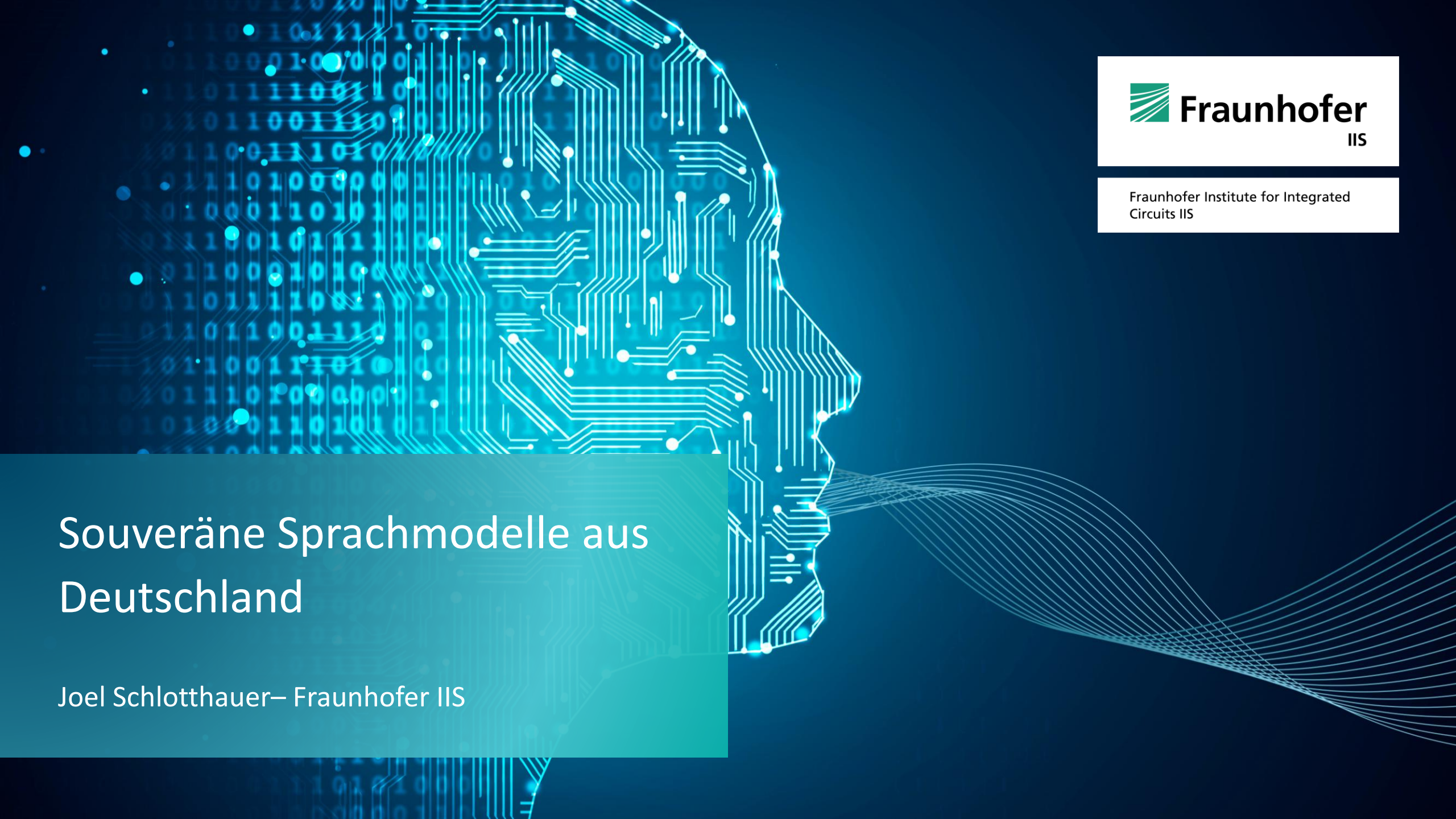


Fraunhofer IIS

Fraunhofer Institute for Integrated Circuits IIS

Souveräne Sprachmodelle aus Deutschland

Joel Schlotthauer– Fraunhofer IIS

Warum brauchen wir deutsche Sprachmodelle?

Souveränität, Wertschöpfung und technologische Basis

1. **Digitale Souveränität**
   Unabhängigkeit bei kritischen Technologien.

2. **Wettbewerbsfähigkeit**
   industrielle Wertschöpfung entlang der gesamten KI-Kette.

3. **Technologische Basis**
   verstehen, anpassen, betreiben und weiterentwickeln können.

4. **Angepasste Modelle**
   für deutsche und industrielle Kontexte.

84%

der Unternehmen, die generative KI nutzen oder es planen, geben an, dass das Herkunftsland des Anbieters „sehr wichtig“ oder „eher wichtig“ ist.

bitkom

Quelle: Bitkom Research 2024

Leistungsniveau auf Standard-Benchmarks vs. industrielle Anforderungen

80 %
Leistungsniveau erreicht

20 %
zur industriellen Reife

Zwei Modellklassen - Ein Kreislauf

Große offene Basismodelle

- An der technologischen Front
- Für komplexe, generalistische Aufgaben
- Erzeugen hochwertige synthetische Daten
- Destillieren und verbessern kleinere Modelle

Kleine spezialisierte Modelle

- In der industriellen Fläche
- Lokal, effizient, latenzarm, kosteneffizient
- Datennah – sensible Daten bleiben im Haus
- Ideal für Geräte, Edge und agentische Systeme

Basiskompetenz • Synthetische Daten • Distillation

Reale Anwendungen • Use Cases • Feedback

ELMOD 2.7B

Sprachmodell aus Deutschland

Prototyp lauffähig auf Smartphone Chip

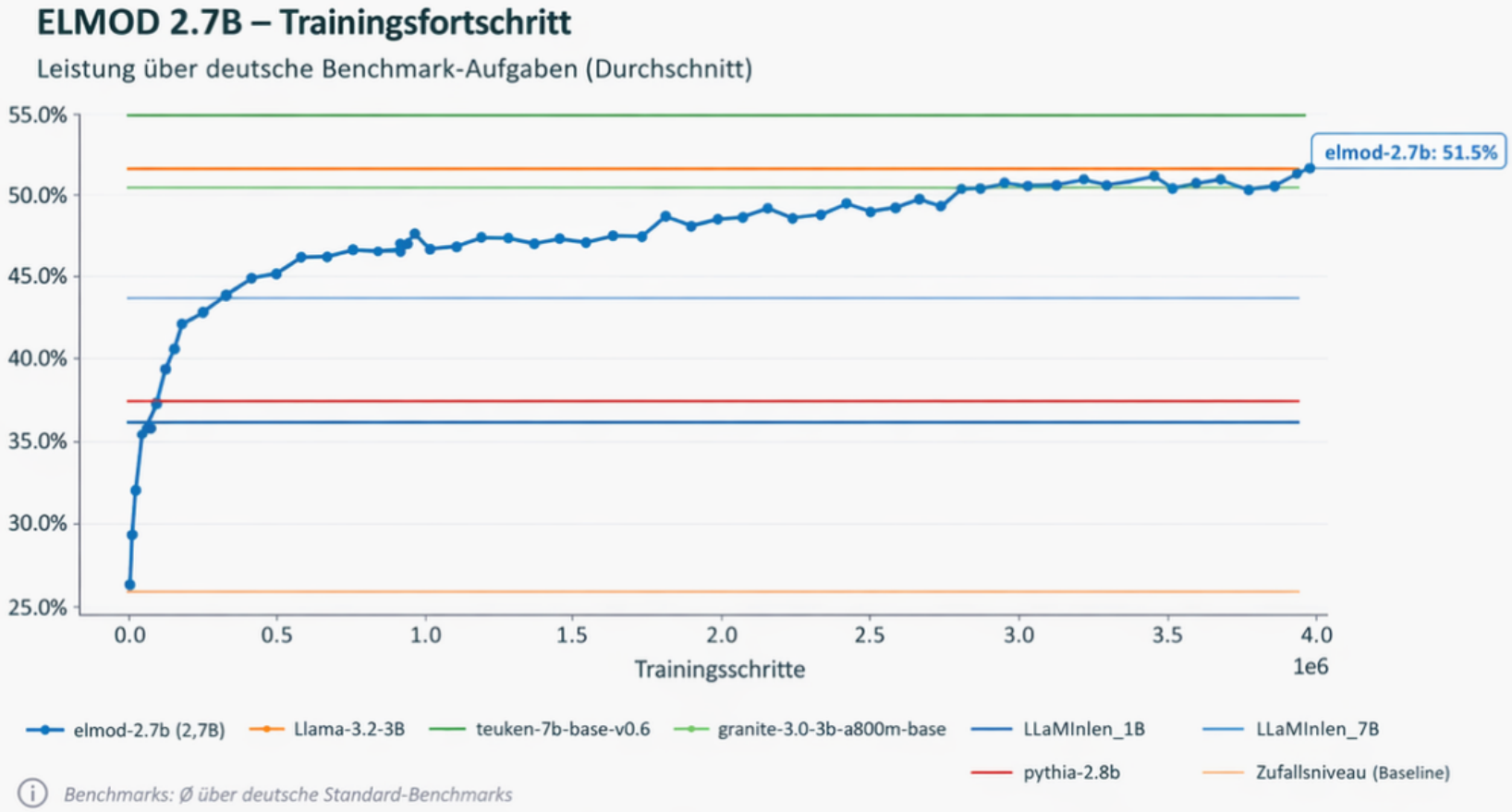
ELMOD 2.7B Referenzprojekt – Lokal-lauffähige Sprachmodelle aus Deutschland

TECHNIK

2,7 Mrd.
Parameter
trainiert auf 3,8T Tokens
(DE + EN + Code)

52k GPUh
auf H100 @ FAU Helma
End-to-end-Pipeline

5,5 PB
kuratierter deutscher
Rohdaten-Korpus



FORSCHUNGSPROTOTYP

ELMOD ist ein **Forschungsprototyp.**
Wir entwickeln heute die Grundlagen für souveräne KI-Modelle, die künftig an vielen Orten sinnvoll eingesetzt werden können.

ELMOD 2.7B – GenAI direkt in der Hosentasche

Referenzprojekt

ANWENDUNGSFALL — GEDANKENSPIEL

Der Servicetechniker beim Kunden.

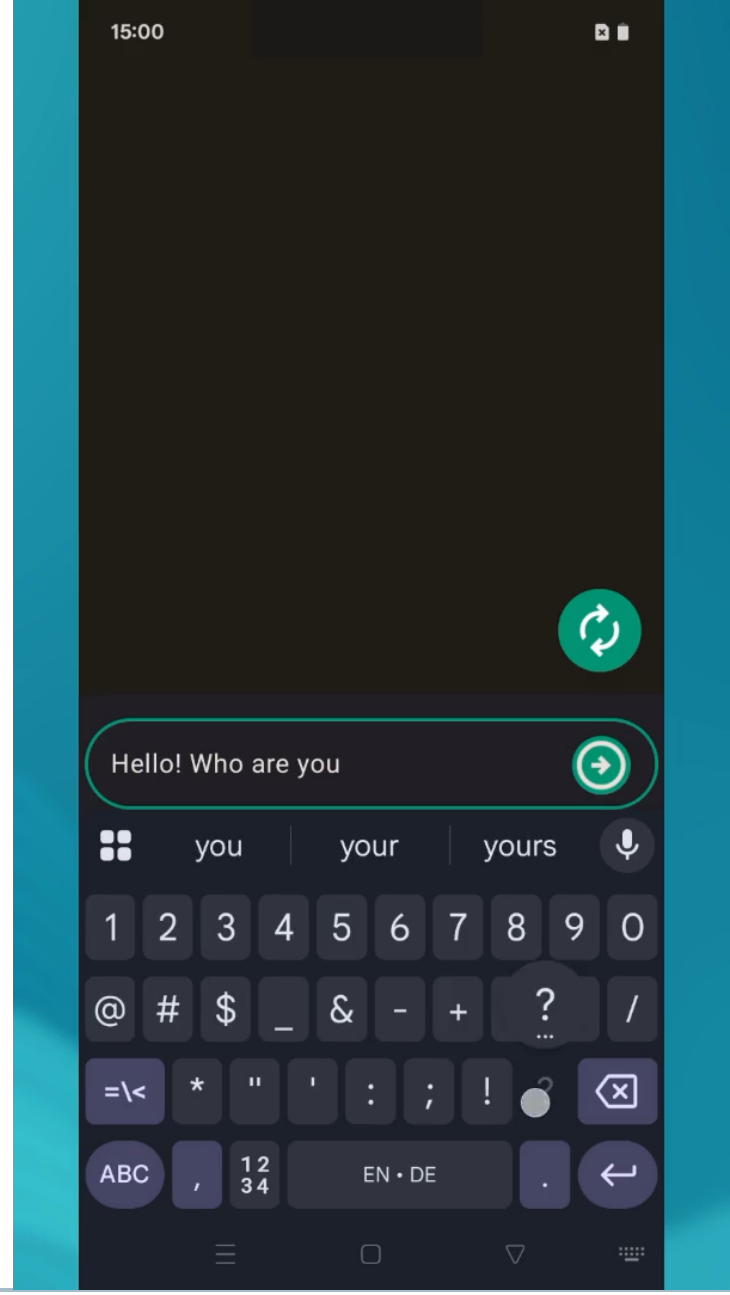
Ein mögliches Zukunftsszenario –
nicht der Entwicklungszweck von ELMOD.

Fehlerlogs | Sensordaten | Zeitreihen | Handbücher

Das Modell **schlussfolgert** über heterogene Quellen.

Offline. Sicher. Auf Deutsch. Datennah.

Lokale Ausführung



SOOFI

Europas Reasoning LLM

Ziel: Europäische KI-Souveränität

SOOFI – Europas Reasoning LLM

Referenzprojekt: Anschluss an die technologische Spitze

Bundesministerium für Wirtschaft und Energie

- SOOFI soll ein offenes KI-Sprachmodell mit rund **100 Mrd. Parametern** entwickeln
- Ziel: **europäische KI-Souveränität** — weniger Abhängigkeit von US-Anbietern.
- Gefördert mit ca. **21 Mio. €** durch das Bundesministerium für Wirtschaft und Energie
- **Reasoning-Modelle** — für komplexe Aufgaben in Industrie & Verwaltung

soofi

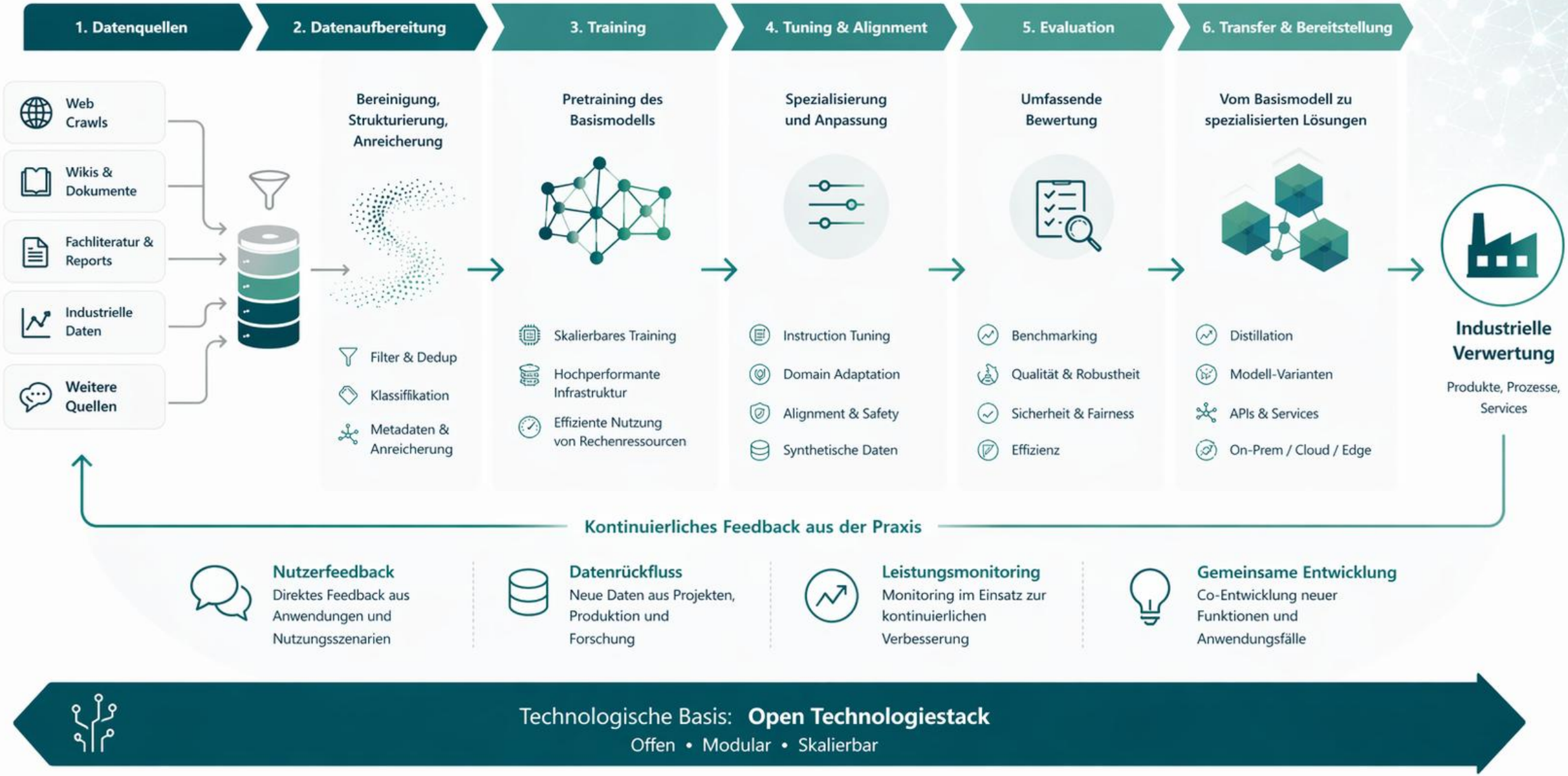
SOOFI – Europas Reasoning LLM

Referenzprojekt

- Training auf der neuen **Industrial AI Cloud der Deutschen Telekom**
- 130 NVIDIA DGX B200 Systemen mit insgesamt über **1.000 GPUs** exklusiv für SOOFI
- Seit **März 2026**

Die gesamte Kette zählt

- Von Daten bis zur Anwendung und zurück



Souveräne KI entsteht durch Kopplung von Forschung und Anwendung

- Große, generalistische Modelle **zunehmend möglich in DE.**
- Deutschland baut **souveräne Fähigkeiten und Infrastruktur** auf.

→ **Forschung ist die Grundlage.**

Im Zusammenspiel entsteht Wirkung.

Nutzung ist entscheidend. Industrielle Modelle für die Praxis.

- **Spezialisierte Modelle** für Industrie und Verwaltung.
- Zugang zu **Daten, Domänenwissen und Anwendungsschicht.**

→ **Anwendung ist der Hebel.**

Die letzten 20% entstehen in der Kopplung von Forschung, Industrie und gemeinsamer Infrastruktur.

Vielen Dank für Ihre Aufmerksamkeit!

Kontakt

Joel Schlotthauer
Gruppenleiter Natural Language Processing
Department Generative KI
joel.schlotthauer@iis.fraunhofer.de

Fraunhofer-Institut für Integrierte Schaltungen IIS
Am Wolfsmantel 33
D-91058 Erlangen

www.iis.fraunhofer.de

Teuken

Teuken 7B – Unser erstes LLM

Referenzprojekt bis Juli 2025

Europäische KI „Made in Germany“ – entwickelt unter der Leitung des Fraunhofer IIS und IAIS

Datenschutzkonform – kann in der eigenen IT betrieben werden; on-premise oder in sicherer Cloud

Mehrsprachig – trainiert auf 24 EU-Sprachen

Energieeffizient – spart Energie beim Training und senkt Kosten bei der Nutzung

Open Source – anpassbar und transparent

ALEPH ALPHA · ControlExpert · dfki ai · Fraunhofer IIS · KI BUNDESVERBAND · Fraunhofer IAIS · IONOS · JÜLICH Forschungszentrum · WDR · ZIH Zentrum für Informationsdienste und Hochleistungsrechnen

Teuken-7B

Downloads und Medienresonanz

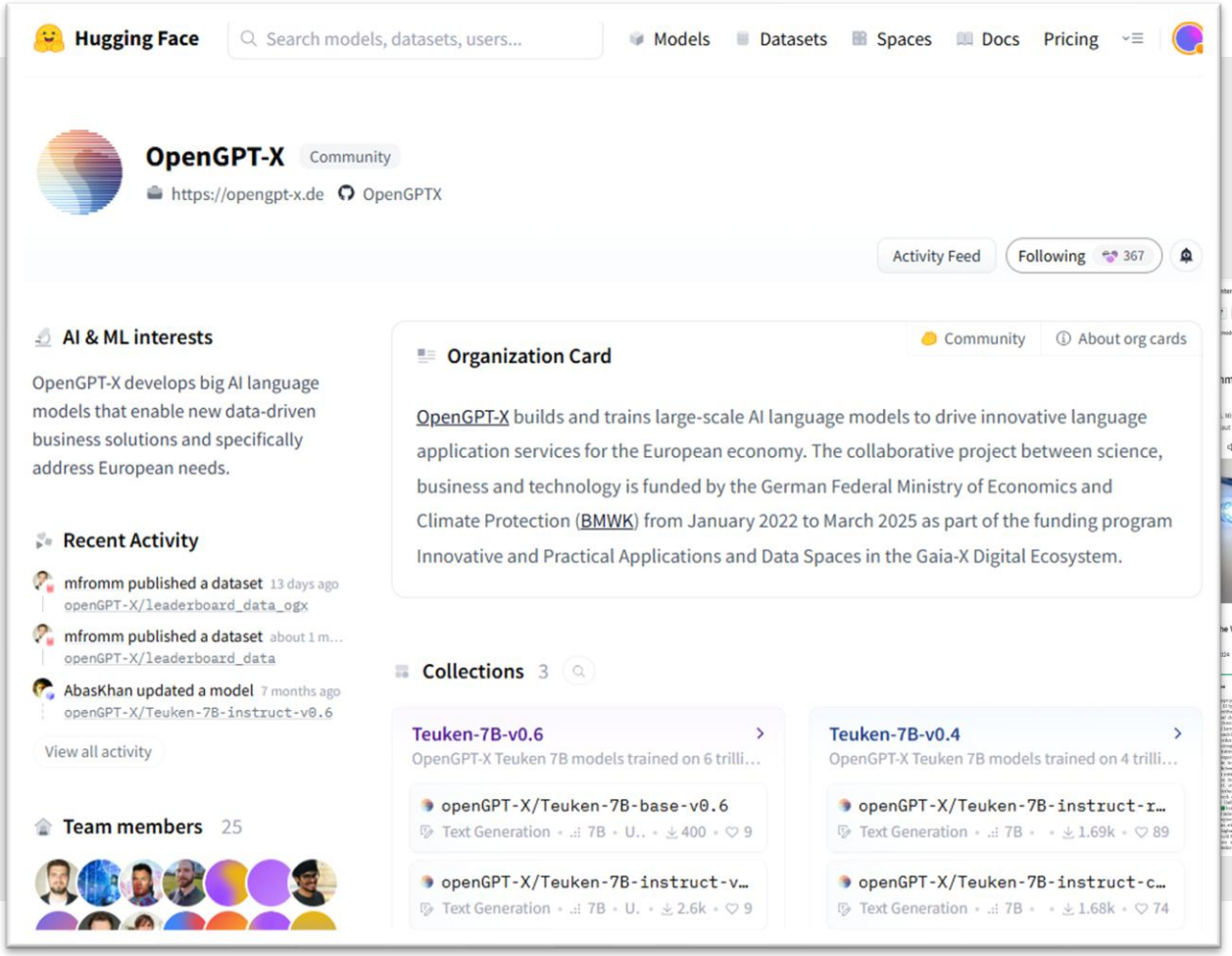
> 60.000 Downloads
der Modelle auf Hugging Face

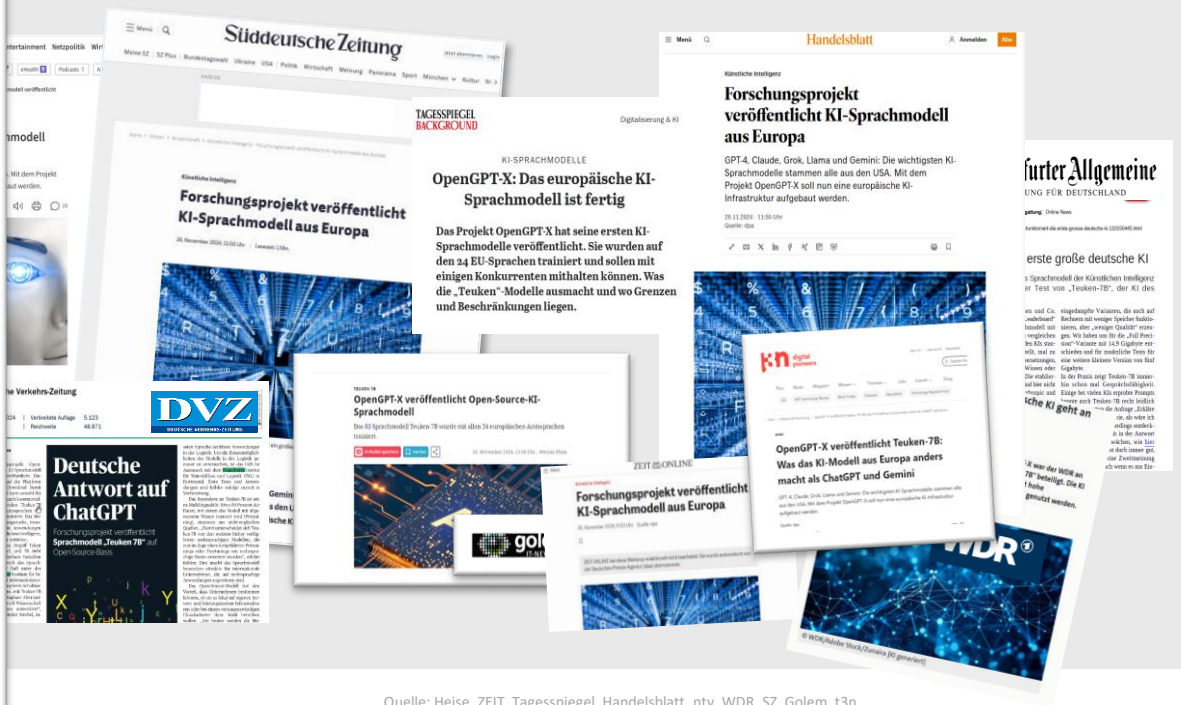
> 50 Presseresonanzen

> 33 Mio. Reichweite
(dpa, heise, FAZ, Handelsblatt etc.)

~ 40.000 Impressions
durch eigen-initiierte Social Media Postings

stand Jahresende 2024





Quelle: Heise, ZEIT, Tagesspiegel, Handelsblatt, ntv, WDR, SZ, Golem, t3n

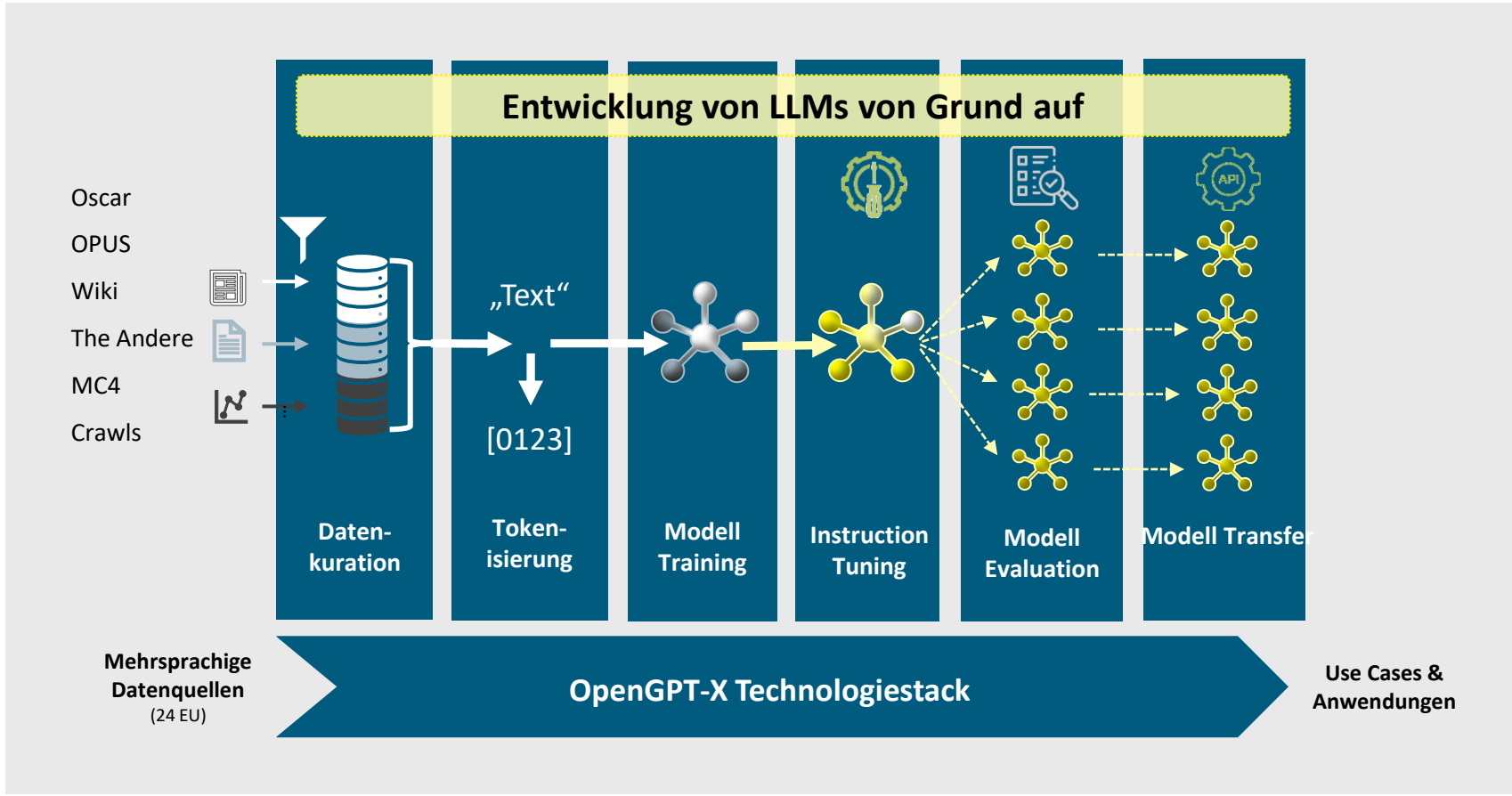
Verteilung der Sprachen

Teuken-7B

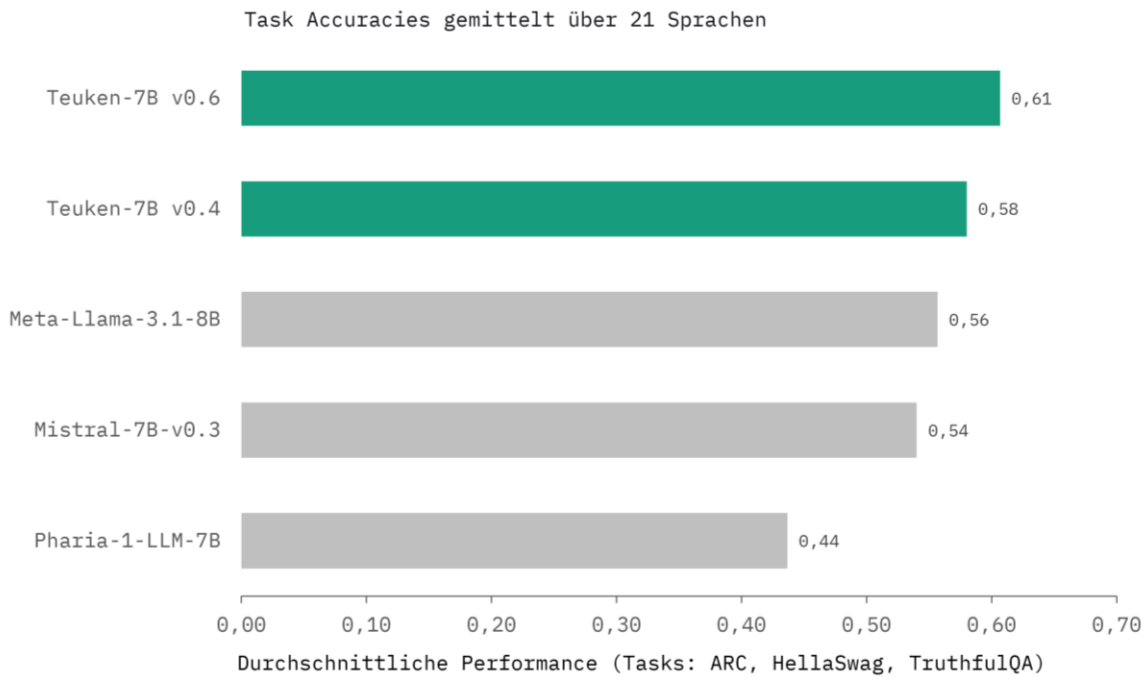
Beherrschung der Technologischen Basis

Technologiestack

- **Große Datensammlung erstellt Mehrsprachig**
  - Nicht nur Common Crawl
- **Datenverarbeitungspipeline**
  - Multilingual
  - Filterungsmethoden produktiv
- **Tokenisierung**
  - Multilingual
- **Skalierung beim Modelltraining**
  - Modalities
  - Megatron
- **Instruction Tuning**
- **Evaluation**
  - European Leaderboard



Teuken-7B: Leistungsfähigkeit gemittelt über 21 Europäische Sprachen



Durchschnitt über 21 Sprachen ohne Maltesisch, Kroatisch, Gaelisch, weil dort keine Evaluierungsdaten vorliegen anhand der drei wichtigsten Benchmarks (HellaSWAG, AI2ARC, TruthfulQA)

- 7 Milliarden Modellparameter
- Multilingual: 24 Europäische Sprachen & multilinguale Tokenisierung
- Open Source (Apache 2.0 + Forschungslizenz)
- Trainiert mit 4 Billionen Tokens und auf JUWELS Booster mittels 512 A100

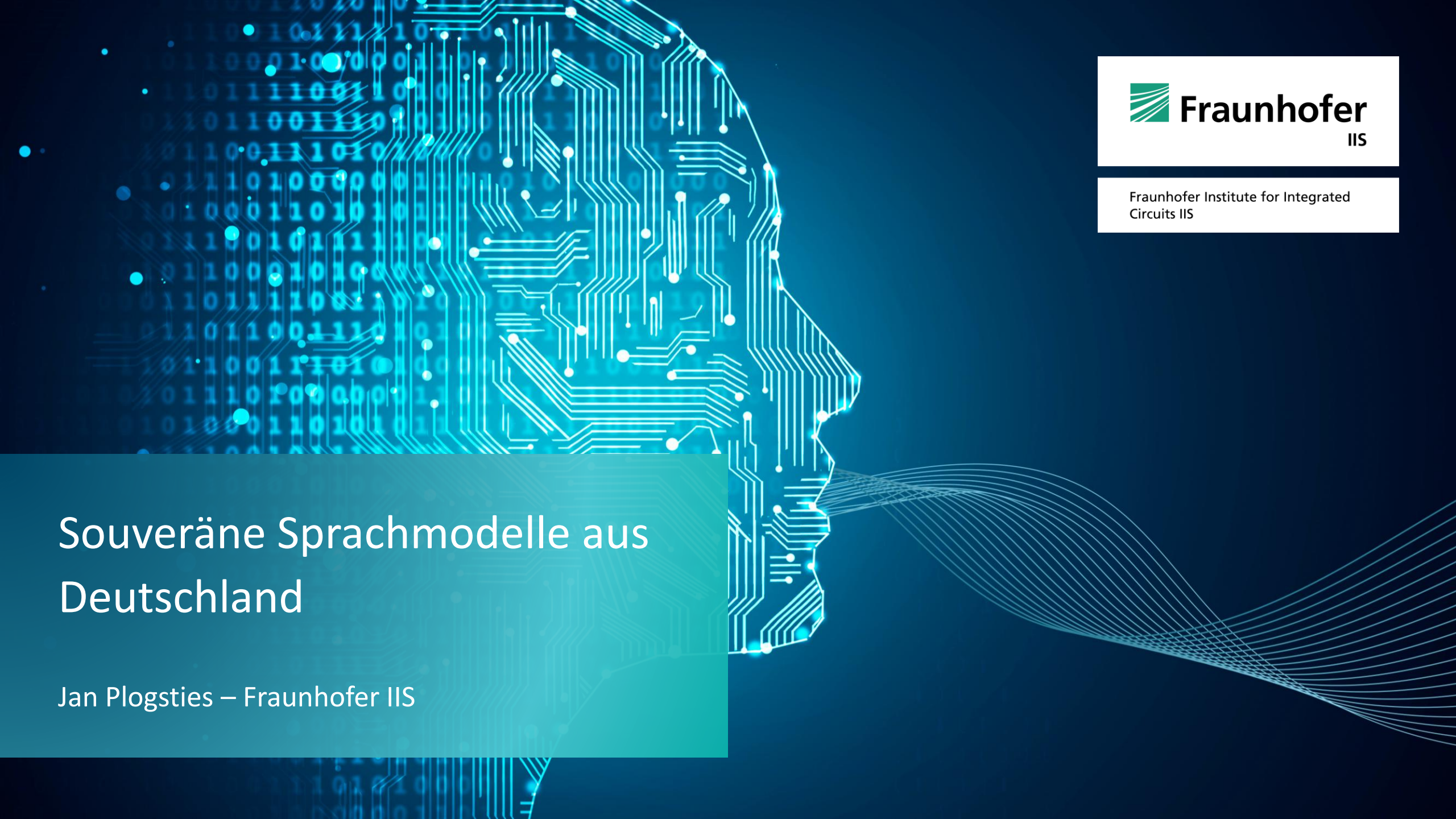
Teuken-7B war im Vergleich mit umfangreicher trainierten Modellen kommerzieller Anbieter wie Mistral (8 Billionen Token) oder Llama3 (15 Billionen Token) kompetitiv

Fraunhofer IIS

Trainingskosten für große Sprachmodelle

Modell	Parameteranzahl	Trainingsaufwand	Quelle
Gemini Ultra	Nicht öffentlich bekannt	Nicht öffentlich bekannt, geschätzte Kosten: 191 Millionen USD	voronoiapp.com
GPT-4	Nicht öffentlich bekannt	Geschätzte Kosten: 78 bis 100 Millionen USD	techradar.com
LLaMA 3	405 Milliarden	30,8 Millionen GPUh	the-decoder.de
DeepSeek V3	671 Milliarden	2,78 Millionen GPUh	the-decoder.de
LLaMA 3.1 8B	8 Milliarden	1,46 Millionen GPUh	huggingface.co
Teuken	7 Milliarden	0,81 Millionen GPUh	

Reine Trainingskosten ohne Experimente und Vorstudien Bekannte Daten zum Trainingszeitpunkt der Modelle; in Deutschland; 1 NVIDIA H100 GPUh einer kostet ca. 2 €



Fraunhofer IIS

Fraunhofer Institute for Integrated Circuits IIS

Souveräne Sprachmodelle aus Deutschland

Jan Plogsties – Fraunhofer IIS

Warum brauchen wir deutsche Sprachmodelle?

Digitale Souveränität statt Abhängigkeit

Wettbewerbsfähigkeit & Wachstum

Technologie beherrschen

Angepasste Modelle

ELMOD 2.7B – Training Progress

Performance across German benchmark tasks

TOKENS

3.8T tokens

Stage 1: 2.4T | Stage 2: 1.4T

COMPUTE

~52k GPUh

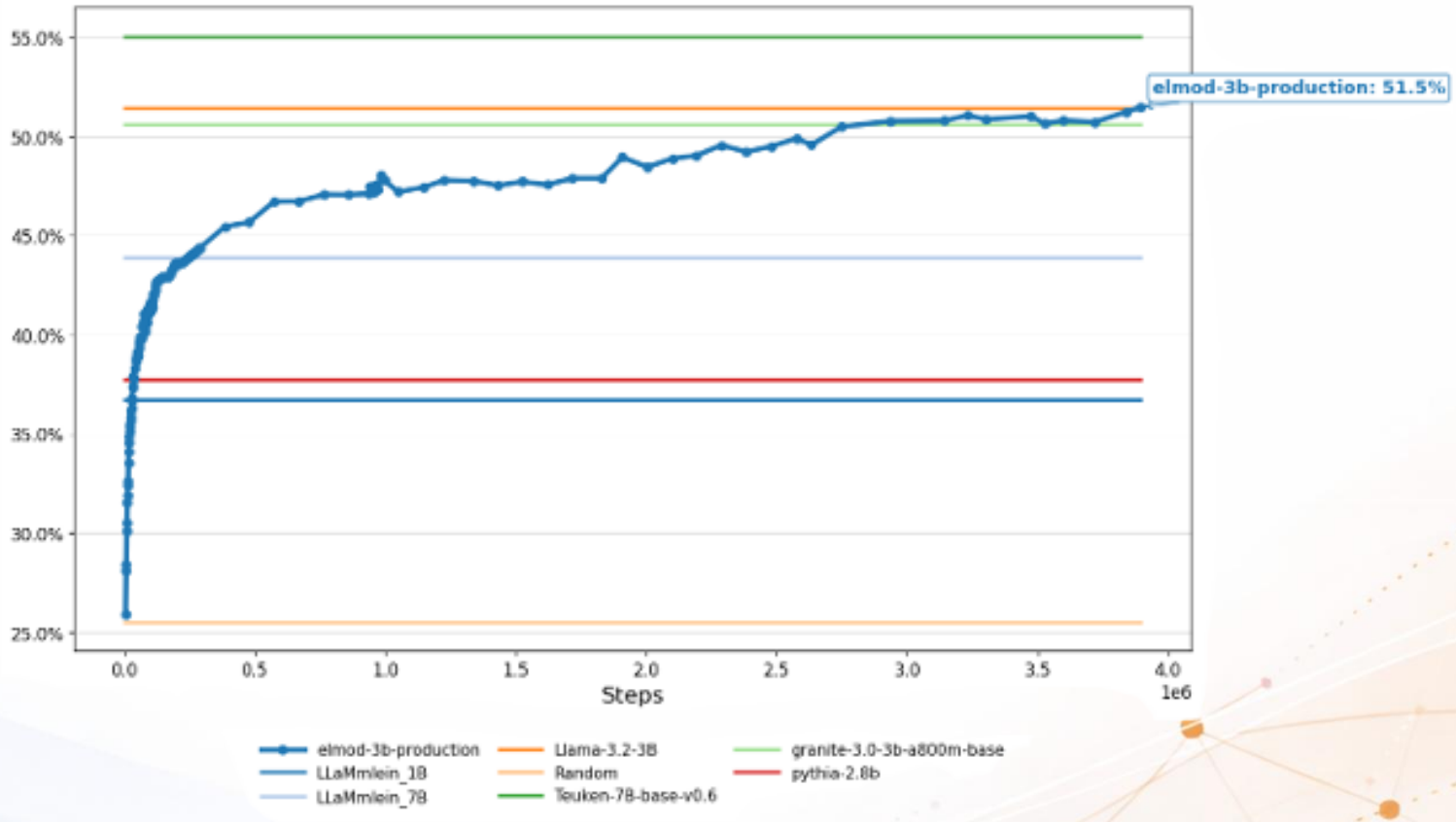
64xH100 @ FAU Helma

MODEL

2.7B parameters

ELMOD

Average across German benchmark tasks

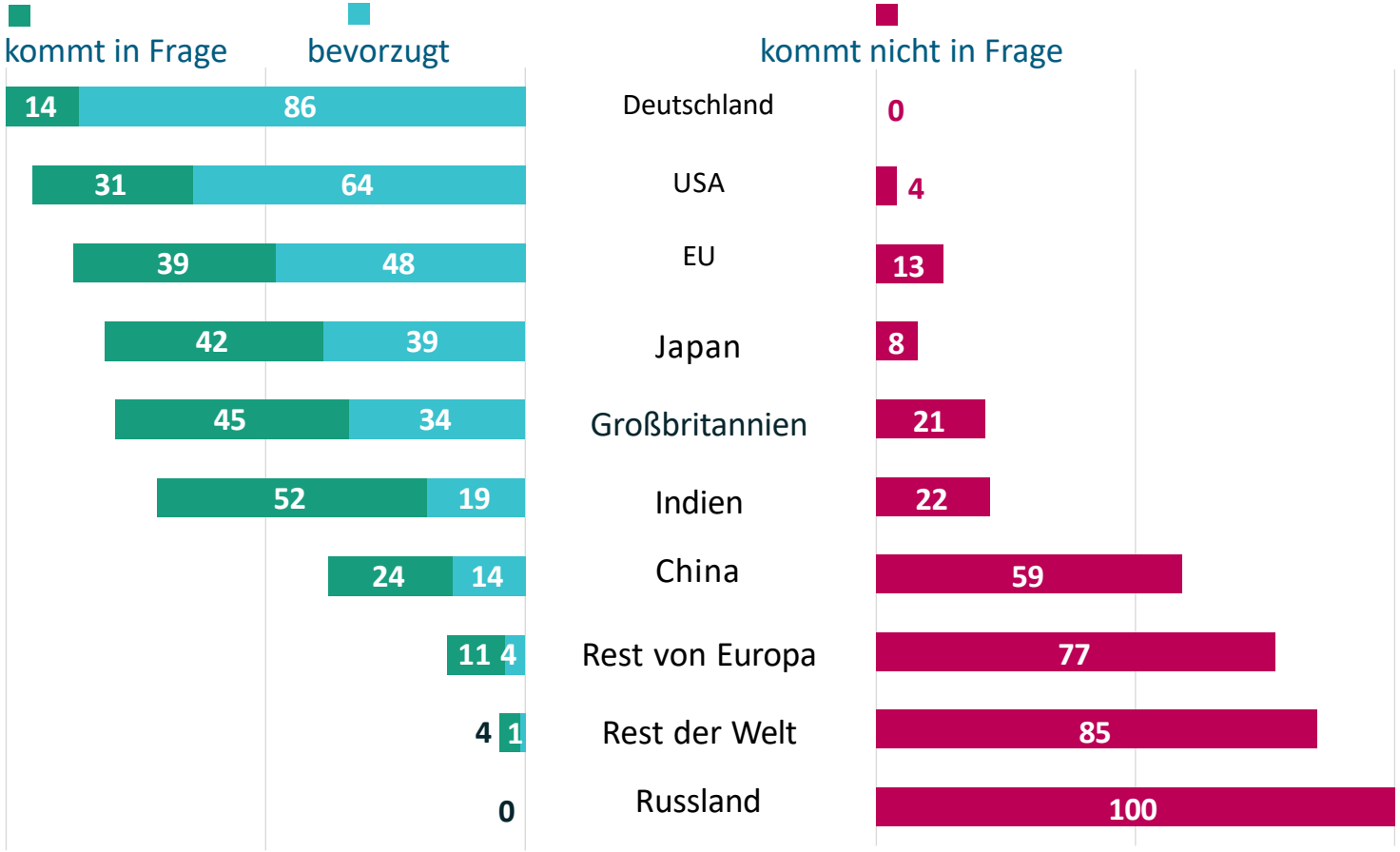


Bayerisches Staatsministerium für Wirtschaft, Landesentwicklung und Energie

Fraunhofer IIS

KI-Anbieter aus Deutschland wären die erste Wahl

Wie würden Sie das Herkunftsland des Anbieters einer generativen KI einordnen?



84%

der Unternehmen, die generative KI nutzen oder es planen, geben an, dass das Herkunftsland des Anbieters „sehr wichtig“ oder „eher wichtig“ ist.

Basis: Unternehmen, für die der Standort des Anbieters einer generativen KI »sehr wichtig« oder »eher wichtig« ist (n=135)

Quelle: Bitkom Research 2024

SOOFI

SOOFI – Europas Reasoning LLM

Referenzprojekt

Bundesministerium für Wirtschaft und Energie

- SOOFI soll ein offenes KI-Sprachmodell mit rund **100 Mrd. Parametern** entwickeln
- Ziel: **europäische KI-Souveränität** — weniger Abhängigkeit von US-Anbietern.
- Gefördert mit ca. **21 Mio. €** durch das Bundesministerium für Wirtschaft und Energie
- **Reasoning-Modelle** — für komplexe Aufgaben in Industrie & Verwaltung

soofi

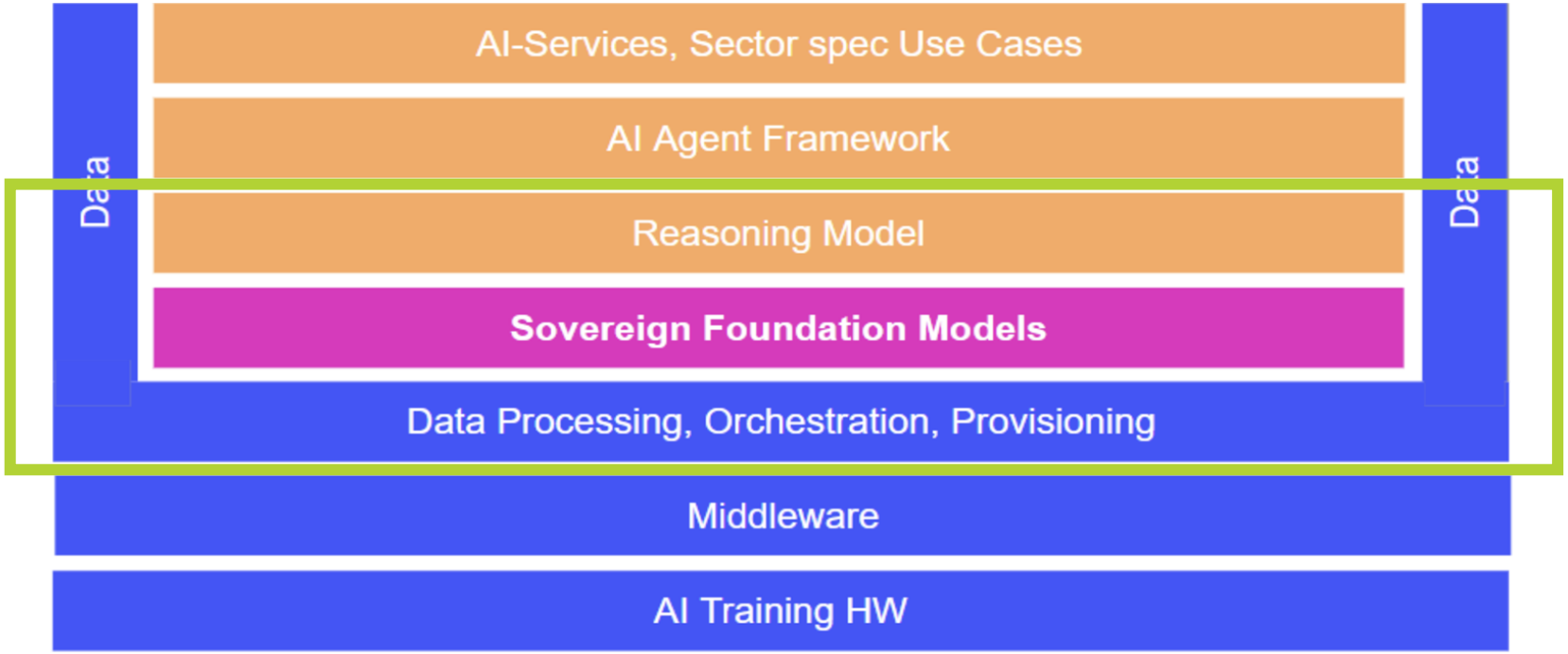
SOOFI – Europas Reasoning LLM

Referenzprojekt

- Training auf der neuen **Industrial AI Cloud der Deutschen Telekom**
- 130 NVIDIA DGX B200 Systemen mit insgesamt über **1.000 GPUs** exklusiv für SOOFI
- Seit **März 2026**

SOOFI - Ziele und Tech Stack

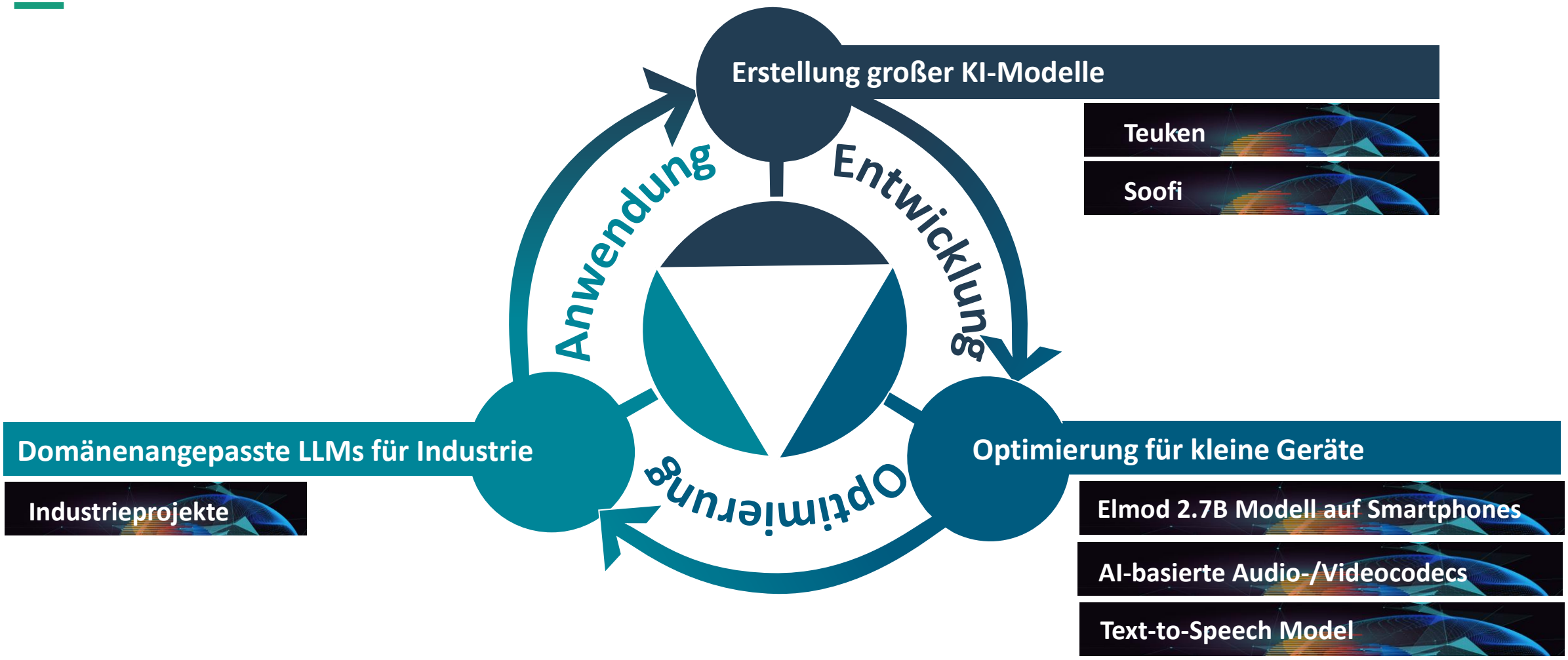
- Soofi entwickelt ein **offenes, souveränes europäisches Ensemble großer Sprach- und Reasoning-Modelle** (LLMs/LRMs) und
- **agentenbasierter Systeme**, das auf europäischen **Cloud- und Edge-Infrastrukturen einsetzbar** ist.
- **Anwendungsfälle** aus realen Nutzungskontexte der **Industrie** werden umgesetzt.



Weiterführung des Projekts als **europäischen KI Moonshot** für die Entwicklung souveräner und wettbewerbsfähiger Foundation Modelle, neuer Modellarchitekturen und agentischer Systeme der nächsten Generation

Unsere Vision und Fokus für Generative KI

Fraunhofer IIS



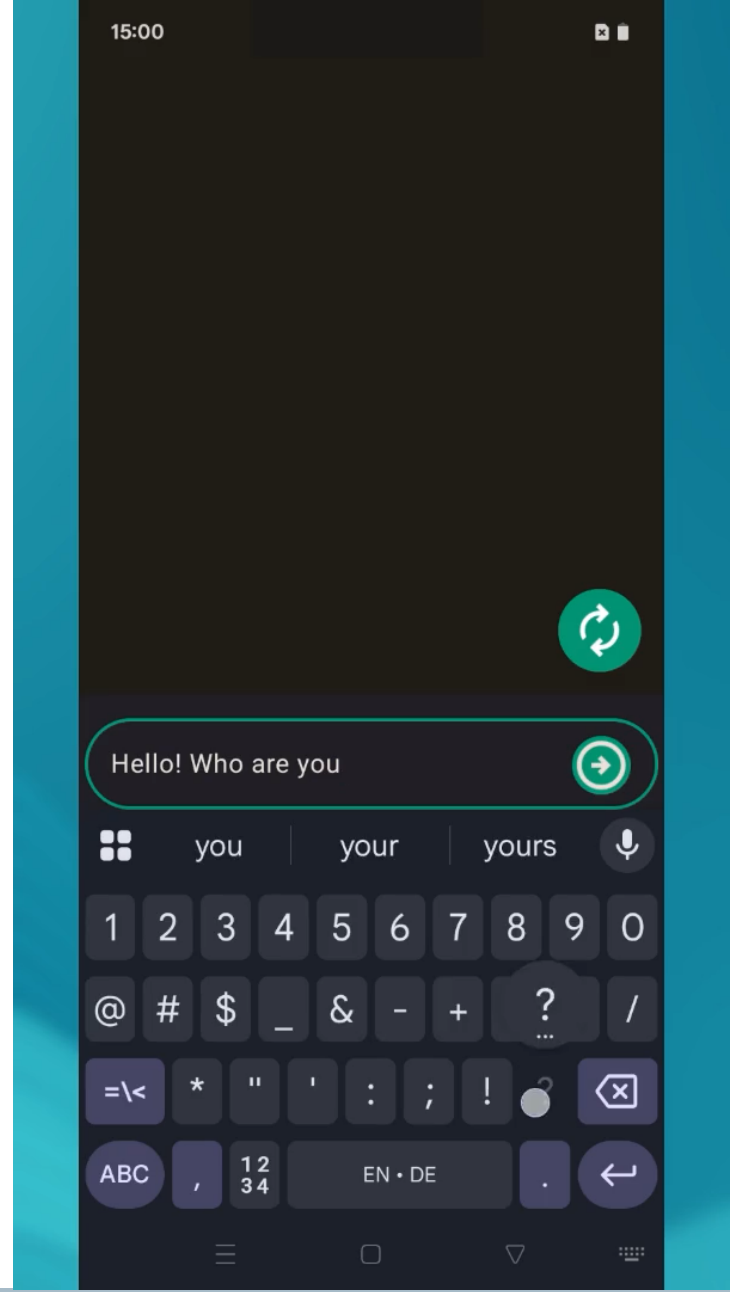
ELMOD 2.7B – GenAI direkt in der Hosentasche

Referenzprojekt

Study LLM, das reibungslos auf einem Smartphone-Chip läuft

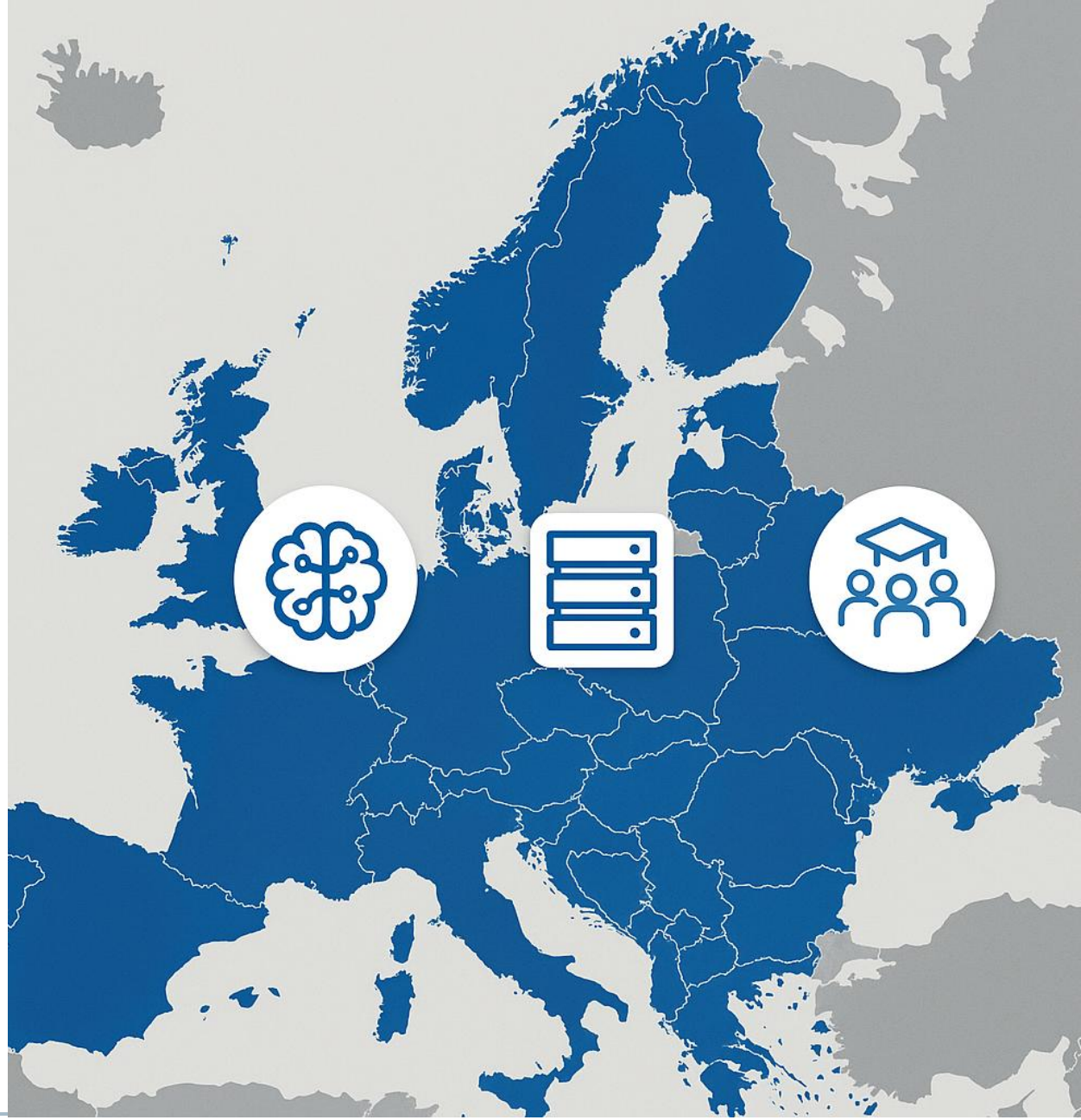
End-to-End-Pipeline, von Grund auf vom Fraunhofer IIS entwickelt

- Kompaktes Modell, geeignet für on-device Deployments
- **2.7B Parameter**, trainiert auf 3.8T Tokens (German, English, and Code)
- **First-of-its-kind deutscher Datensatz**
- Etablierung von Prozessen, Frameworks und Know-How für weitere Basismodelle
  - Erhöhung der Datenqualität
  - Optimierung von Speicherbedarf und Rechengeschwindigkeit
  - …



Wie wir beim Digitalisierungs- und KI-Rennen mitspielen – oder verlieren

- Rechenzentren
- Projekte und Modellentwicklung
- Talente
- Start-up Ökosystem
- Politische Rahmenbedingungen
- Regulatorische Hürden



Vielen Dank für Ihre Aufmerksamkeit!

Kontakt

Jan Plogsties
Abteilungsleiter Generative KI
jan.plogsties@iis.fraunhofer.de

Fraunhofer-Institut für Integrierte Schaltungen IIS
Am Wolfsmantel 33
D-91058 Erlangen

www.iis.fraunhofer.de

Unser Leistungsangebot

Fraunhofer IIS

Vollständige Datenaufbereitung für LLM-Training

- Konzeption einer passenden Datenstrategie
- Erstellung kuratierter, hochwertiger Pretraining-Datensätze (inkl. multilingual)
- Domänenspezifische Datenaufbereitung und Datenoptimierung (branchen- & firmenspezifische Trainingsdaten)
- Erhöhung und Sicherstellung der Datenqualität

Modellentwicklung

- Konzeption und Vorstudie
- Erstellung großer KI-Modell (Training von Basismodellen)
- Expertise mit Techniken für Multilingualität und Multimodalität
- Expertise mit Flexible Adapter Experten Architekturen
- Training (angepasster) Embedding-Modelle

Domänenspezialisierung von Basismodellen

- Spezialisierung von Basismodellen an Branchen- und firmenspezifische Anforderungen (von Fine Tuning bis Continued Pre-Training)

Effizienz & Ressourcenoptimierung

- Effizientes, stabiles und kostengünstiges Training
- Optimierung für kleine Geräte (Speicher, Geschwindigkeit, Quantisierung)

Evaluation & Benchmarking

- Benchmarking und Vergleich diverser Basismodelle
- Entwicklung geeigneter Evaluationsmetriken für Modell- und Applikationsbewertung
- Praxisnahe Evaluation: Tests unter realen Einsatzbedingungen

Produktintegration & Systeme

- Aufsetzen und Optimierung von Applikationsarchitekturen
  - RAG-Systemen
  - Multi-Agent-Systeme
- Entwicklung von Lösungen für lokalen/offline Einsatz (On-prem und Edge-Deployment)
- Integration in produktive Plattformen & Anwendungen