Fraunhofer

Dr. Nicolas Flores-Herr & Team

GOSIM AI | Paris 2025

How to Build Competitive Large Language Models "Made in Europe"?

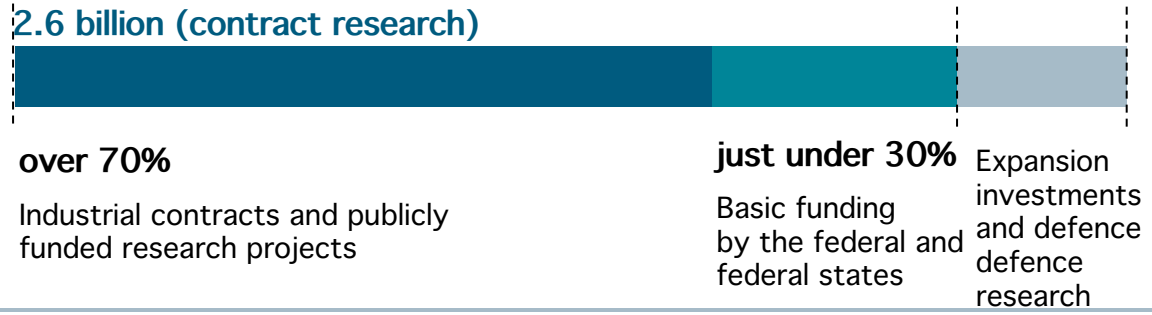
The Fraunhofer-Gesellschaft at a glance

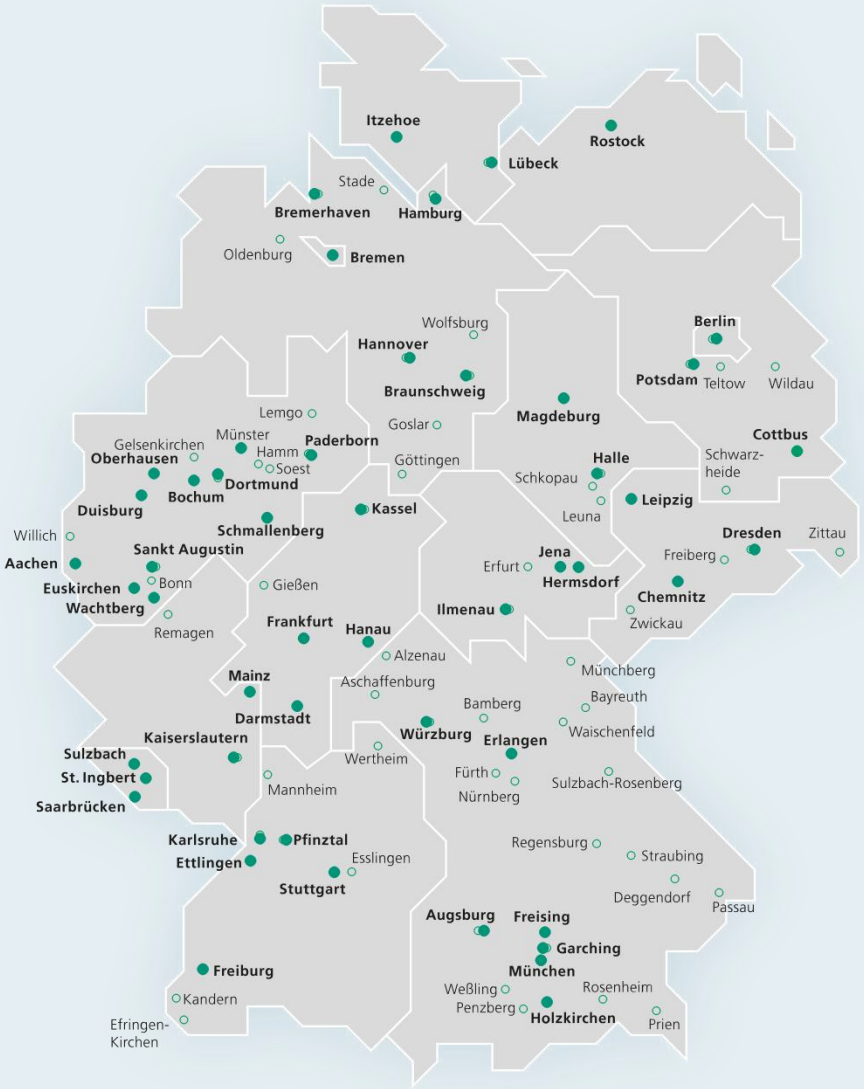
Research For the direct benefit of the economy and the society

30,000+ employees

76 institutes and research facilities

3.0 billion € financial volume





Status: 31.12.2022

Fraunhofer

LLM Ecosystem for Europe
That reliably builds competitive GenAI models

Fraunhofer

GenAI & Digital Sovereignty
Is the "race" ongoing or decided?

US - Main Players: OpenAI, Antrophic, Google, Microsoft, Meta etc.

- **Huge AI Infrastructure:** Azure, GC, AWS, Microsoft
- **Clear statement**: US wants to dominate the AI market in the world - **America first**
- **Stargate** project: 500 Billion$ for AI

China: strong teams and models, like **DeepSeek**, **Qwen**, **Kimi** etc.

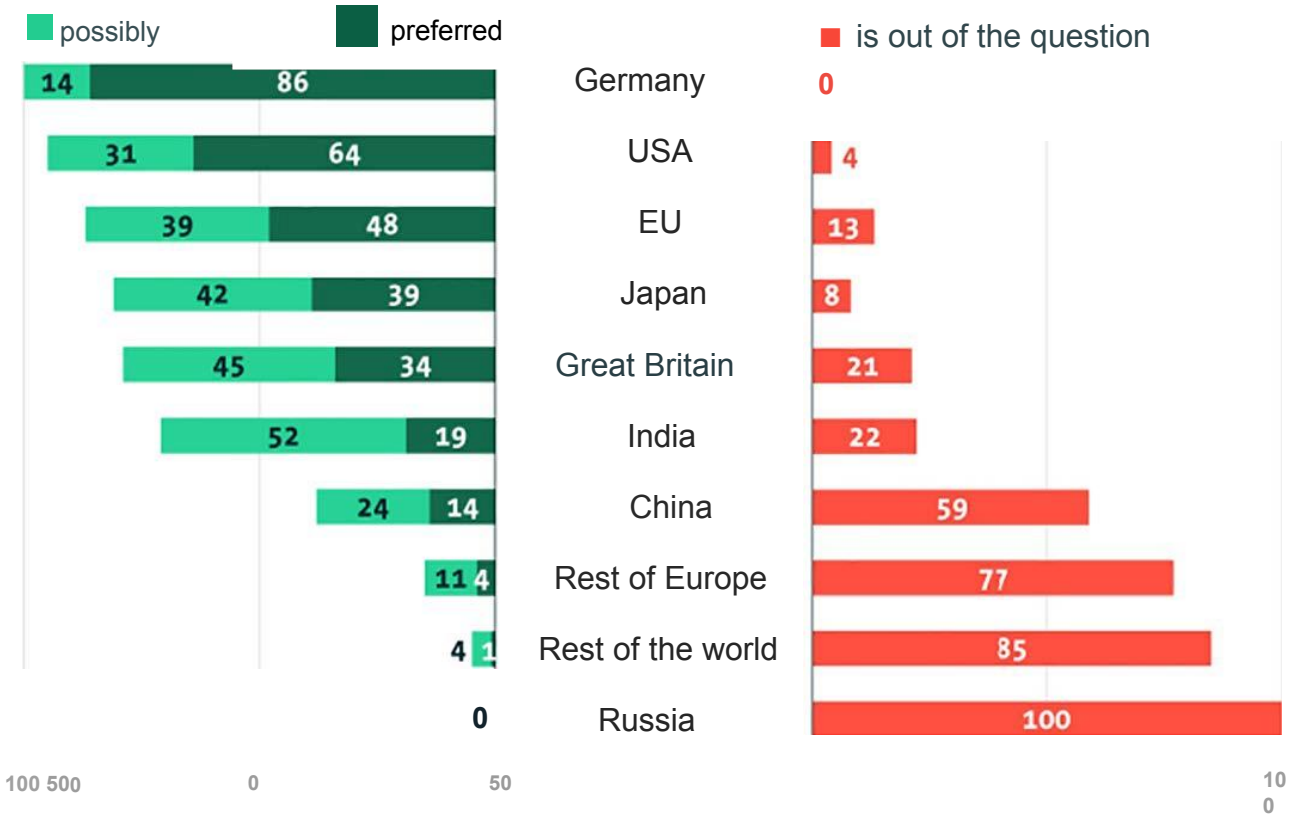
EU: Small start-ups building models like Mistral, iGenius, Aleph Alpha etc. => dozens more needed

- **Strong resources:** EuroHPC, industrial data but not yet used with a great focus on competitive AI models
- **European and National Initiatives**: great potential but speed, governance and administrative overhead pose challenge
- **European AI Act:** regulation, rules, laws causing more red tape
- **AI Summit in Paris / AI Continent Plan:** Announcements like Gigafactories give hope but may take time

AI providers from Germany would be the first choice

Bitkom survey of German industry from October 2024

How would you categorise the country of origin of the provider of generative AI?



84%

of companies that use or plan to use generative AI, state that the provider's country of origin is "very important" or "rather important".

Basis: Companies for which the location of the provider of a generative AI is "very important" or 11somewhat important (n=13S) 1missing values to 100%: .don't know/k. A... | QtJelle: 8itkom A:ese;,rch 2024

Economic potential for Germany

36%

of German companies **used AI in 2023**, compared to 28% in 2022

61%

of German companies believe that **AI** will **significantly change their industry** in **the** next 5 years

€ 668 billion

for the German economy until 2030

92%

of companies that use AI reported **higher efficiency**

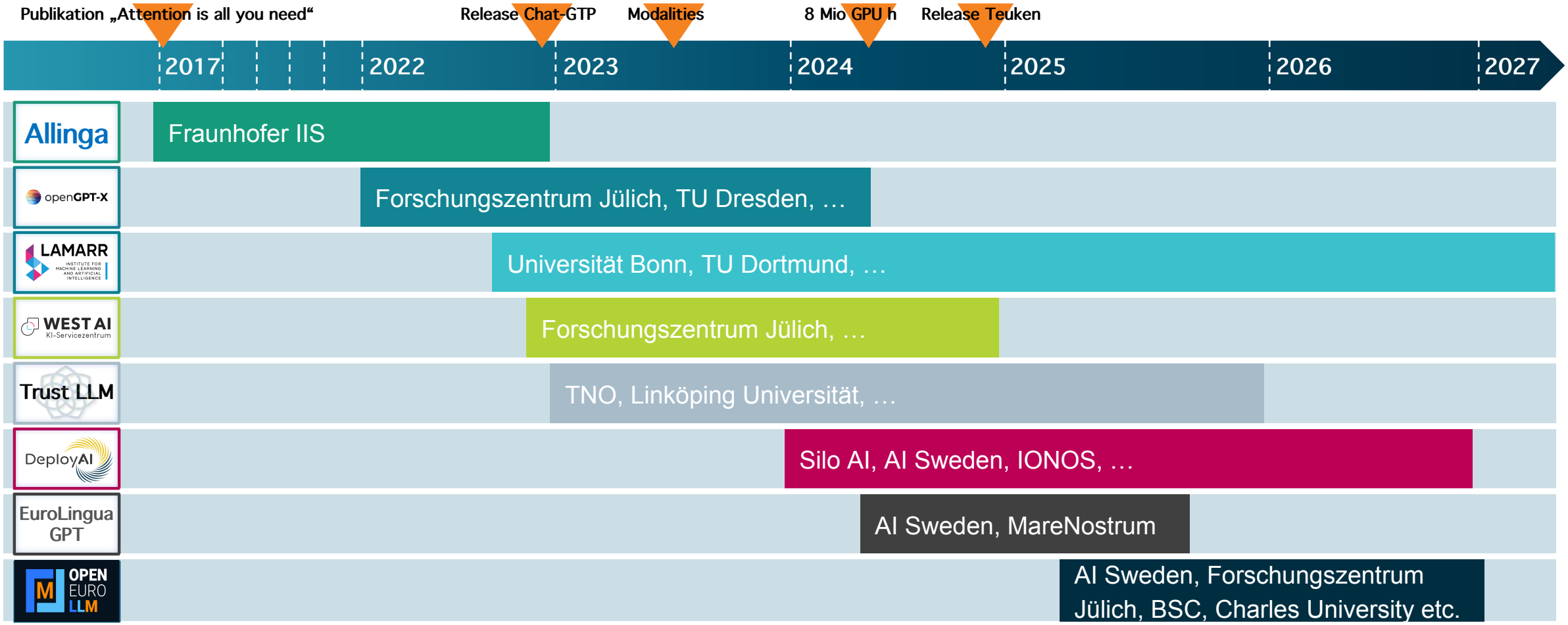
72%

of companies using AI reported **more effective processes**

Source: Germany Country Report 'Unlocking Germany's AI Ambitions in the Digital Decade', AWS and Strand Partners, February 2024 (https://www.unlockingeuropesaipotential.com/germany)

Cornerstone of European digital sovereignity

Research projects and collaborations



Fraunhofer

Fraunhofer IAIS in Germany & Europe
Sovereign multilingual language models

- **OpenGPT-X** started in 2022 and was the **largest German consortium** for the development of LLMs
  - Teuken-7B 0.4 with 24 European languages
  - Open source of model weights under Apache 2.0
  - Focus on digital sovereignty of industry and public sector

- **European LLM flagship projects**
  - European AI platform
  - 35 languages
  - Open source

Development of a high-performance LLM family from Europe according to European values

Fraunhofer

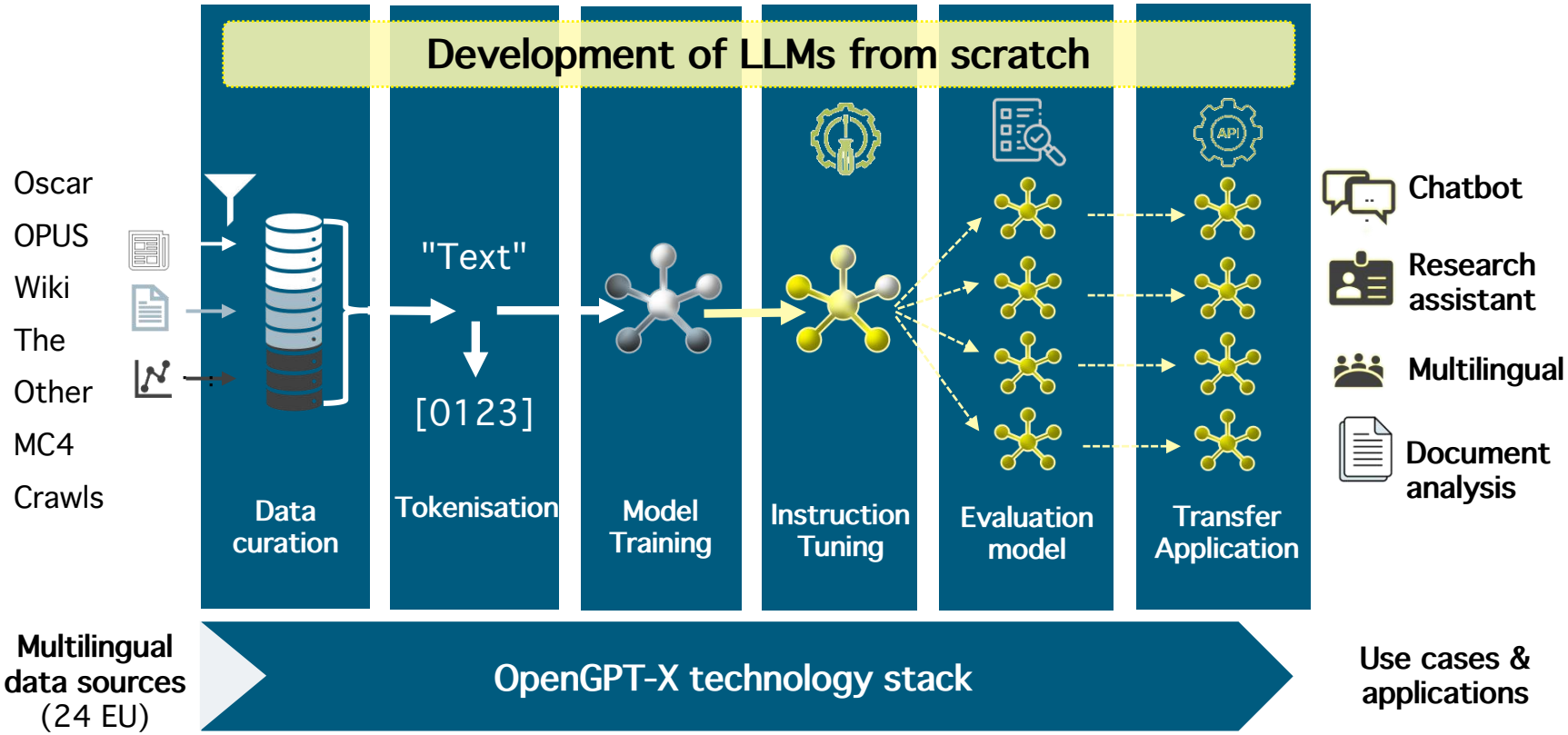

Multilingual and Open **Source**
Teuken-7B
ALEPH ALPHA
ControlExpert
dfki ai
Fraunhofer IIS
KI BUNDESVERBAND
Fraunhofer IAIS
IONOS
JÜLICH Forschungszentrum
WDR
ZIH
Zentrum für Informationsdienste und Hochleistungsrechnen

Fraunhofer technology overview

Technological basis for the development of competitive European language models

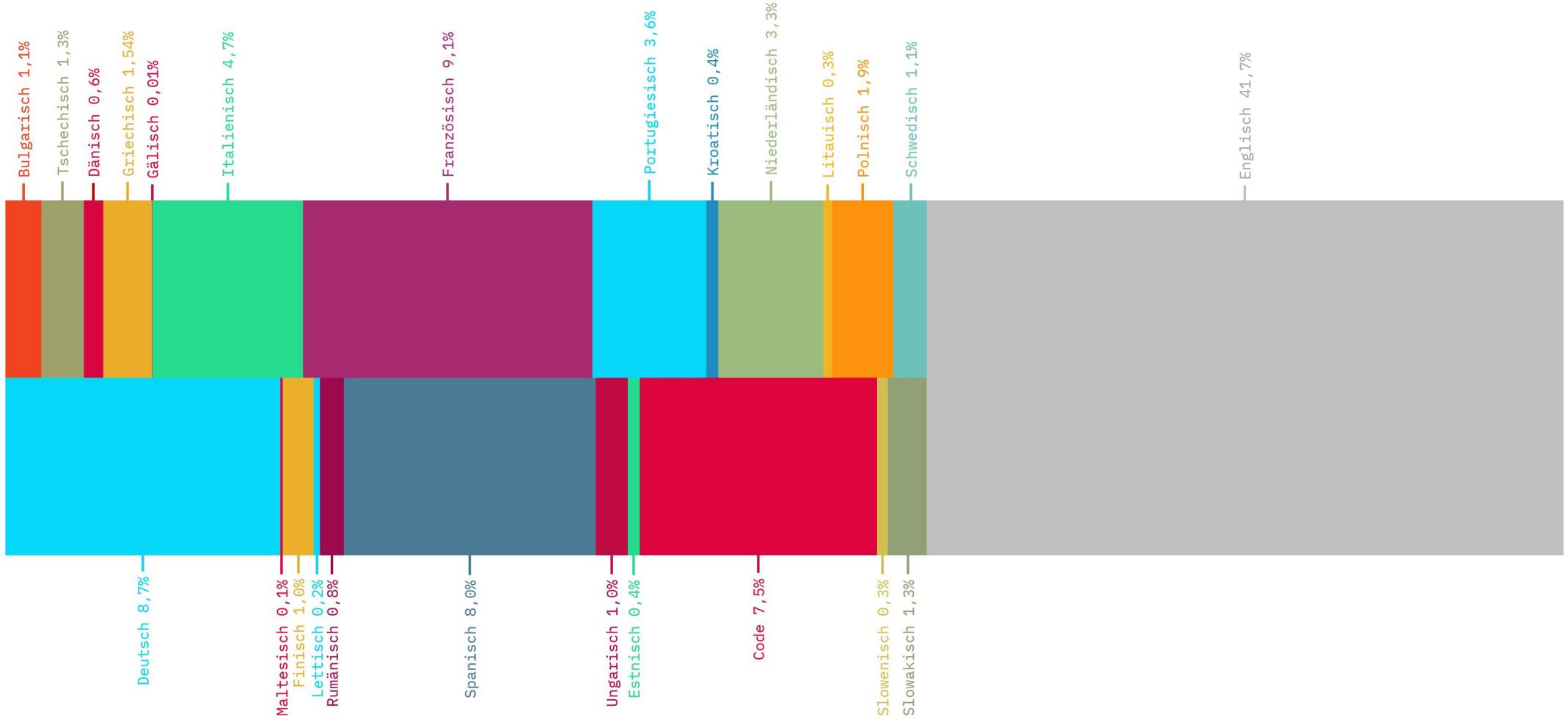
Technology stack

- **Large data collection created**
  - **Multilingual**
  - Not just Common Crawl
- **Data processing pipeline**
  - Multilingual
  - Productive filtering methods
- **Tokenisation**
  - Multilingual
- **Scaling in model training**
  - Modalities
- **Instruction Tuning**
- **Evaluation**
  - European Leaderboard



Fraunhofer IAIS

Teuken-7B
Language Distribution



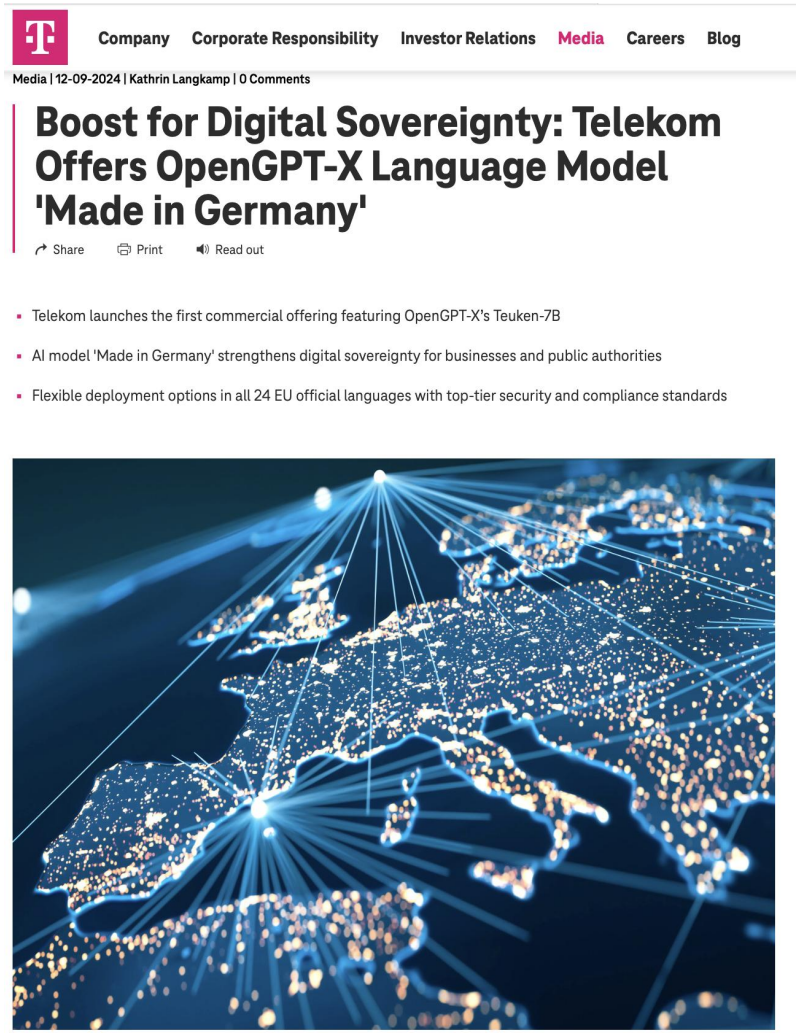
Teuken-7B - 4T in a nutshell (November 24)

Features

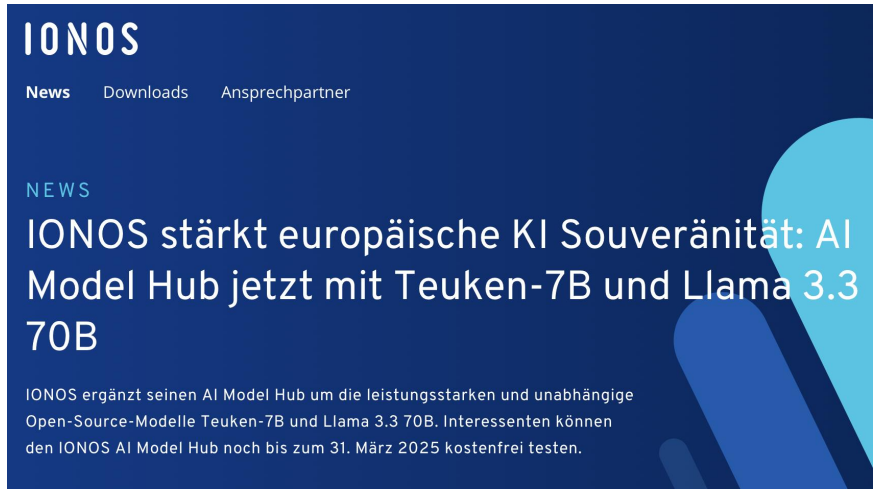
- 7 billion parameters
- Multilingual: supports 24 European languages with multilingual tokenisation
- Open Source Weights (Apache 2.0)
- Trained on 4 trillion tokens using the JUWELS Booster with 512 NVIDIA A100 GPUs

Hardware requirements

- Hardware: NVIDIA A100 PCIe 80GB
- GPU Memory: 37GB
- Deployment Method: vLLM



Huggingface Download



Fraunhofer IAIS

Teuken-7B Media Response

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

> 60,000 downloads
of Teuken models at Hugging Face

> 50 Press articles
Major german news outlets

> 33 million outreach
(dpa, heise, FAZ, Handelsblatt etc.)

Discord
Growing community

Status at the end of 2024

Fraunhofer IAIS

Teuken 7 B 6T benchmarks: comparison with Open Source LLMs

6 trillion tokens, 800,000 GPU-hours

Model	Token	**Mean value**	**ARC**	**GSM8K**	**HellaSwag**	**MMLU**	**TruthfulQA**
Instruction Tuned Versions	*Number of tokens in trillions*		Scientific multiple choice questions	Maths exercises	Generally comprehensible questions about everyday events	Multiple choice knowledge questions	Factual correctness
Meta-Llama-3.1-8B	*15*	**0,56**	0,56	0,56	0,58	0,58	0,53
Teuken 7B	*6*	**0,51**	0,59	0,29	0,66	0,45	0,57
Mistral-7B v0.3	*8*	**0,48**	0,53	0,30	0,54	0,49	0,55
Pharia-1-LLM-7B	*7,7*	**0,35**	0,40	0,06	0,44	0,37	0,47

Effects of training an LLM on a high percentage of german text

Nach wem wurde die Pfarrer Eras Straße in Utzenhofen benannt?

61 character(s)

Open Source | Llama-3.3-70B-Instruct Chat | License: Llama3.3

Die Pfarrer-Eras-Straße in Utzenhofen wurde nach dem katholischen Pfarrer Eras benannt. Leider konnte ich keine spezifischen Informationen über den Pfarrer Eras finden, da es möglicherweise nicht viele Quellen über ihn gibt oder er nicht sehr bekannt ist.

Es ist jedoch wahrscheinlich, dass Pfarrer Eras eine wichtige Rolle in der Geschichte von Utzenhofen gespielt hat, möglicherweise als Seelsorger oder als Initiator von sozialen oder kulturellen Projekten in der Gemeinde. Die Benennung einer Straße nach ihm ist wahrscheinlich ein Zeichen der Wertschätzung und des Respekts, den die Gemeinde ihm entgegenbringt.

Wenn du mehr über Pfarrer Eras erfahren möchtest, könntest du dich an die lokale Gemeindeverwaltung oder das Stadtarchiv von Utzenhofen wenden, um zu sehen, ob sie weitere Informationen über ihn haben. Es ist auch möglich, dass es in der Gemeinde oder in lokalen Geschichtsbüchern Hinweise auf sein Leben und Wirken gibt.

940 character(s)

OpenGPT-X | OpenGPT-X-Teuken-7B-instruct-research-v0.4 Chat | License: Other

Die Pfarrer Eras Straße in Utzenhofen wurde nach dem ehemaligen Pfarrer der örtlichen Kirche, Johann Eras, benannt. Er war von 1958 bis 1976 Pfarrer in Utzenhofen und hat sich besonders für die Jugendarbeit und die Förderung der Kultur in der Gemeinde engagiert. Nach seinem Tod wurde die Straße, die direkt an der Kirche vorbeiführt, zu seinen Ehren benannt.

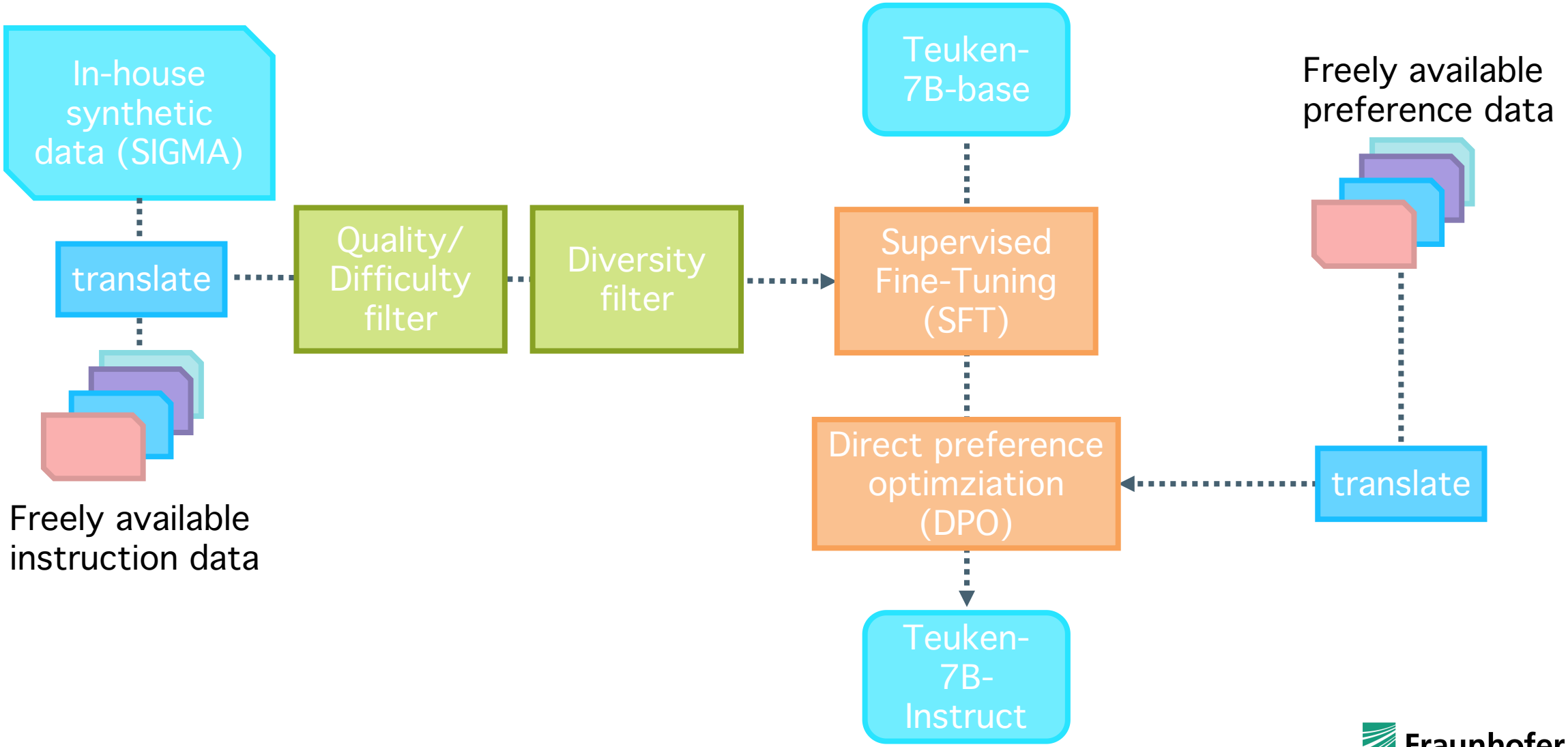
360 character(s)

Open Source | Mistral-Small-3.1-24B-Instruct-2503 Chat

Die Pfarrer Eras Straße in Utzenhofen, einem Ortsteil von Markt Nordheim im Landkreis Neustadt an der Aisch-Bad Windsheim in Bayern, ist nach Pfarrer Johann Michael Eras benannt. Johann Michael Eras war ein bedeutender Geistlicher und Pädagoge, der sich um die Bildung und das Wohl der Gemeinde verdient gemacht hat. Straßenbenennungen nach lokalen Persönlichkeiten sind oft eine Anerkennung für deren Verdienste und ihren Beitrag zur Gemeinschaft.

448 character(s)

Teuken – Instruction tuning

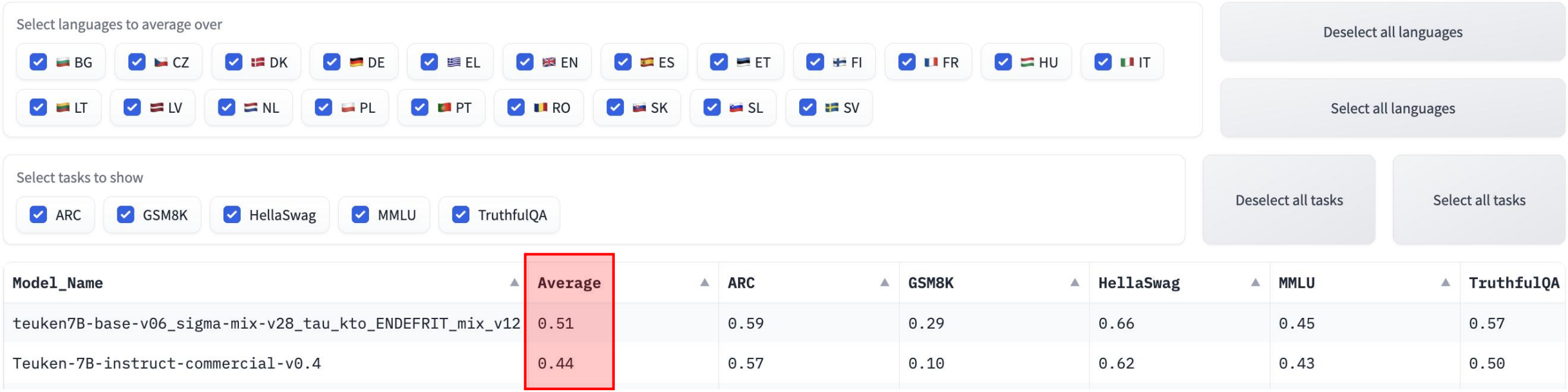


Fraunhofer IAIS

Teuken 7 B benchmarks: effect of continued pretraining & finetuning

6 trillion tokens, 800,000 GPU-hours

European Leaderboard OpenGPT-X



Model_Name	Average	ARC	GSM8K	HellaSwag	MMLU	TruthfulQA
teuken7B-base-v06_sigma-mix-v28_tau_kto_ENDEFRIT_mix_v12	0.51	0.59	0.29	0.66	0.45	0.57
Teuken-7B-instruct-commercial-v0.4	0.44	0.57	0.10	0.62	0.43	0.50

By improving instruction tuning datasets and continued pretraining on 2 T high-quality training data Teuken-7B 6T improved by 7 percent in relation to Teuken-7B 4T released in November

Fraunhofer IAIS

OpenGPT-X
Lessons learned

Lighthouse project of the BMWK
Budget: 20 million euros - Duration: 2022 to 2024
Headed by Fraunhofer IAIS

- **Largest German consortium** for the development of large AI language models
- **Compute / AI Infrastructure:** one of the main challenges
- **Data processing and learning schedule**
- **Legal aspects**
- **Governance: research projects versus companies**

Successful and trustworthy collaborations are key to build

competetive models

EuroLingua-GPT

"Extreme Scale Access" EuroHPC project

8.8 million H100 GPU hours
06/2024 - 06/2025

- Partners are SwedenAI and Fraunhofer
- Compute on Mare Nostrum 5 in Barcelona
- Focus on very large multilingual models
- Based on OpenGPT/X, TrustLLM, DeployAI
- Open source model weights and more if possible

Teuken-2 Downstream Performance – Initial Results

Model	Average	ARC	HellaSwag	MMLU	TruthfulQA
Teuken 7B Base v04 (4T)	0.50	0.55	0.61	0.35	0.47
Teuken-2 8B (2T)	0.53	0.56	0.63	0.44	0.47

- Teuken is trained on 24 languages
- Teuken-2 is trained on 37 languages (using Modalities)

Scientific Contributions for Each Phase of Model Development

Datenqualität – Paper:
“Enhancing Data Quality: Can Language Models Be Trusted as Multilingual Judges?“

Modell – Paper:
“Teuken-7B-Base & Teuken-7B-Instruct: Towards European LLMs“

Multilinguale Evaluation – Paper:
„Towards Cross-Lingual LLM Evaluation for European Languages“

Data curation → Token ization → Modell training → Instruction-Tuning → Model Evaluation → Transfer of Models

Tokenisierung – Paper:
“Tokenizer Choice For LLM Training: Negligible or Crucial? “
Akzeptiert bei NAACL Findings 2024 (A Konferenz)

Instruction Tuning - Paper:
“Investigating Multilingual Instruction-Tuning: Do Polyglot Models Demand for Multilingual Instructions?“
Akzeptiert bei EMNLP 2024 (A* Konferenz)

Bias Evaluation - Paper:
"Do Multilingual Large Language Models Mitigate Stereotype Bias?"
Akzeptiert bei ACL-Workshop 2024

Scientific Publications

Technological development
Roadmap for 2025/2026

EuroHPC (MareNostrum)
EuroLingua

Jupiter of the FZJ:
26,000 GPUs + AI Factories

Teuken-7B
Publication in November 2024

Teuken-7B
June 2024

- 24 EU languages
- 7 billion parameters
- 4 trillion tokens
- On a par with comparable models

Teuken-7B
December 2024

- 24 EU languages
- 7 billion parameters
- Training with 6 trillion tokens
- Improved filtering of the training material
- Trained on JUWELS

Teuken-2 8/15/30B
September 2025

- EuroLingua
- 37 Languages, dialects, codes
- 8/15/30 billion parameters
- Training data filtered by LLM (iterative)
- Training with up to 10 trillion tokens

Teuken-2 100B
December 2025

- EuroLingua
- 37 Languages, dialects, codes
- >70 billion parameters
- Training with up to 10 trillion tokens

Teuken Enterprise
End 2025 / 2026

- EuroLingua
- Multimodal
- Improved reasoning skills
- Execution of tools

Summary
How to Build Competitive Large Language Models "Made in Europe"?

Digital Sovereignty:

- Europe is under strong pressure to compete in the AI market, but the race for GenAI is still open
- What is the position of the German/Europe industry? Germany and Europea has to wake up and **provide more resources**
- **Competetive European models** and a **strong startup scence** are key to make sure transfer into real-world applications is achieved
- This is only possible with **strong and trustworthy** cooperations

Fraunhofer perspective:

- Teuken is only the beginning with follow-up versions will come soon
- Research team will significantly expand

Our outlook:

- Building European AI ecosystems (AI Factories, DeployAI, LEAM, Fraunhofer Eco-System)
- Fostering **scientific collaborations**
- Fostering **collaborations** to improve **open source distribution** and **(vertical) application of models**

Research Areas
WE‘RE HIRING ☺

Contact

-

Dr. Nicolas Flores-Herr
Team Manager Foundation Models
Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS
LinkedIn: https://www.linkedin.com/in/floresherr/

Dr. Mehdi Ali
Innovation Group Leader on Foundation Model Research
Email: mehdi.ali@iais.fraunhofer.de
www.lamarr-institute.org

Thank you!

Scientific Publications

Teuken

Discord

Modalities

European Leaderboard

Example

SFT data

Training the model to follow instructions

Given the current temperature (22.2°C) and the desired temperature (21.1°C), calculate the amount of heat energy needed to be removed to reach the desired temperature, in joules.

First, we need to know the mass of the object or the volume/size of the room where [...]

Preference data

Training the model to be helpful and harmless

Give me a list of 5 nightshade plants, **without additional explanations.**

1. Tomato (Solanum lycopersicum)
2. Potato (Solanum tuberosum)
3. Paprika (Capsicum annuum)
4. Aubergine (Solanum melongena)
5. Belladonna (Atropa belladonna)

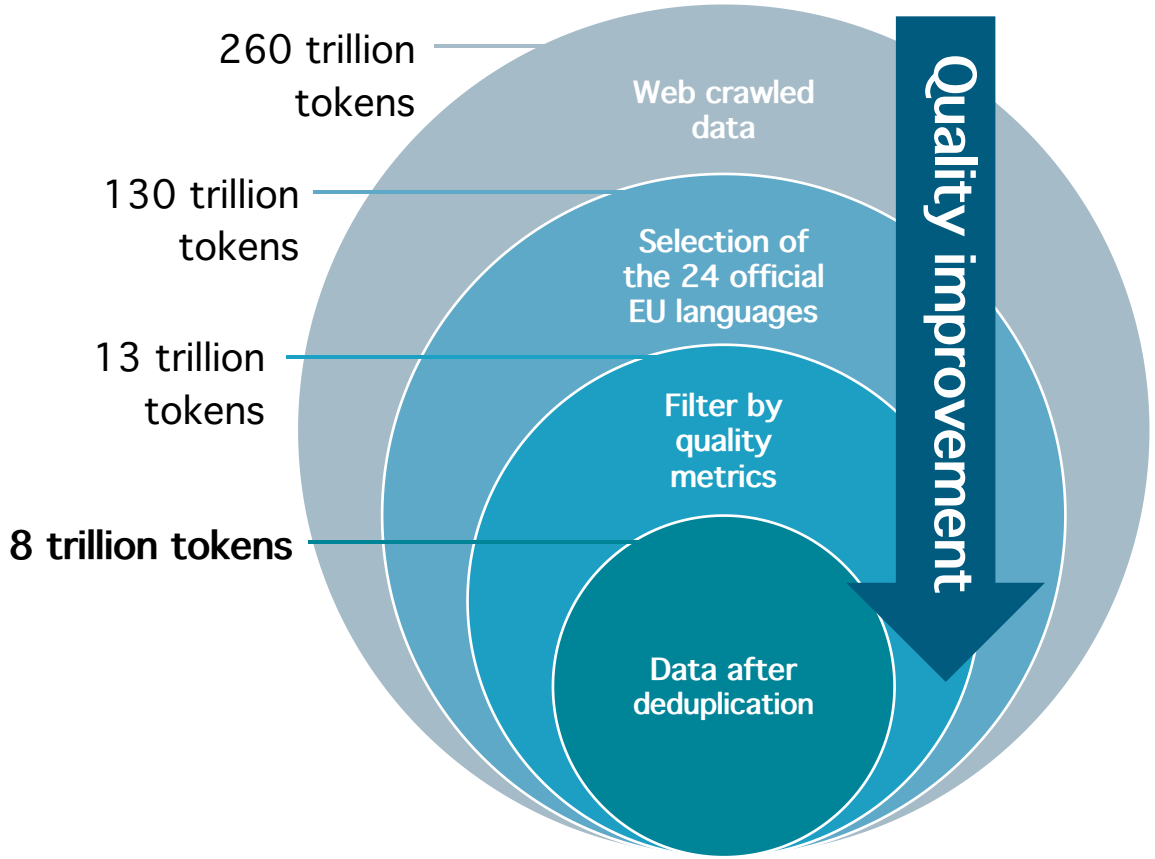
Here are five examples of nightshade plants:

1.Tomato (Solanum lycopersicum) - One of the best-known cultivated plants, rich in vitamins [...]

Multilingual data set from 24 official EU languages

Preparation of enormous data sources to improve quality

Scalable data pipeline for efficient processing



Use of curated data:

- 75 different data sources
- 24 languages, strong German component
- Licences all clarified by BHO/Licences department

Domain	Percentage [%]
Source Code	40.46
Law and Administration	20.40
Web	13.69
Medical	8.94
Math	8.92
Forum	4.05
Books	2.08
News	1.05
Knowledge Base	0.23
Culture	0.08
Recreation	0.08

Pipeline for the development of base models
From raw data to deployed models