



Making available a high performing open-source European foundation model for fine-tuning Info Session



Cecile Huet - Head of Unit
Miguel Rubio – Policy Officer
Robotics & AI Innovation and Excellence
European Commission DG CONNECT

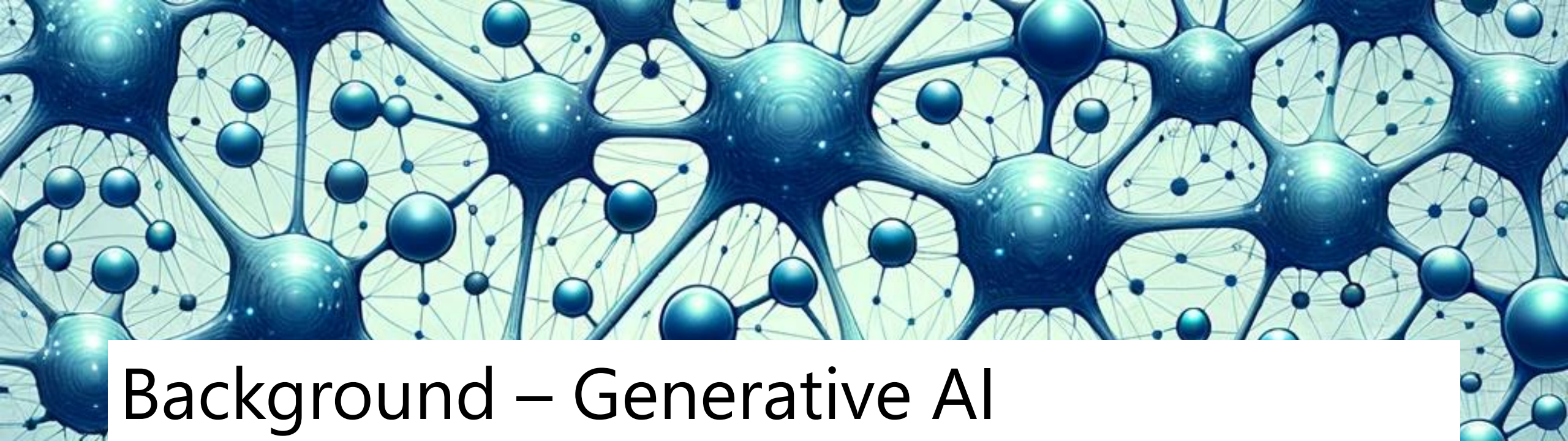
#DigitalEuropeProgramme

March 2024



Outline

- Background on Generative AI
- Other European Commission Initiatives
- Digital Europe call Call DIGITAL-2024-AI-06-FINETUNE: Making available a high performing open-source European foundation model for fine-tuning



Background – Generative AI

- **Generative AI models** (such as large language models) are a new wave of AI models adaptable to various domains and tasks.
- These models have immense **potential** to revolutionise multiple sectors.
- Mastery of this technology is of **strategic importance for Europe** to reduce dependency on non-European companies and ensure **sovereignty**.
- Despite their advantages, generative AI models, have **capabilities and risks** that are still being uncovered.

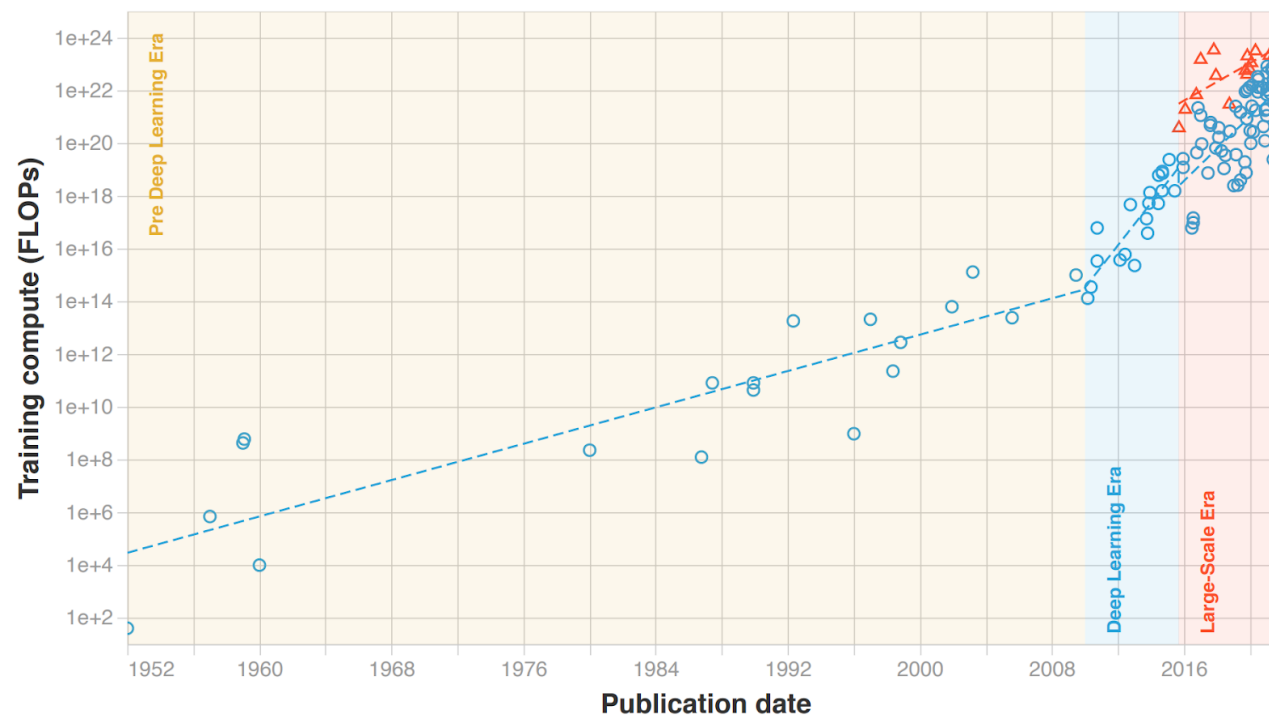


New Technological wave in AI

- **GPT-3 (June 2020)**
 - First widely-known general-purpose model.
 - Emergence of new and unexpected capabilities with increased size

Training compute (FLOPs) of milestone Machine Learning systems over time

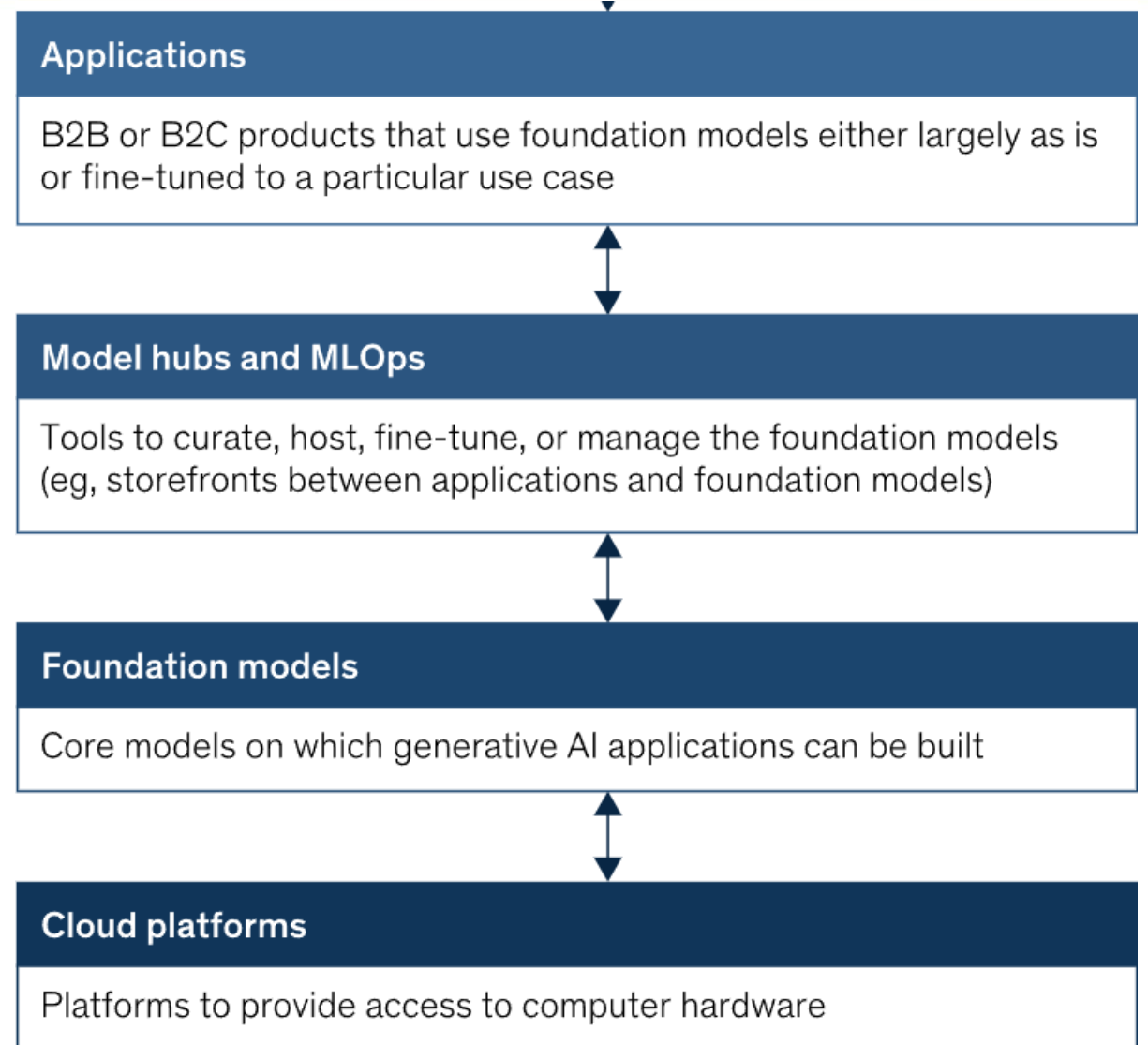
n = 121





Generative AI Value Chain

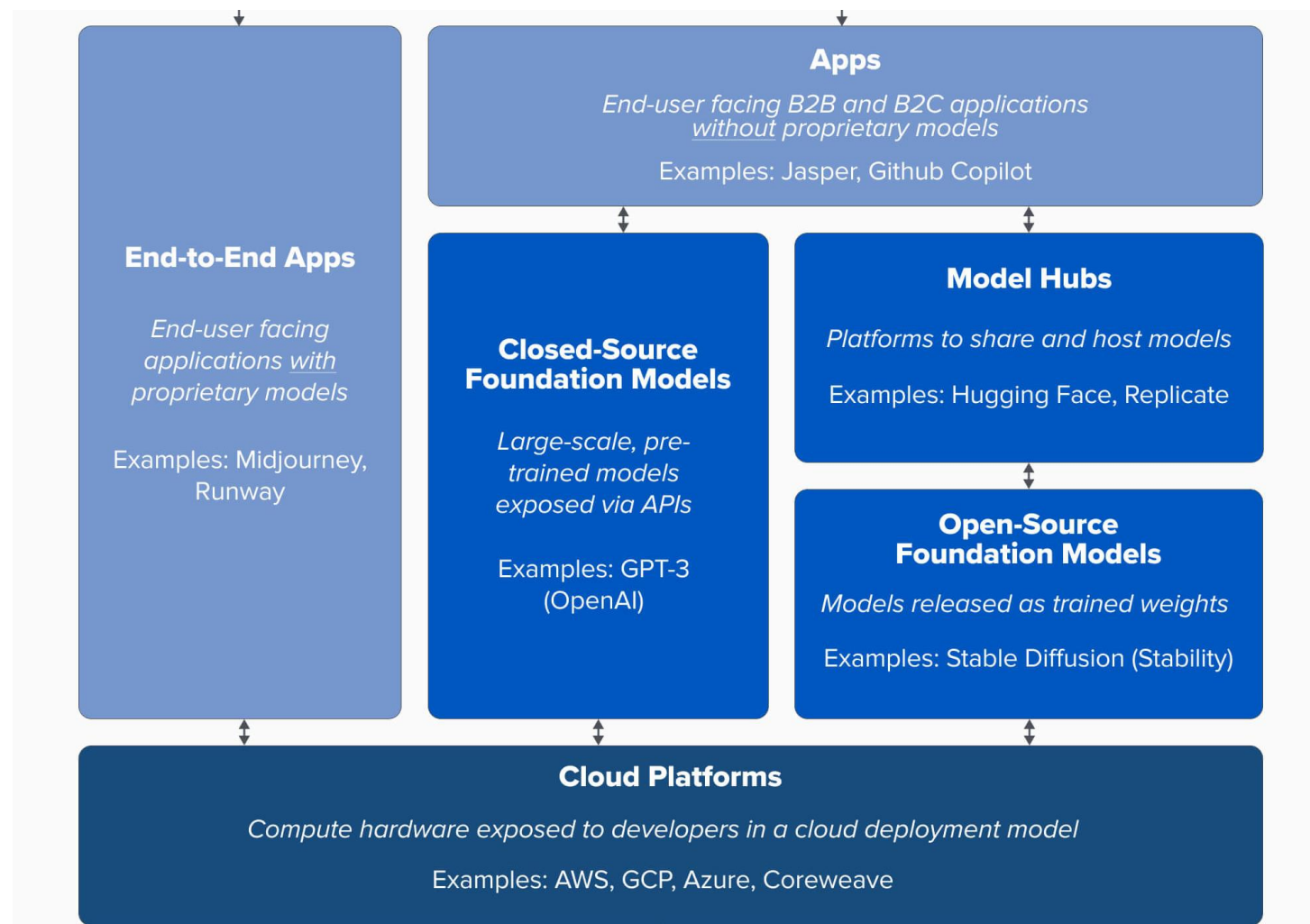
- Applications
- Model Hubs and MLOps
- **Foundation Models**
- Cloud Platforms





Generative AI Business Models

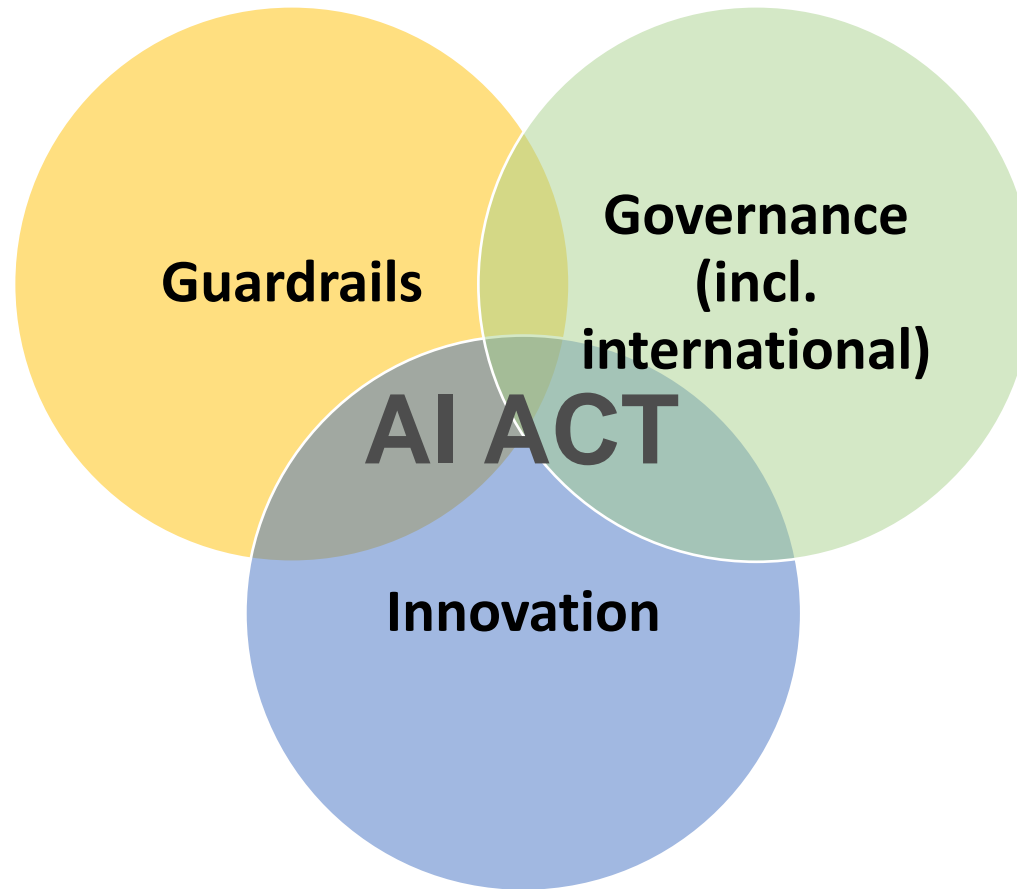
- End-to-End Apps
- Closed-Source Foundation Models
- Open-Source Foundation Models + Model Hubs





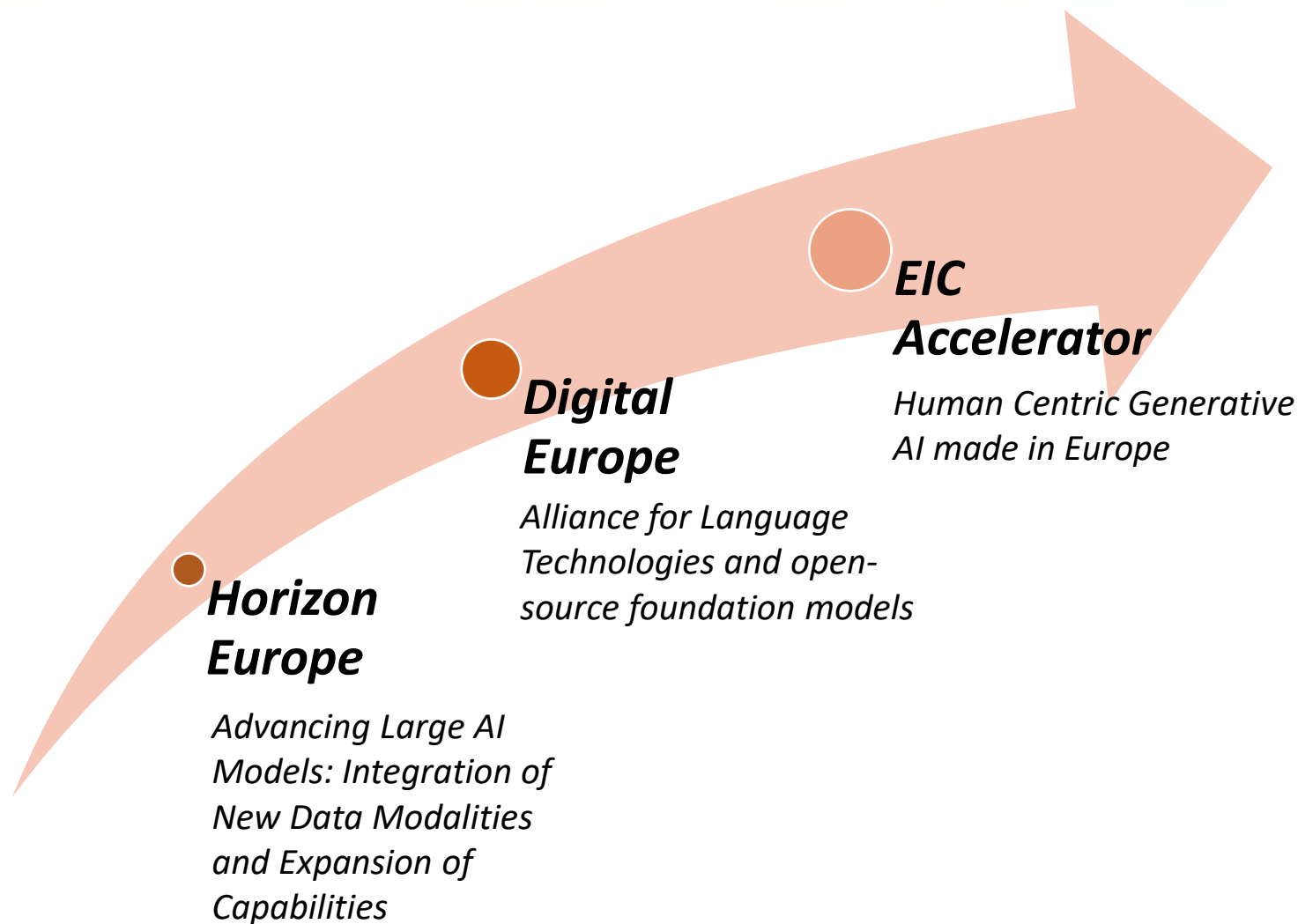
Other European Commission Initiatives

Global AI Framework





Innovation: From the lab to the market





Horizon Europe

Advancing Large AI Models: Integration of New Data Modalities and Expansion of Capabilities (Draft)

Supports the development of foundation AI models:

- Innovative Data Modalities and State-of-the-Art
- Multimodal Models

Explainable and Robust AI (Draft)

- Supports the development of AI systems that are more robust, transparent and explainable.





European Innovation Council

Accelerator Challenge: Human Centric Generative AI made in Europe - €50 million

Support the development of foundation language and multimodal 'frontier' models that reach performances at least equivalent to the most powerful state of the art models.

Targeted applicants

SMEs developing models themselves, and SMEs providing innovative infrastructure, development tools, and critical support.

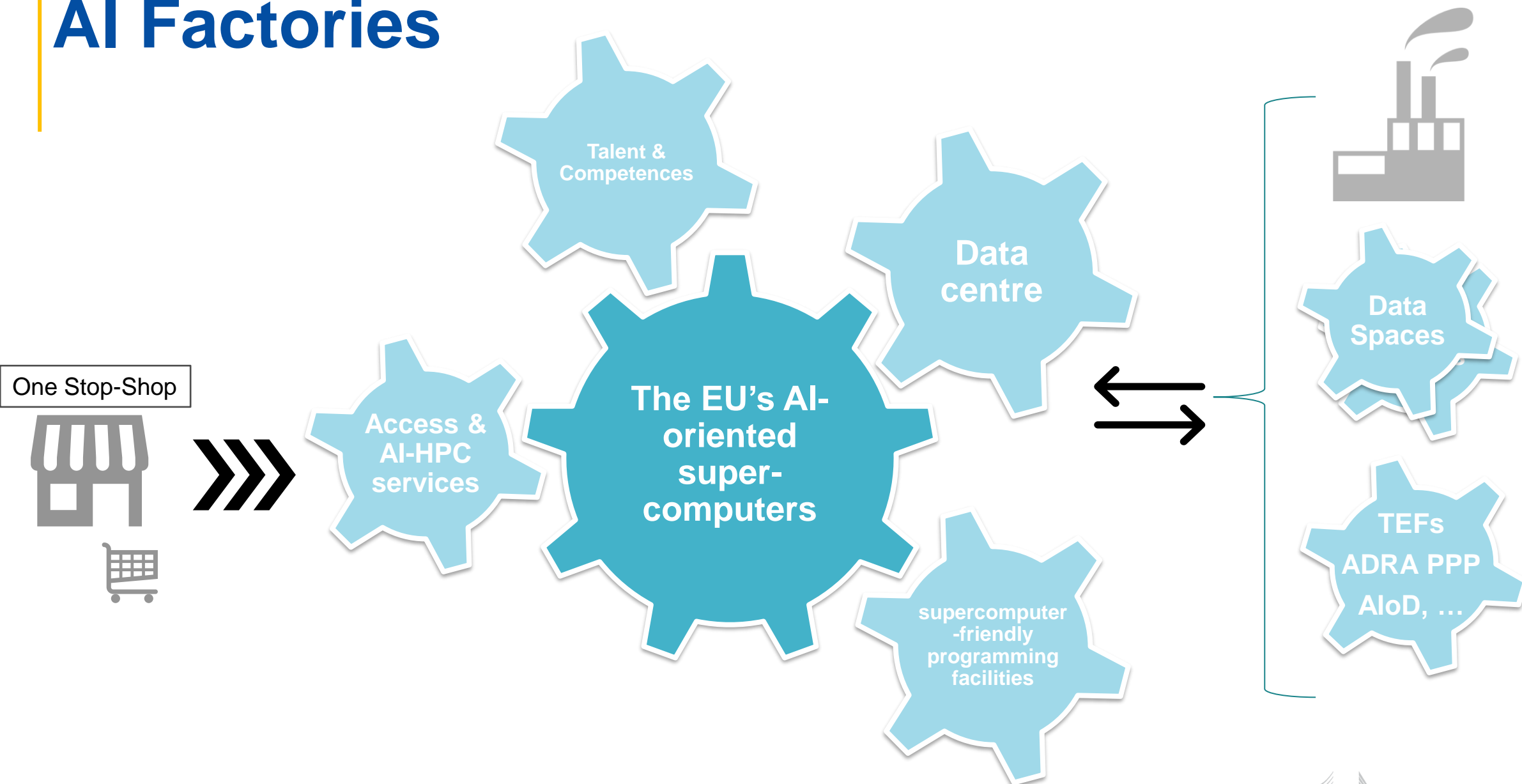


Communication on boosting startups and innovation in trustworthy artificial intelligence

Two Main Objectives:

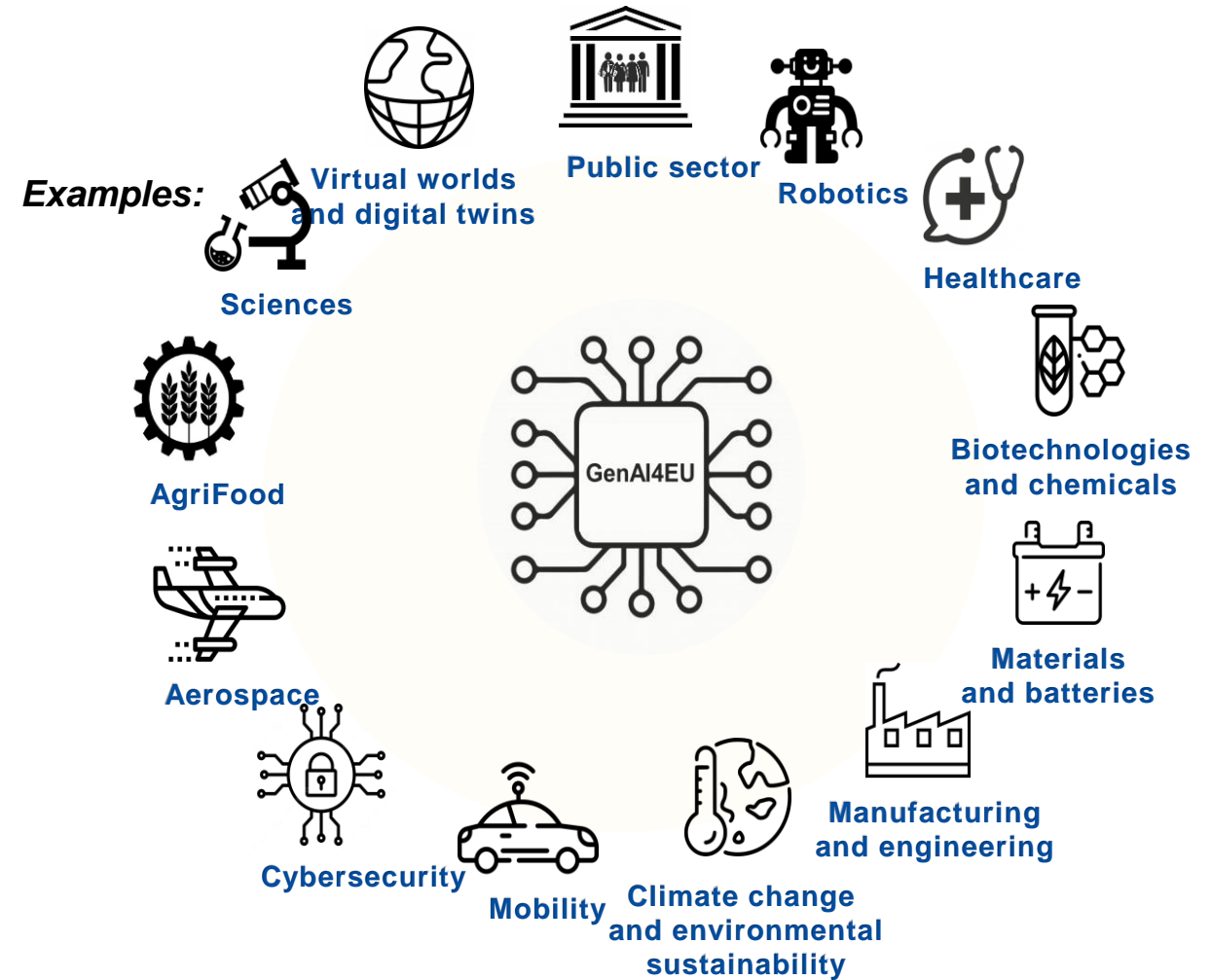
- **AI Factories**
 - Making available HPC computing capacity to facilitate the development of GenAI models
- **GenAI4EU**
 - Stimulating the development in strategic sectors of novel and innovative applications based on GenAI models and facilitating their uptake.

AI Factories



GenAI4EU Initiative

- **GenAI4EU** initiative to stimulate the widespread uptake of generative AI across the Union's *fourteen strategic industrial ecosystems*.
- Startups and innovators can work closely with industrial users, attract investments in the Union and have access to the key ingredients of AI - *data, computing, algorithms and talent*.





DIGITAL-2024-AI-06-FINETUNE:

Making available a high performing open-source European foundation model for fine-tuning



DIGITAL-2024-AI-06-FINETUNE

Open-source European foundation model



Objective: To develop and make available one open-source large language foundation model as an infrastructure designed to be largely used by public or private users.



Type: SME Support Action (SME)



Size : EUR 25 Million
- Duration: 24 to 36 Month



Participation: EU + EEA + other participating countries (! ARTICLE 12.6 restriction – entity controlled by non-eligible countries)



DIGITAL-2024-AI-06-FINETUNE

Expected outcomes

A high performing
European large
language
foundation model;
**including all
relevant
components**

Development of a
strong community
with the
establishment of a
coordination
framework

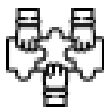


DIGITAL-2024-AI-06-FINETUNE

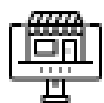
Scope



European large language foundation model: Scaling-up of a selected, open-source foundation model



Coverage of all the official EU languages.



Computing resources for the pre-training of the model should be sourced from HPC facilities, such as EuroHPC



The model should be deployed and made available through the AIO platform and the ALT action and/or other



DIGITAL-2024-AI-06-FINETUNE

Other requirements

Coordinate and build on related actions under HE and DEP

Security restrictions apply (Article 12.6 of DEP regulation)

Attention should be paid to the performance, transparency and security, in compliance with the future AI Act.

SME Support Actions:
50% and 75% (for SMEs)
funding rate



DIGITAL-2024-AI-06-FINETUNE

Other requirements

What are you looking for?

- ✓ A language foundation model competent in all EU languages, ensuring basic functionality and understanding.
- ✓ A language foundation model that can facilitate exploitation and fine-tuning to European SMEs.
- ✓ To foster the creation of an open community around European foundation models.
- ✓ Actors with significant technological know-how and experience in the field.
- ✓ Links with other actions (AloD, ALT-EDIC, etc.)

What do you NOT want?

- Foundation models that cover only some of the EU languages
- Foundation models where only some of the artifacts can be released
- Lack of relevant players with significant experience developing foundation models.
- A proposal focused on research and not on deployment



DIGITAL-2022-CLOUD-AI-03-AI-ON-DEMAND

Key actors

The consortium that will carry this action should be composed by entities with experience in developing foundation models: private companies, including SMEs and start-ups, research and technology organisations, higher education entities and EDIC, or a combination of these.



DIGITAL-2022-CLOUD-AI-03-AI-ON-DEMAND

More Information

Call text

https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/digital/wp-call/2024/call-fiche_digital-2024-ai-06_en.pdf

Work Program

<https://ec.europa.eu/newsroom/dae/redirection/document/100740>



Thank you



© European Union 2020

Unless otherwise noted the reuse of this presentation is authorised under the [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/) license.

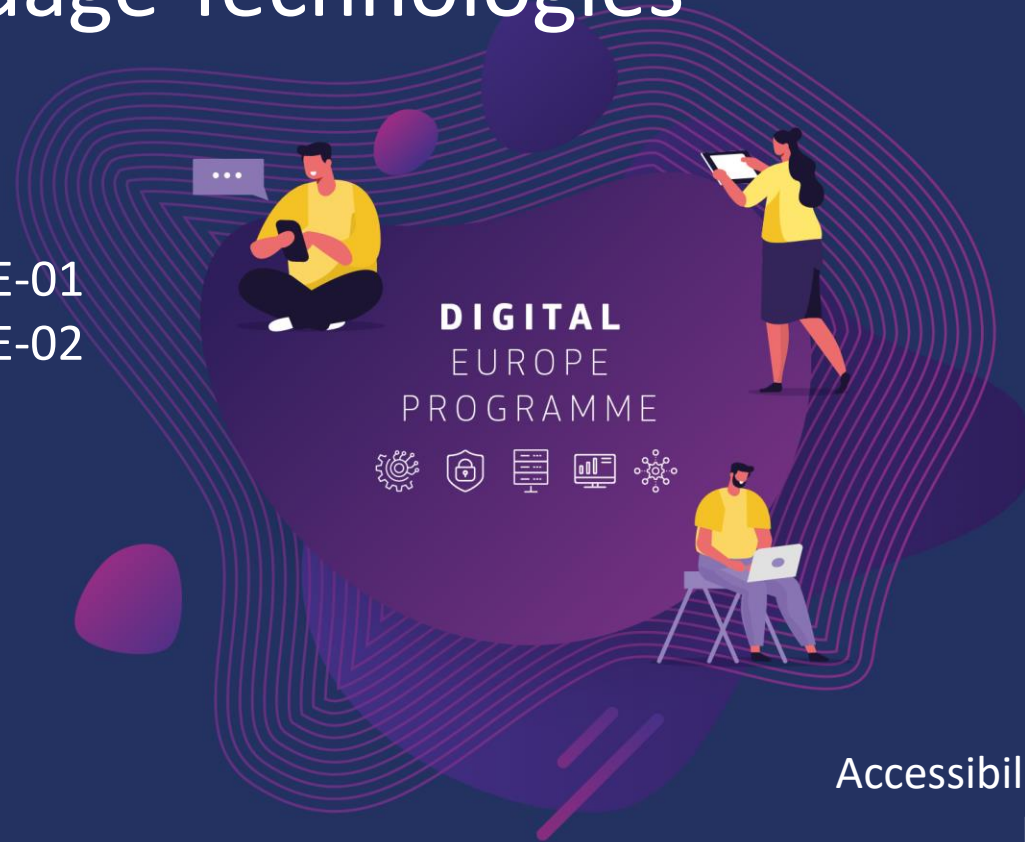


Alliance for Language Technologies Calls:

DIGITAL-2024-AI-06-LANGUAGE-01
DIGITAL-2024-AI-06-LANGUAGE-02

Info Session

March 2024



#DigitalEuropeProgramme

Dhafer LAHBIB
Programme Officer
Accessibility, **Multilingualism** and Safer Internet
European Commission – DG CONNECT



Outline

1. Background and Other European Commission Initiatives
2. DIGITAL-2024-AI-06-LANGUAGE-01 – Alliance for Language Technologies – Data and Finetuning (Simple Grant)
3. DIGITAL-2024-AI-06-LANGUAGE-02 – Alliance for Language Technologies – Ecosystem (CSA)



BACKGROUND

Other initiatives

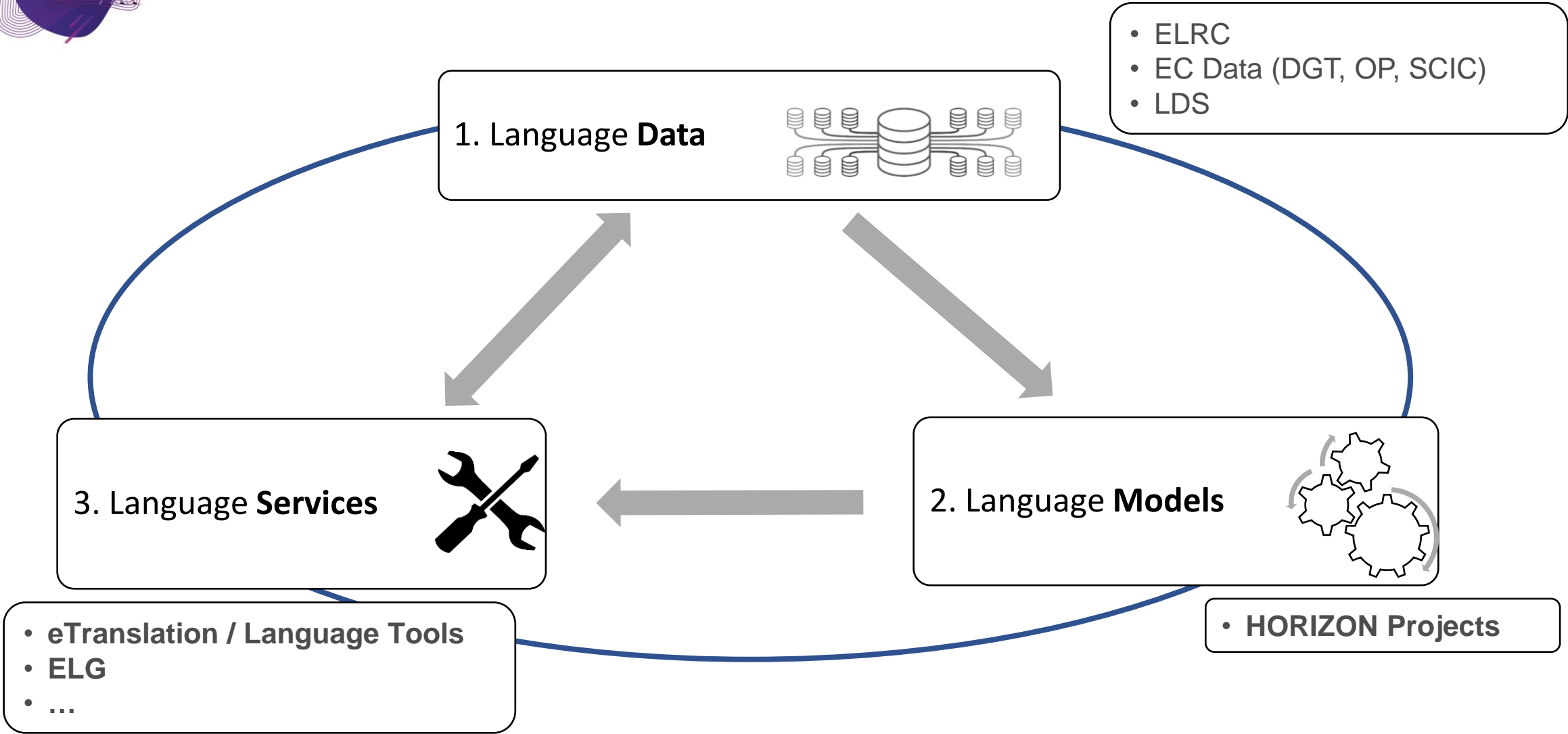


Background

- Language technology plays a crucial role in promoting linguistic diversity, cultural exchange, and socio-economic development within the EU.
- Most advancements originate from regions outside Europe, leading to unequal language coverage.
- Establishing an ecosystem around large language and AI models in Europe will enhance autonomy in the domain and reduce dependence on non-European technologies.



Background & Other Initiatives





Background

- The European Commission, in collaboration with Member States, has set-up the Alliance for Language Technologies European Digital Infrastructure Consortium (ALT-EDIC).
- In alignment with this objective, the European Commission is launching calls within the DIGITAL Programme to support the development of language technologies.



DIGITAL-2024-AI-06-LANGUAGE-01

Alliance for Language Technologies – Data and Finetuning (Simple Grant)



DIGITAL-2024-AI-06-LANGUAGE-01 - Alliance for Language Technologies

Data and Fine-tuning



€ 20 Mio



Simple Grants
50% funding
rate



24-36 months



Restricted on the
basis of Article **12(6)**
of the Regulation
(EU) 2021/694



DIGITAL-2024-AI-06-LANGUAGE-01 – Data and Fine-tuning Scope

Support the Collection of Language Data

- Ensure compliance with legislation like Copyright and GDPR.
- Provide sufficient quality and quantity of data to build large language models covering all official EU languages and socially/economically relevant ones.
- Establish a repository of European Large Language foundational models and models to be adapted to specific languages/domains.

Support **Fine-tuning** of Large Language Models

- Adapt, evaluate, and optimize foundation models for specific languages, domains, or industries.
- Facilitate efficient deployment across industries,
- Support ongoing maintenance and enhancement for adaptability to evolving tasks and domains.
- Provide dedicated support and services, especially for SMEs, for fine-tuning existing models.



DIGITAL-2024-AI-06-LANGUAGE-01 – Data and Fine-tuning

Expected Outcomes

Facilitating access to language data for developing and adapting large language models while addressing data privacy, security, and disinformation risks.

Repository of Existing Large Language Foundation Models, for public and industrial reuse within the EU

Repository of Language Models fine-tuned to specific languages, domains, or industries

Infrastructure and Services for Fine-Tuning



DIGITAL-2024-AI-06-LANGUAGE-01 – Data and Fine-tuning Requirements

What We are Looking For

- Deployment Endeavour to ensure practical implementation and impact
- Engage Industry and SMEs
- Cover all EU Languages for both Data and Models
- Actors with significant technological know-how and experience in the field.
- Synergies with other National and European initiatives

What We do not Want

- Proposal focused on research and not on deployment
- consortia dominated by academic institutions without sufficient representation from industry and SMEs
- Siloed Development and lack of stakeholder's engagement



DIGITAL-2024-AI-06-LANGUAGE-02

Alliance for Language Technologies – Ecosystem (CSA)



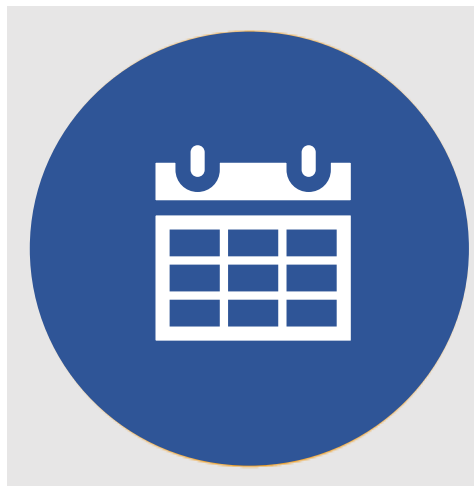
DIGITAL-2024-AI-06-LANGUAGE-02 - Alliance for Language Technologies Ecosystem – CSA



€ 4 Mio



Coordination and
Support Action
Grant
100% funding rate



24-36 months



Restricted on the
basis of Article **12(6)**
of the Regulation
(EU) 2021/694



DIGITAL-2024-AI-06-LANGUAGE-02 – Ecosystem Objectives & Scope

Objectives

- Support the Coordination of a European Language Technologies Ecosystem.

Scope

- Serve as an advisory point, offering awareness raising and community building.
- Support SMEs in integrating large language models into their production workflows.
- Align with national initiatives, Data Spaces, the AI on-demand platform, Testing and Experimentation Facilities, Digital Innovation Hubs, and the EuroHPC Joint Undertaking's supercomputing facilities,
- Ensure synergy with call “DIGITAL-2024-AI-06-FINETUNE” on open-source foundation models.



DIGITAL-2024-AI-06-LANGUAGE-02 – Ecosystem Requirements

What We are Looking For

- The consortium should be composed by representatives of Member States.
- Inclusive Ecosystem
- Sustainability Plan

What We do not Want

- Fragmented Initiatives
- Lack of Alignment with National and European Priorities



Thank you



© European Union 2020

Unless otherwise noted the reuse of this presentation is authorised under the [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/) license.

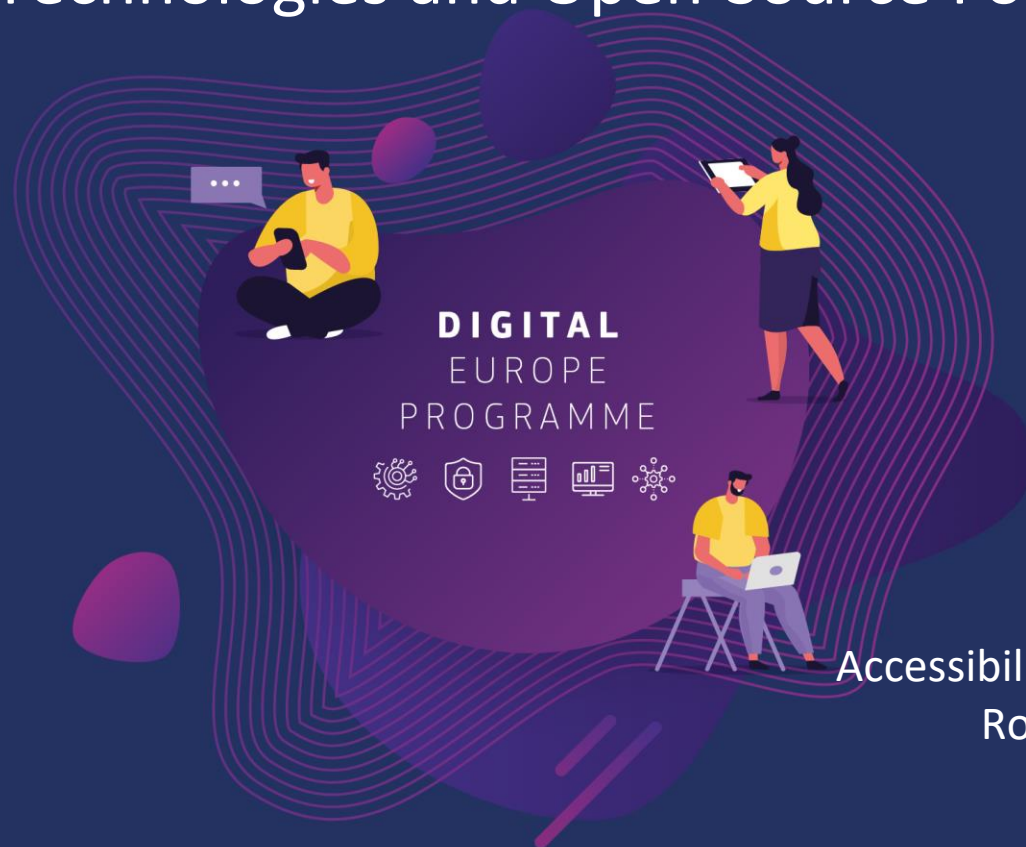


This session will be recorded.
You are invited to switch off
your camera if you wish





Alliance for Language Technologies and Open-Source Foundation Model Open Pitch



Accessibility, Multilingualism and Safer Internet
Robotics & AI Innovation and Excellence
European Commission DG CONNECT

#DigitalEuropeProgramme

March 2024



Agenda

- **ADR – partnership information day and brokerage event**
 - **17 April 2024 (10 – 16.00 CET)**
 - **Breydel Building, European Commission, Brussels, Belgium**
- Information about funding opportunities in Horizon Europe, Digital Europe and the EIC accelerator.
- Opportunity to **matchmake** in order to identify possible collaborations.
- Speakers from the European Commission representing all three programmes



Agenda

- SESTEK - Conversational Automation Company
- HiTZ - Basque Center for Language Technology
- Laion - Ellis
- Le Voice Lab - The French Voice Tech Marketplace
- Luxembourg Institute of Science and Technology
- Tilde - AI powered language technologies
- Open LLM
- Nijta - Speech & text anonymization
- Fraunhofer IAIS - Natural Language Understanding Team



Thank you

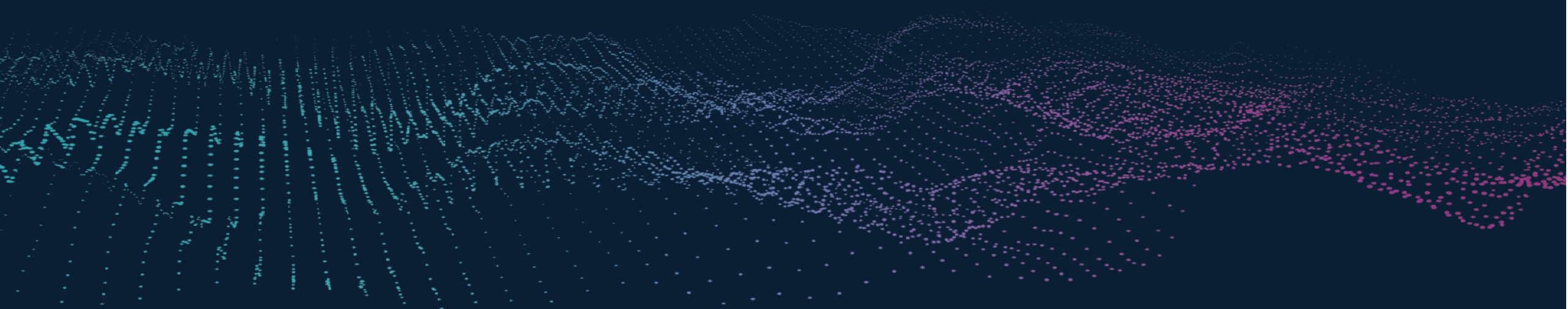


© European Union 2020

Unless otherwise noted the reuse of this presentation is authorised under the [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/) license.

SESTEK

Conversational Automation Company



SESTEK

R&D CAPABILITIES

2

R&D CENTERS

88%

OF REVENUE =
R&D → COMMERCIALIZATION

+35

FUNDED R&D PROJECTS

INCLUDING

- 1 H2020
- 1 EUROSTARS
- 1 GLOBALSTAR
- 4 EUREKA

11

PATENTS

+90

PUBLICATIONS

100% In-
house
developed
products

400
customers in
20 countries

Recognized by leading
consultancy firms

Gartner

DMG
CONSULTING LLC

opusresearch

SESTEK



Virtual Assistants

Speech Recognition (SR)

Text-to-Speech (TTS)

Natural Language Processing (NLP)

Large Language Models (LLM)



Active Verification

Passive Verification

Blacklist identification



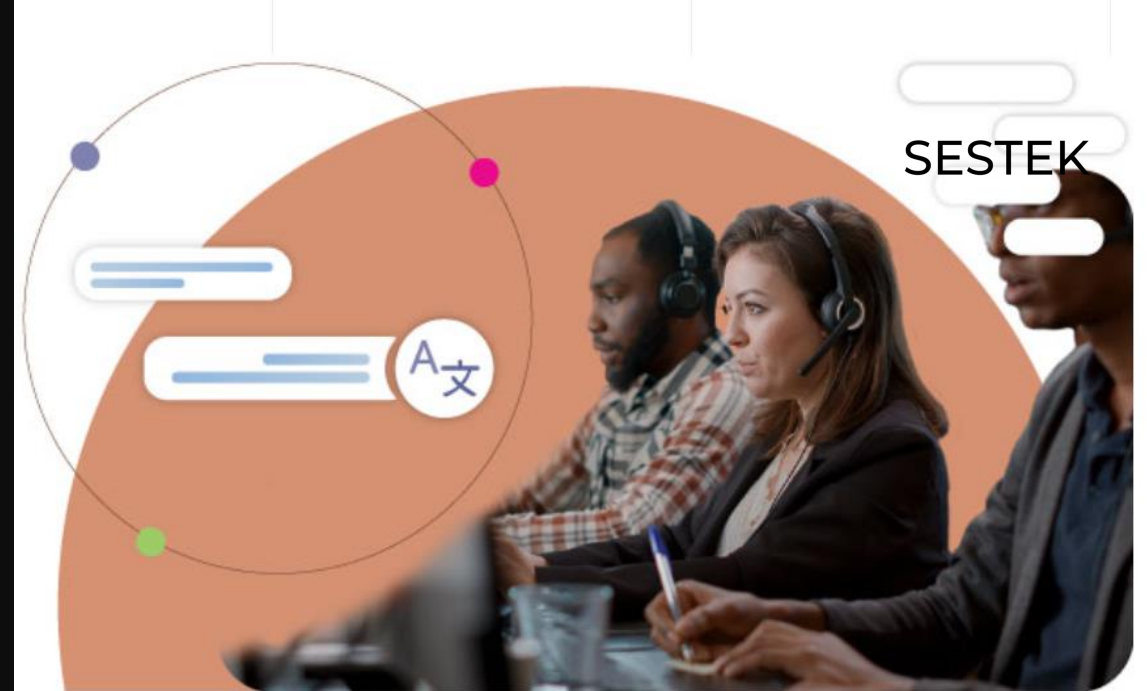
Speech & Text Analytics

Real-time Guidance

Automated Quality Management

EXPERTISE OFFER - DIGITAL-2024-AI-06-LANGUAGE-01

- FINETUNE
- LLMs tailored for user interactions in multi/bilingual different languages
 - ChatGPT3.5 Turbo, ChatGPT4.0, Gemini, Aya, Lama2, Mixtral, Falcon
- Context aware **dialog data generation**
- Addressing Multilingual Communication Challenges – LLM backed virtual translator
- Prompt Engineering
- Training Simulation
- Persona Creation

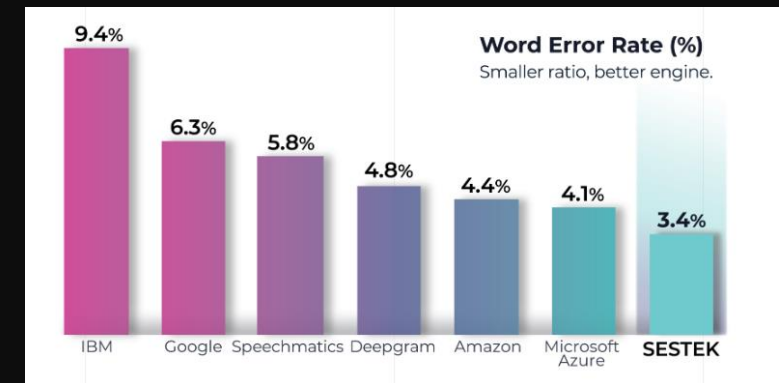
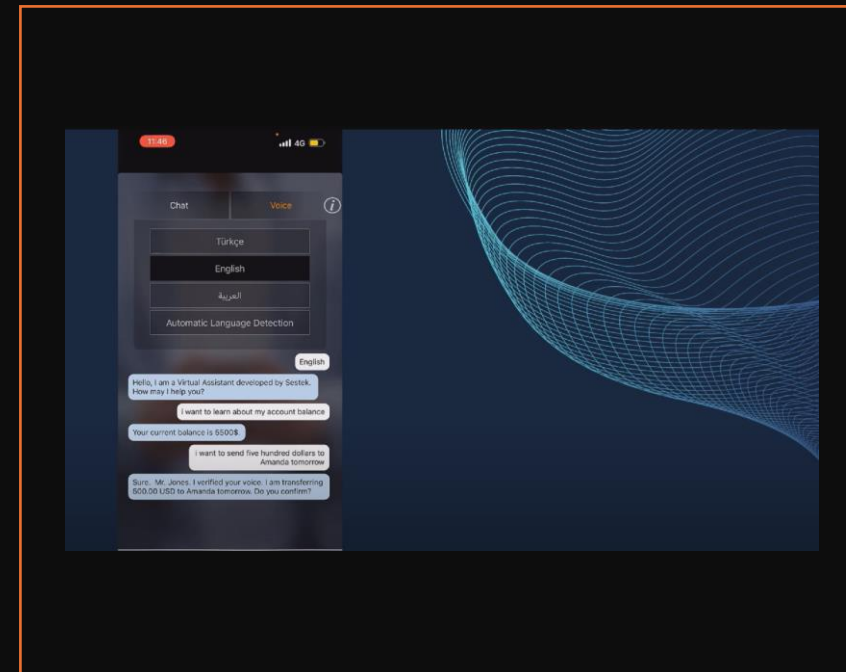


It had all my credit cards and passport in it! Can you help me?



EXPERTISE OFFER - DIGITAL-2024-AI-06-LANGUAGE-01

- FINETUNE
- LLM backed human-machine interaction
 - Content summarization
 - Document/web page-based generative QA
- Speech recognition → text correction for novel multilingual pre-trained language models
- Personally Identifiable Information (PII) tracking & anonymization
 - person/location/organization entities exist in the training set
- Entity mapping
- Dialog based finetuning
- Input – output moderation / open-source guarding / transactional tasks





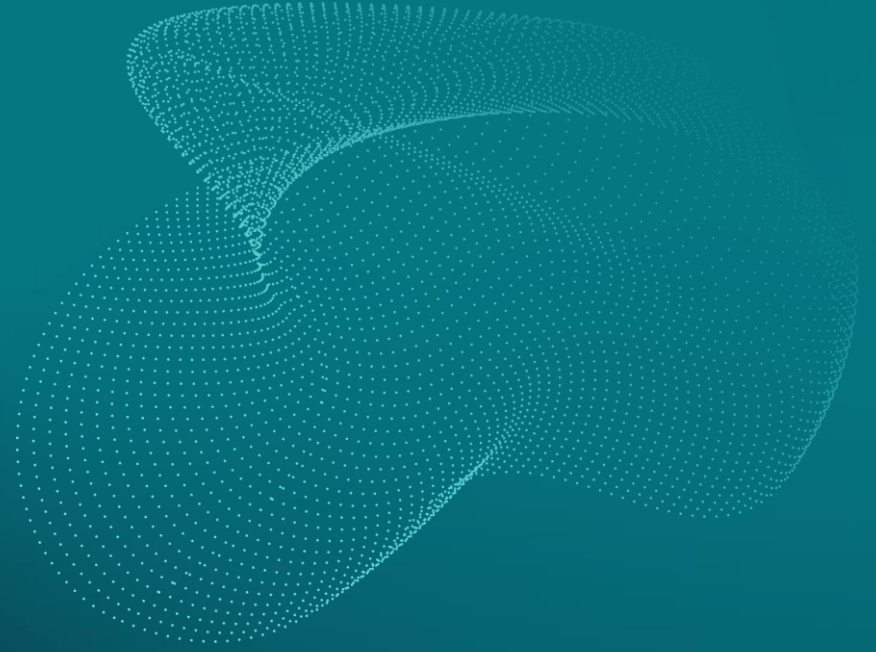
Tuba ARSLAN KIR
R&D&I Director
tuba.arslan@sestek.com



Ayşe GÜVENTÜRK
R&D Project Coordinator
ayse.guventurk@sestek.com



Ömer Faruk KÜRKÜ
R&D Coordination Specialist
omer.kurklu@sestek.com



Thank you!

SESTEK

sestek.com

sestek.com/demos

[in /sestek](https://www.linkedin.com/company/sestek)

 **knovvu**
Virtual Agent

 **knovvu**
Biometrics

 **knovvu**
Analytics

 **Garanti BBVA**

vodafone

 **QNB**
FINANSBANK

 **VakıfBank**

hepsiburada

 **astellas**
Leading Light for Life

 **BNP PARIBAS**

 **بنك دبي الإسلامي**
Dubai Islamic Bank

ode

Dem

to



Universidad
del País Vasco

Euskal Herriko
Unibertsitatea

HiTZ

Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

HiTZ

Basque Center for Language Technology UPV/EHU

Eneko Agirre

Director

hitz.eus



HiTZ Basque Center for Language Technology

University of the Basque Country, Computer Science Faculty (UPV/EHU Spain)

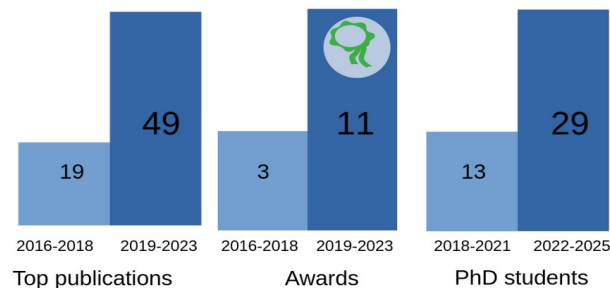
- Ixa NLP (founded 1988) and Aholab Speech (1998)
- 88 members, multidisciplinary with computer science core
- Owns strong infrastructure: 100 A100 GPUs
- Yearly budget: 2,5M € (plus overheads covered by Univ.)

Reference center in NLP research in the Basque country and Spain

- Research & Applications on English, Spanish and other languages
- Spanish coordinators of two European infrastructures CLARIAH-ES
- European Language Equality (core partner)
- Consultants and relevant actors in Basque and Spanish LT plans

HiTZ – Scientific excellence

- Spain: most publications on top conferences, most cited LT researchers
- Two Spanish national prizes in Computer Science
- Fellow of the Association for Computational Linguistics
- Best PhD on AI 2020 (EurAI).
- Six best papers. Three Google Awards.
- Expertise in foundational models:
 - Co-founder of Reka is active member (Mikel Artetxe)
 - Alumni at LLM startups co-supervising PhDs: Cohere, Reka



HiTZ – Areas



**Information Extraction
and Information
Retrieval**



Machine Translation



**Human-Computer
Interaction**



**Speech Synthesis and
Speech Recognition**



Text Analysis



**Speech and Language
Resources**



**Medical and Legal
domains**



**Digital humanities
and education**

HiTZ – key projects

Discovery of unsupervised Machine Translation

New research area

JOURNALS ▾ | COVID-19

Science

NEWS | SOCIAL SCIENCES

Artificial intelligence goes bilingual—without a dictionary

“Unsupervised” machine learning could help translate between uncommon languages

28 NOV 2017 • BY MATTHEW HUTSON

Google Scholar

unsupervised neural machine translation



Articles

About 107,000 results (0.08 sec)

Unsupervised neural machine translation

[M Artetxe](#), [G Labaka](#), [E Agirre](#), [K Cho](#) - arXiv preprint arXiv:1710.11041, 2017 - arxiv.org
... clear how it would perform in our more challenging **unsupervised** scenario, as it might be difficult to learn the **translation** relations between subword units. For that reason, we also experiment at the word level in this **unsupervised** scenario, limiting the vocabulary to the

☆ Save ⓘ Cite Cited by 627 Related articles All 7 versions ⌕

When and why is **unsupervised neural machine translation** useless?

[Y Kim](#), [M Graça](#), [H Ney](#) - arXiv preprint arXiv:2004.10581, 2020 - arxiv.org
... the current state-of-the-art **unsupervised** methods in **neural machine translation** (NMT) for **translation** tasks with various data settings, we analyze the conditions under which the **unsupervised** methods fail to produce reasonable **translations**. We show that their performance

☆ Save ⓘ Cite Cited by 31 Related articles All 7 versions ⌕

Unsupervised neural machine translation with weight sharing

[Z Yang](#), [W Chen](#), [F Wang](#), [B Xu](#) - arXiv preprint arXiv:1804.09057, 2018 - arxiv.org
... **Unsupervised neural machine translation** (NMT) is a recently proposed approach for **machine translation** which aims to train the model without using any labeled data. The model proposed for **unsupervised** NMT often use only one shared encoder to map the pairs of ...

☆ Save ⓘ Cite Cited by 102 Related articles All 6 versions ⌕

☐ include citations

☒ Create alert

HiTZ – key projects

BETTER program (Gob. EEUU – IARPA)

- 3 years and a half (2019-23)
- 11 American universities
- MITRE, NIST, BBN, Raytheon

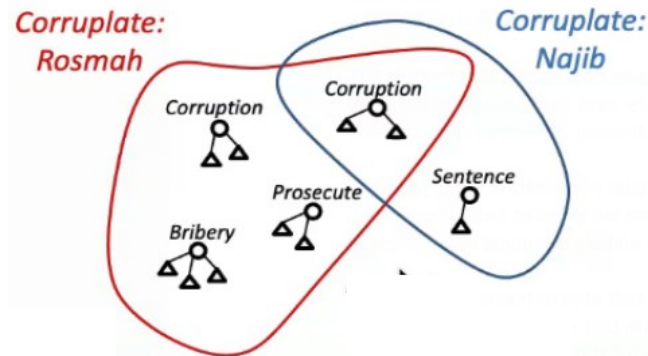
Information Retrieval and Extraction

LT research in the frontier:

- Human in the Loop
- Few-shot with pre-training models (GPT-3)



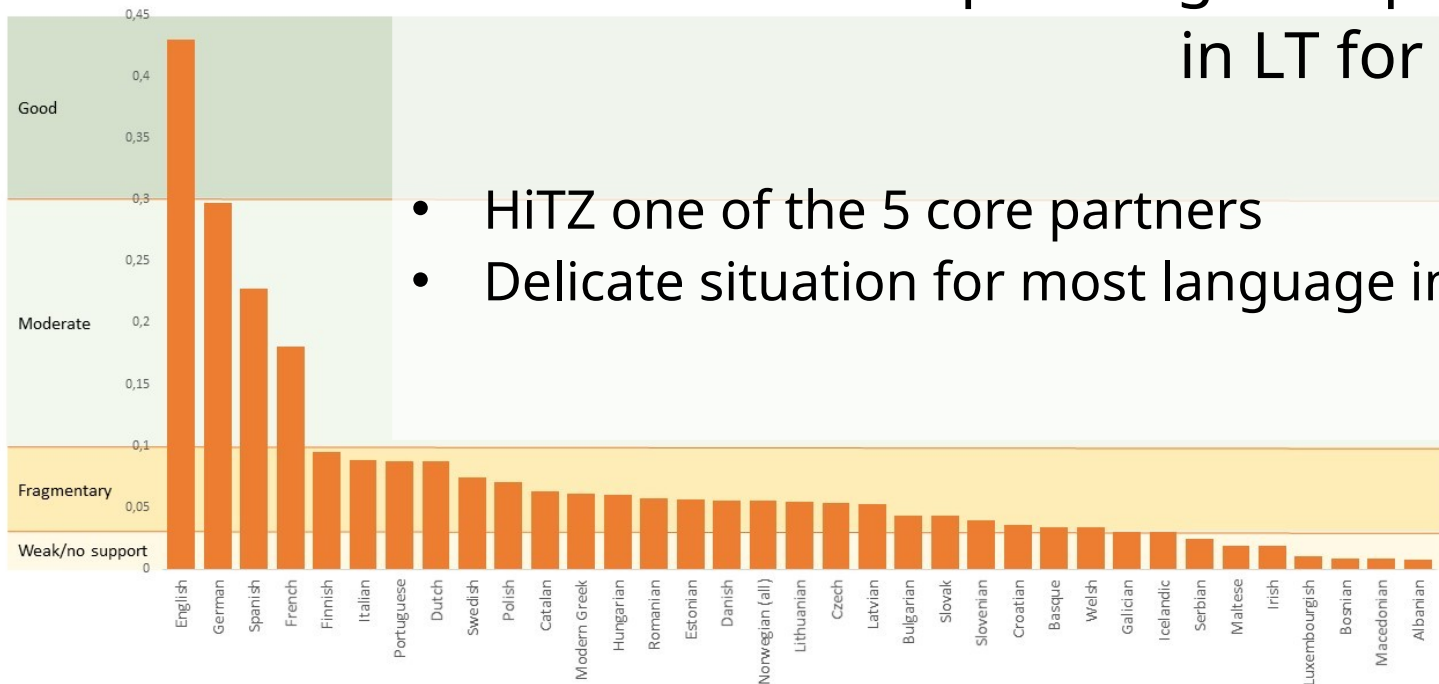
A Malaysian court on Thursday ordered Rosmah Mansor, the wife of former prime minister Najib Razak, to enter a defence in her corruption trial, where she is accused of soliciting and receiving bribes. Both Najib and Rosmah have been accused of corruption; on 08 July 2020, two years after he lost the general elections and stepped down as prime minister, Najib was sentenced to 12 years' imprisonment.



HiTZ – key projects

Roadmap for digital equality in LT for 2030 Europe

- HiTZ one of the 5 core partners
- Delicate situation for most language in Europe



HiTZ – key projects

German Rigau (co-director) lead the entry of Spain in:

CLARIN: Common LAnguage Resources and Technology INfrastructure

DARIAH:
Digital Research Infrastructure for the Arts and Humanities

HiTZ coordinates the infrastructures in Spain.
German has been officially appointed representative of Spain by Ministerio de Ciencia



Estados de la UE que son miembros oficiales de las infraestructuras CLARIN y/o DARIAH.

Fuente: <https://www.clariah.de/>

HiTZ – Ongoing research avenues

- Lack of **reasoning, grounding** to reality and **truthfulness**:
 - Negation, factual accuracy
 - **Visiolinguistic models**
 - Speech recognition and synthesis
- Issues with **structured** modalities
 - Tables and pdfs in the input
 - Extracting Knowledge Bases from text
- High-cost for **domain adaptation**:
 - Combining specialised and general models, cross-lingual transfer
 - Specific models for medical domain / machine translation
- Building **foundation models for low-resource languages**

HiTZ – key projects on building LLMs

IKER-GAITU and ILENIA

Spanish and Basque Governments (3.5 M€, 2023-2025)

- LLMs for text, speech, multimodal, multilingual
- Data sources and evaluation of LLMs
- High impact use cases on industry
- Focus on Basque, include Spanish, Galician, Catalan (with partners)

EUROHPC 1.4M GPU hours

- Research on building LLMs for low-resource languages

Research on LLMs for low-resource lang.

Basque ranks 50th in Common Crawl (from ca. 1000)

Is it possible to build a high-performance monolingual base LLM for Basque?

- Extend pre-train corpus: 4.3M documents and 1.2B words
- Continual pre-training > scratch monolingual
- Evaluation: Proficiency, Reading comprehension, Trivia, Exams

LATXA: Open Framework to research on LLMs
for Basque based on LLAMA2

- Largest basque LLM built to date **7B, 13B, 70B**
(Largest LLM trained in Spain)
- Includes: Corpora, code, models, evaluation
- Obtains state-of-the-art results
- Soon available at arxiv (ACL submission)



Research on LLMs for low-resource lang.



Currently working on related questions for low-resource languages:

- Does it apply to other low-resource languages?
- How to build a high-performance multilingual LLM
- How to instruct and align

Interested on:

DIGITAL-2024-AI-06-LANGUAGE-01

DIGITAL-2024-AI-06-LANGUAGE-02

DIGITAL-2024-AI-06-FINETUNE

Thank you!

Eneko Agirre

@eagirre

hitz.eus/eneko



Open- ψ
Collective



Open-source foundation models with strong reasoning, multi-modality & multi-linguality

Open-Sci Collective with LAION, ELLIS & Friends

Large-scale Artificial Intelligence Open Network (LAION)

European Laboratory for Learning and Intelligent Systems (ELLIS)

Research communities for open foundation models

- **Large AI foundation models** (GPT, CLIP): until 2022 - **closed** data & models, despite crucial role in the ML/AI field and beyond;
→ **non-reproducible!**
- Rise of **grassroot research communities** to open-source and study those
 - **EleutherAI** (USA), **BigScience** (EU, France), **LAION** (EU, Germany)
 - **LAION**: core formed (spring-summer 2021) around grassroots dataset composition efforts and model training on **JUWELS Booster** (state funded supercomputer in Germany, hosted by **Juelich Supercomputing Center**, JSC, Research Center Juelich, Helmholtz Association)
 - **Open large datasets and open foundation models** originating in **EU**



JÜLICH
Forschungszentrum

JÜLICH
SUPERCOMPUTING
CENTRE

Open foundation models for broad community

- Problem – composing & studying whole pipeline for open foundation models is challenging: requires
 - **large-scale data** (at least 100M of samples)
 - **large-scale compute** (GPU years per single experiment)
 - **expertise** in large-scale machine learning
 - → Broad research community cut off from training & studying strong foundation models at larger scales

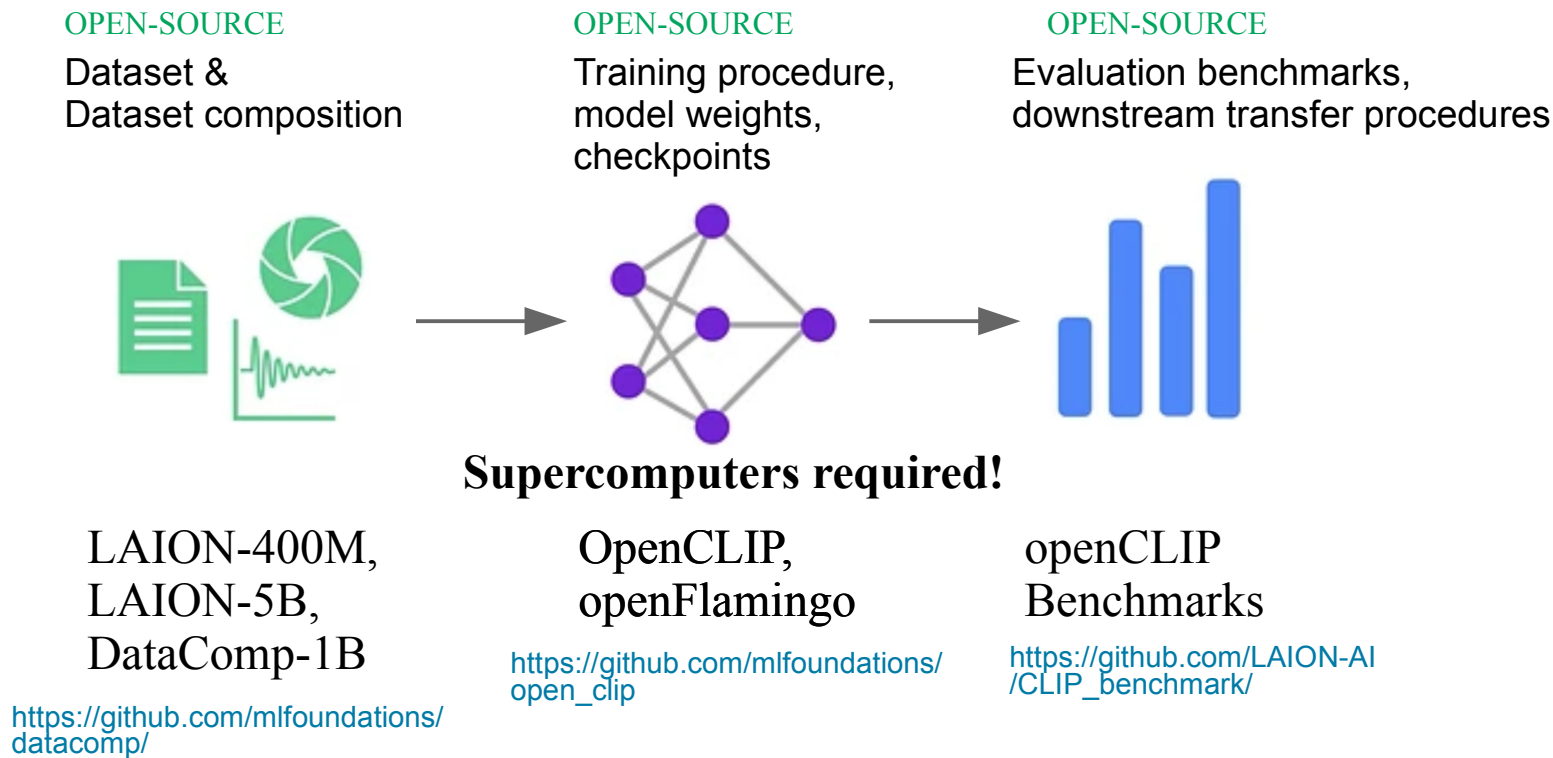
→ Solution



Registered as **non-profit research LAION e.V.** since 2021 in Hamburg

Open-source foundation models & datasets

- Making **whole pipeline** – dataset composition, model training, benchmarks & evaluation – **fully reproducible**



Open-source foundation models & datasets

- Making **whole pipeline** – dataset composition, model training, benchmarks & evaluation – **fully reproducible**

OPEN-SOURCE

Dataset &
Dataset composition



BigScience 

EleutherAI

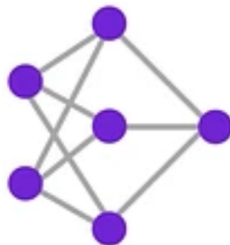
together.ai 

Ai2

Pile,
RedPajama,
Dolma

OPEN-SOURCE

Training procedure,
model weights,
checkpoints



BigScience 

Ai2

Pythia, Together-
INCITE, Olmo

EleutherAI



OPEN-SOURCE

Evaluation benchmarks,
downstream transfer procedures



Lm-eval-harness, bigcode-evaluation-harness,

<https://github.com/EleutherAI/lm-evaluation-harness>

EleutherAI

Ai2

BigScience 

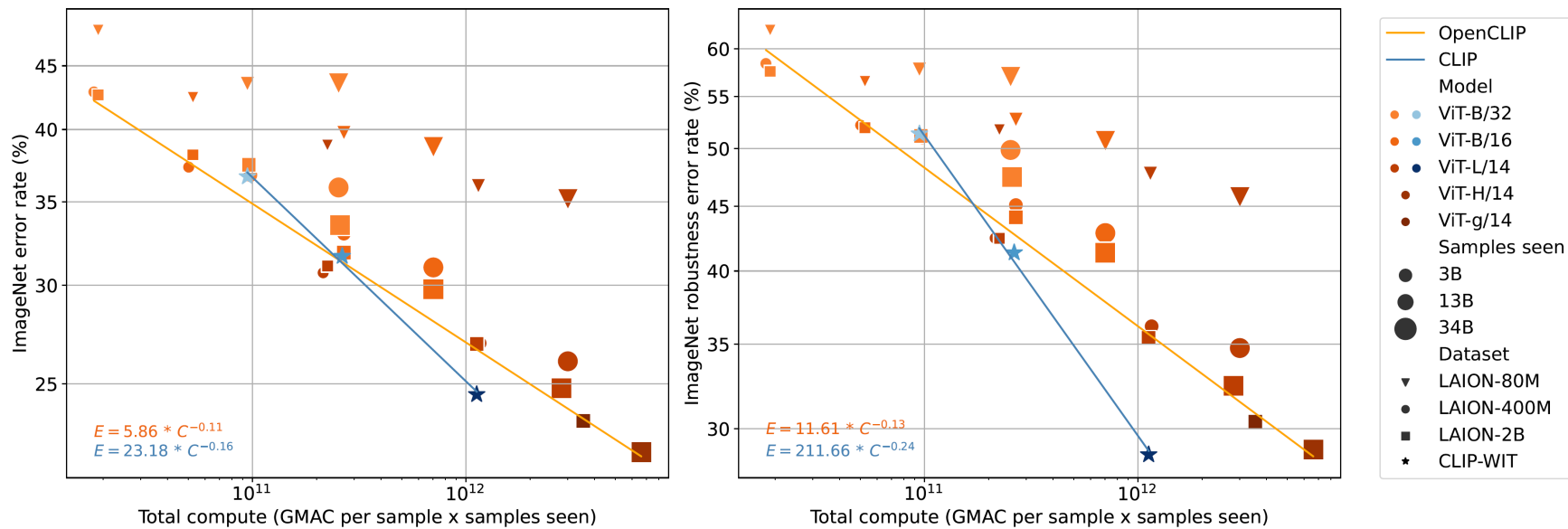
Open datasets & foundation models @ LAION

- Open-source foundation models
 - OpenCLIP ViT B/32 - G/14: **language-vision learning** at larger scale
 - openCLIP-CoCa: image-to-text generative
 - Stable Diffusion, openImagen, Paella, Wuerstchen: **text-to-image generative models**
 - OpenFlamingo-3B/4B/9B: **interleaved language-vision learning**, image-text sequences, text generative
 - together-INCITE-3B/7B; OA-falcon-7B/40B; LeoLM-3B/7B/70B (German tuned LLaMA 2): **language models (from scratch and fine-tuned)**
 - LAION-CLAP, MusicLDM: **language-audio learning**
- Open reference datasets, required for foundation models training
 - LAION-400M, LAION-5B, DataComp-1B (used by: openCLIP, Stable Diffusion, FLAVA, EVA, ...)
 - LAION-audio-630k: language-audio (CLAP, MusicLDM)
 - RedPajama v1 (Together AI, Ontocord.ai, ETH, University Montreal, Stanford – together-INCITE language models & many others)



Reproducible scaling laws for foundation models

- Scaling laws with LAION-400M/2B and openCLIP: open-source data, models and code - reproducible science of foundation models



Open science for large-scale foundation models

- **LAION: Large-scale Artificial Intelligence Open Network**
 - **compute**: applying for publicly funded supercomputers
 - **JUWELS Booster**, Germany: Gauss Center for Supercomputing
 - **Summit**, USA: INCITE Leadership computing call
 - **LUMI** (Finland), **Leonardo** (Italy): **EuroHPC** calls



Open science for large-scale foundation models

- Supercomputers in EU – hubs for large-scale basic AI research
- Open science for advancing powerful, safe generic AI tools for public



*Stable Diffusion 1.5, trained on **LAION-5B** image-text dataset.*

Prompt: "An epic scene of a supercomputing center building of the future, embedded in a rich wild green exotic blooming jungle forest, nearby a lake"



LAION: strong grassroots research community

- Collaborative work of broadly distributed community: **Outstanding NeurIPS 2022 paper award**, strongly impacting open source releases
- **Falling Walls Award: Scientific Breakthrough 2023**
- LAION public Discord server: > **27k members**



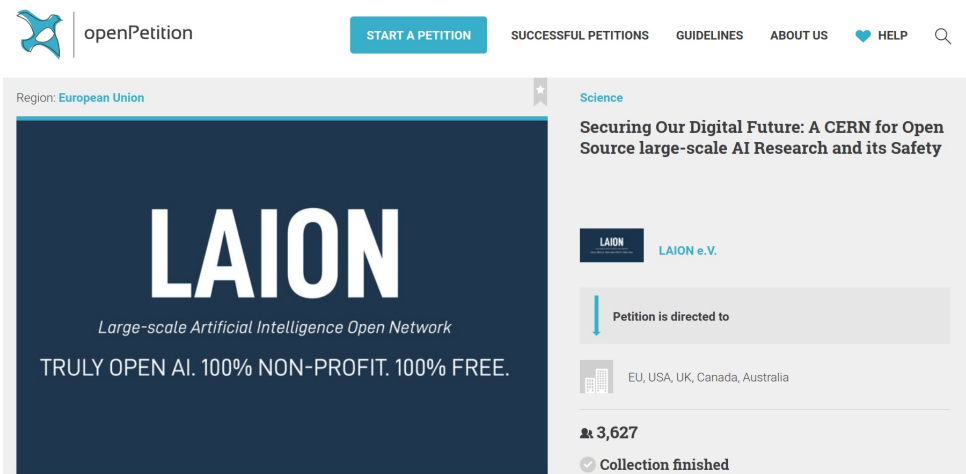
Bringing large-scale AI to research community

- **LAION: Large-scale Artificial Intelligence Open Network**
 - **expertise**: strong grassroots research community skilled in large-scale experiments and distributed training
 - Juelich Supercomputing Center, University of Washington, Allen AI Institute, UC Berkeley, Stanford U, U Tel Aviv, U Montreal, ...
 - Making whole pipeline – **dataset composition, model training, evaluation** – **fully reproducible** for important foundation models
 - Able to handle distributed training on very large machines
 - Impacting open-source releases in various public repositories:
 - HuggingFace: <https://huggingface.co/laion>
 - LAION repository: <https://github.com/LAION-AI/>
 - ML Foundations repository: <https://github.com/mlfoundations/>



LAION: research community & alliances

- Various alliances in EU: **ELLIS**, Tuebingen AI Center, MPI for Intelligent Systems, WestAI, Hessian AI & TU Darmstadt, HuggingFace, FAIR, U Turku & SILO AI (**HPLT**), German National Library, ...
- Various alliances worldwide: U Washington, Allen AI Institute, Stanford, Together AI, Ontocord AI, U Montreal, Tokyo Tech, U Berkeley, U Tel Aviv, ...



Open-sci: foundation models & strong reasoning

- **Open-Sci Collective**: recent ongoing effort to create **strong open base language model family** for open research & development
 - **strong reasoning** by using math, code, natural science datasets (high quality books, exercises, papers); **synthetically generated data** (à la Phi)
 - injecting **transferable multi-linguality** by covering broad language family (ca. main 11 families)
 - training to get models **re-usable & extendable**: continued training from pre-trained checkpoints
- Establishing strong reproducible base for further research & development

Open- ψ
Collective

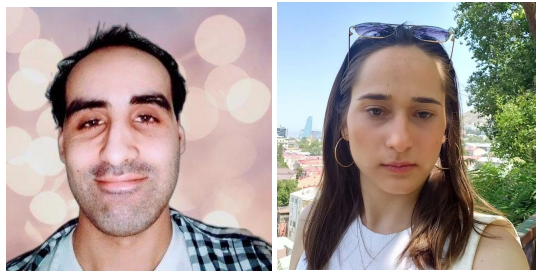


Open foundation models: outlook

- „Moonshot“: build upon open-sci towards **open-sci-MM - open multi-modal foundation model, learning with any modality – text, vision, audio, ...**
 - Strong impact across various disciplines beyond core machine learning
 - Focus on reasoning, coding, complex workflow automation
 - Foundation models for science & semi-automated scientific discovery
 - Customized AI assistants for citizens, for governance, for education, ...
researched, developed and deployed in EU from open base validated by broad community
- „**LAION/ELLIS/BigScience 2.0**“ : Germany/France (Italy/Spain/Netherlands/Finland/Israel/...) - EU consortium for building large open foundation models that are powerful, transparent and **validated by research community for safe fine-tuning and deployment**



Acknowledgements



Dr. Mehdi Cherti, Marianna Nezhurina,
JSC



LAION community & friends (Romain Beaumont, Ross Wightmann, Huu Nguyen, ...)



Prof. Ludwig Schmidt, UoW



Christoph Schumann

Visit <https://laion.ai/>
Join public LAION Discord server
for more projects
and research tracks
> 27k members !

**Let's build open, strong, safe
AI foundations together!**



FEDERATED LANGUAGE INFRASTRUCTURE FOR LOW-RESOURCE LANGUAGES

Jordi Cabot

Head of the Software Engineering RDI Unit
LIST

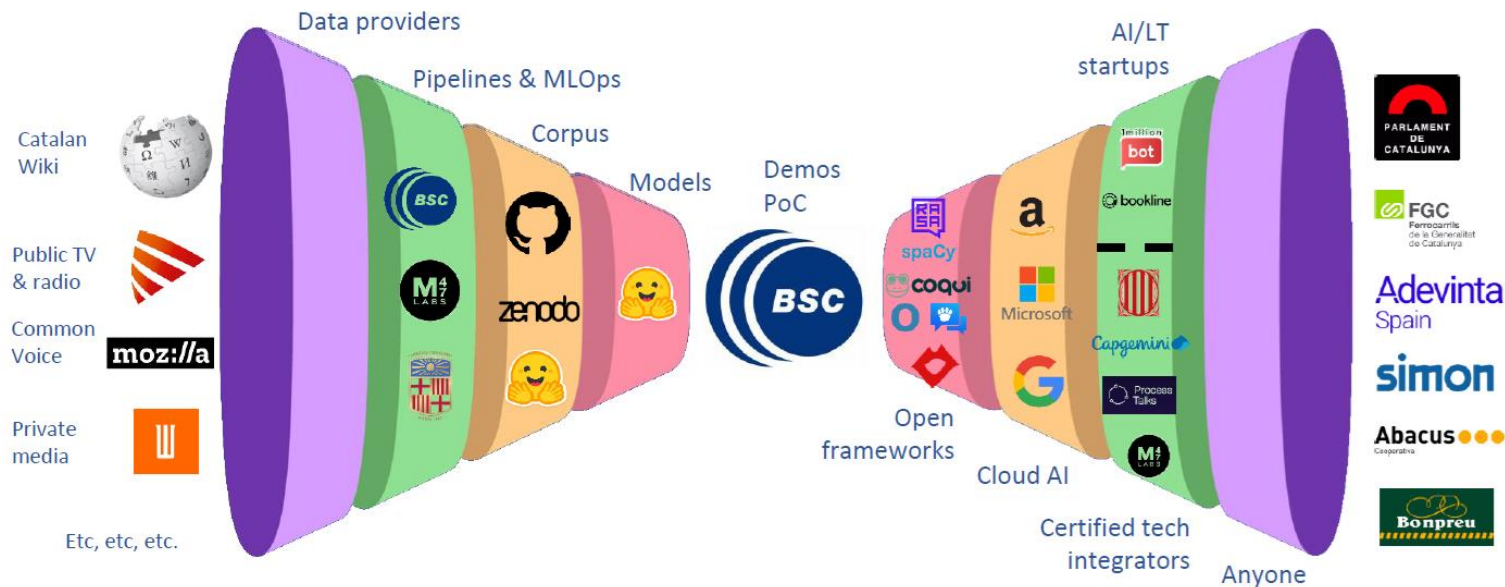
Jordi.cabot@list.lu

05/03/2024

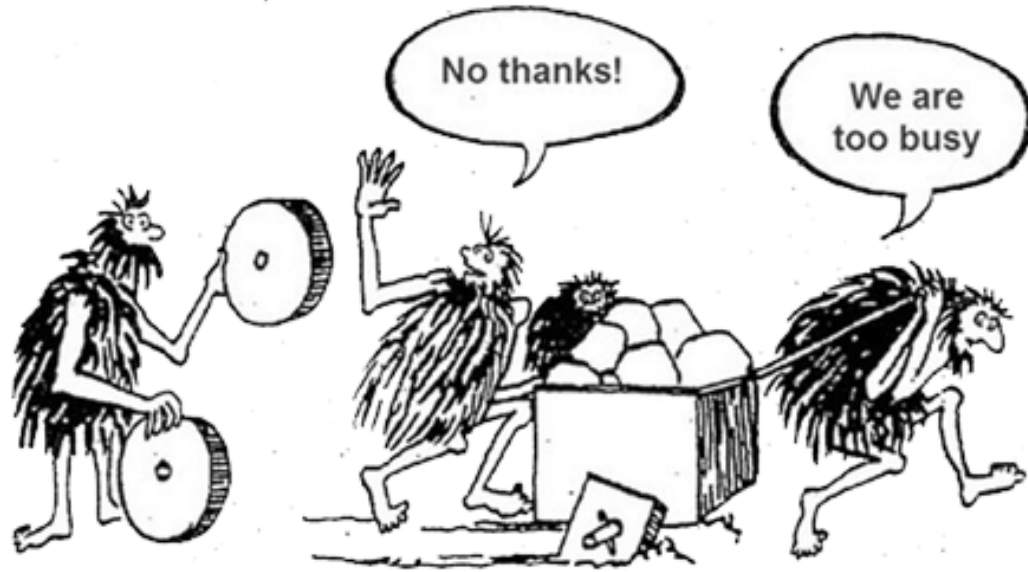
***Low-resource languages** are languages with limited linguistic data and resources, which compromise their presence in our current AI-driven world*

Many EU official languages fit into this category (and if we include Regional or Minority Languages, all EU countries are likely affected)

MORE THAN A MODEL, WE NEED A LANGUAGE INFRASTRUCTURE



NOT INVENTED HERE SYNDROME



Let's cooperate instead of
reinventing the wheel!

TOWARDS A FEDERATED LANGUAGE INFRASTRUCTURE

Joint-development

Of strategies, best practices, techniques and methods, e.g. for data augmentation, model training,...

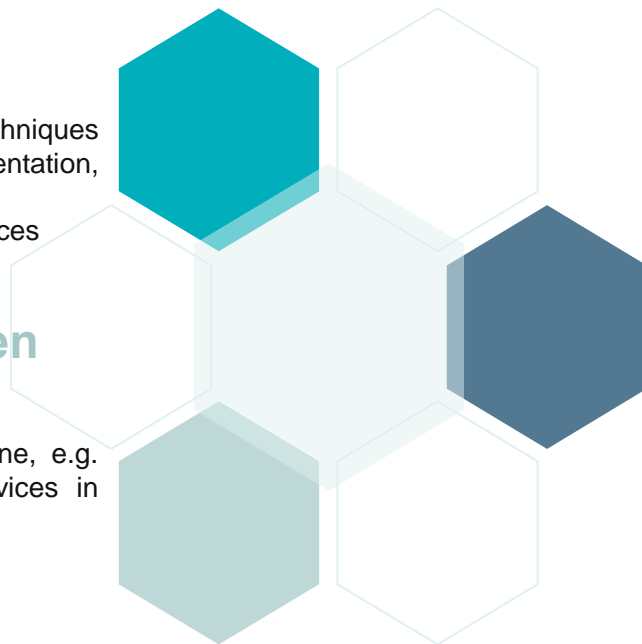
Sharing also computational resources

Open source & open access

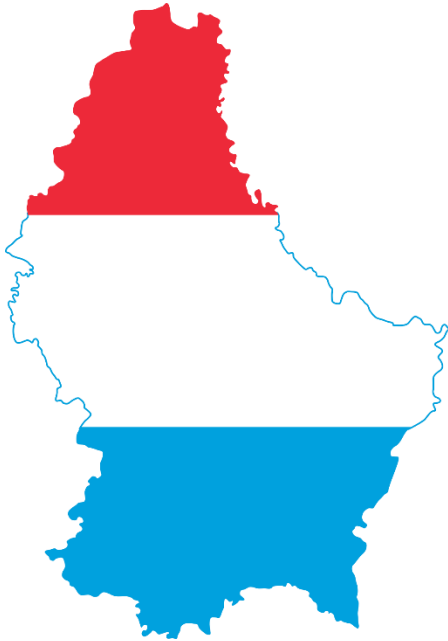
Sovereign and available to anyone, e.g. SMEs interested in offering services in any of the languages.

Living Labs approach

To co-create with the whole ecosystem of actors and society in general



WHAT WE COULD BRING



Multi-lingual and multi-cultural environment (3+1 core languages in a small country)

Dynamic and diverse language community involving all types of actors

MORE SPECIFICALLY...



Expertise in data modeling and processing

- Involved in the Language Data Space, European Language Equality, European Language Grid projects (Dimitra Anastasiou)
- Involved in “Croissant” an MLCommons initiative to describe ML datasets (e.g. including provenance)

Technical expertise in ML/AI topics

- Developed a flexible open-source chatbot framework
- Test suite for testing of biases in LLMs
- Low-code techniques for building smart systems
- ...

Interested? Let's chat!



Jordi.cabot@list.lu



Tilde – **AI powered** language technologies



Andrejs Vasiljevs

Co-Founder and Board Member
andrejs@tilde.com



AI-powered solutions

Tilde is one of the leading language technology companies in Europe



Market leading products and technologies

Tilde Machine Translation, Tilde Term, Tilde Virtual Assistants, Tilde Speech Solutions



Award-winning MT

Four times winner at the global Machine Translation competition WMT



Passionate team

Offices in 3 Baltic capitals
150+ employees



500+ business clients worldwide

Including the European Commission, Finland's Prime Minister Office, Estonian Government, SAP, Microsoft, IBM, Oracle and many others



Excellence in research & innovation

11 PhDs, 220+ scientific publications.
Cooperation with 30+ universities and research centres



Trusted technology partner for EU institutions

Tilde has supported 8 Presidencies of the Council of the European Union by providing custom machine translation solutions



AI-powered language technologies



MACHINE TRANSLATION

Crossing language barriers



SMART VIRTUAL ASSISTANTS

Intelligent chatbots

MULTILINGUAL LANGUAGE MODELS



NATURAL LANGUAGE UNDERSTANDING

Access and share knowledge



SPEECH TECHNOLOGIES

Voice communication &
accessibility

Expertise in Language Models



- LLM Fine-Tuning
- Retrieval Augmented Generation
- Language Models for Machine Translation, Speech Recognition, NLU
- Foundational LLM for Balto-Slavic Languages
preparatory work
- Latvian National Large Language Model
preparatory work

Expertise in Processing Language Data



- Data crawling
- Data cleaning
- Data anonymization
- Data curation
- Data management
- Synthetic Data



Core consortium member of the key **Language Data** initiatives

- ALT-EDIC (Alliance for Language Technologies European Data Infrastructure Consortium)
Member of Latvia Consortium
- European Language Data Space
- ELG - European Language Grid
HORIZON 2020 Programme
- ELRC - CEF DIGITAL European Language Resource Coordination Action
- META-SHARE
Multilingual Europe Technology Alliance

ALT-EDIC



LLMs as part of National Language Technology Infrastructure



LARGE LANGUAGE
MODELS

ANONYMIZATION

MACHINE
TRANSLATION

CHATBOTS

SPEECH
PROCESSING

NATIONAL LANGUAGE TECHNOLOGY
PLATFORM

Contribution to the project



- Experts and expertise in **Language Models**
- Adapting and assessing LLMs for **real-world applications**
- Tools and services for multilingual **data processing**
- **Multilingual data** resources
- Synergy with **Language Data Infrastructures**
- Synergy with **National initiatives**
- Link to the **Big Data** community (BDVA)
SMEs & Startups

Thank you!



Andrejs Vasiljevs

Co-Founder and Board Member
andrejs@tilde.com



OPEN LLM



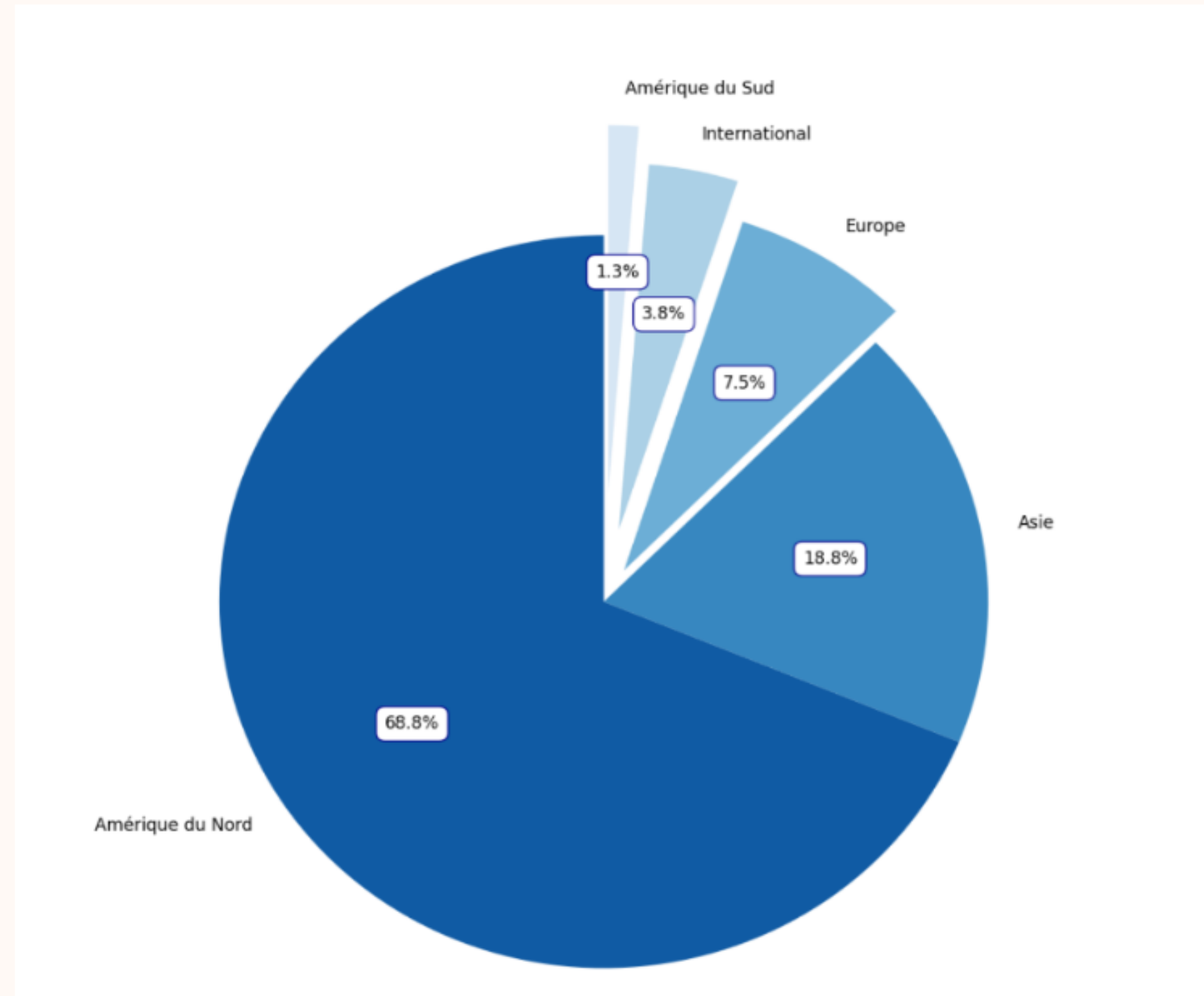
One LLM to free them all

THE NORTH AMERICAN BIAS IN CURRENT LLMS

Language	Percent	Language	Percent
en	89.70%	uk	0.07%
unknown	8.38%	ko	0.06%
de	0.17%	ca	0.04%
fr	0.16%	sr	0.04%
sv	0.15%	id	0.03%
zh	0.13%	cs	0.03%
es	0.13%	fi	0.03%
ru	0.13%	hu	0.03%
nl	0.12%	no	0.03%
it	0.11%	ro	0.03%
ja	0.10%	bg	0.02%
pl	0.09%	da	0.02%
pt	0.09%	sl	0.01%
vi	0.08%	hr	0.01%

LLAMA V2 : Language distribution in pretraining data

THE NORTH AMERICAN BIAS IN CURRENT LLMS



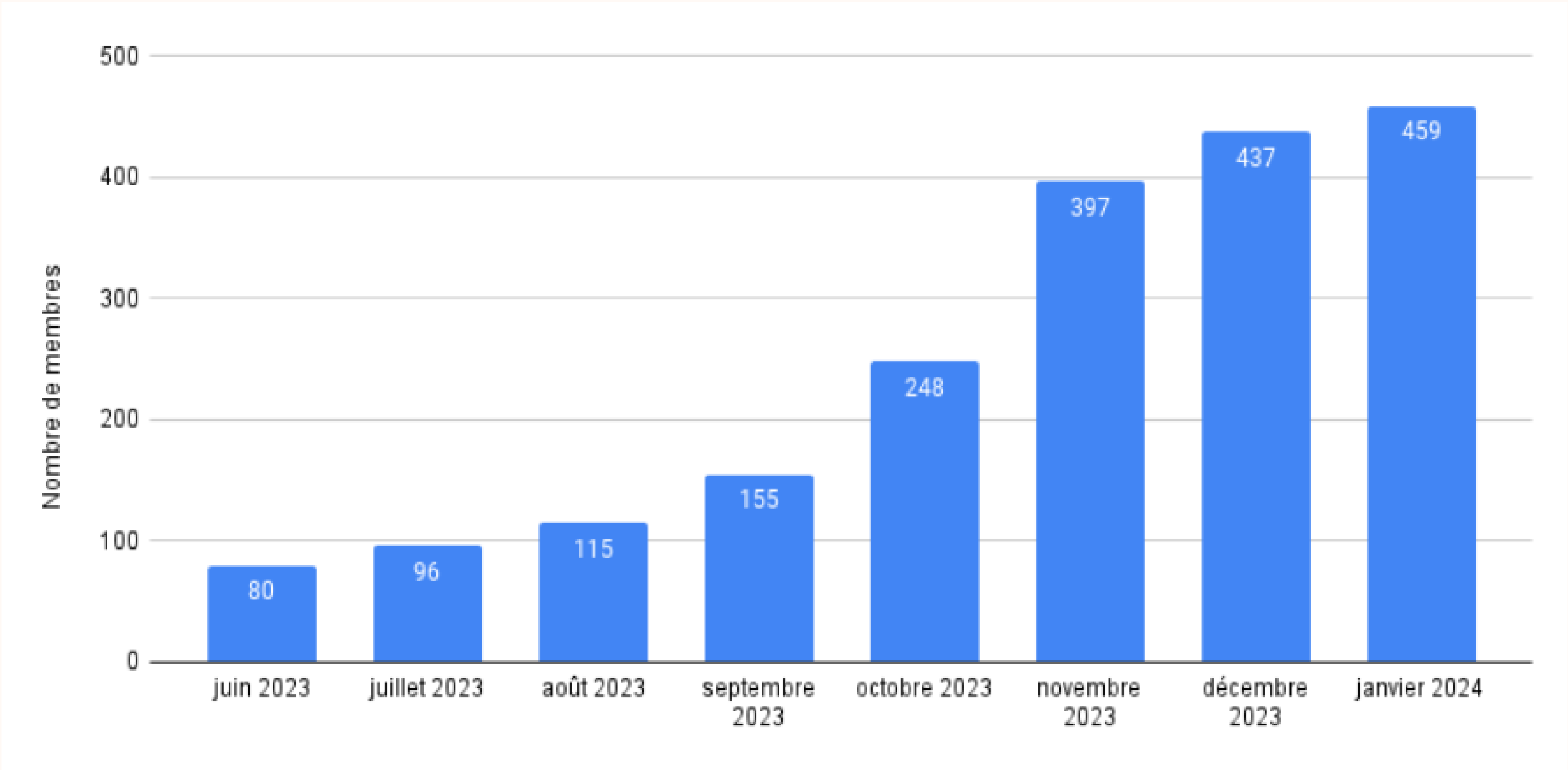
Geographical distribution of LLMs with more than one billion parameters since 2018

OUR MISSION

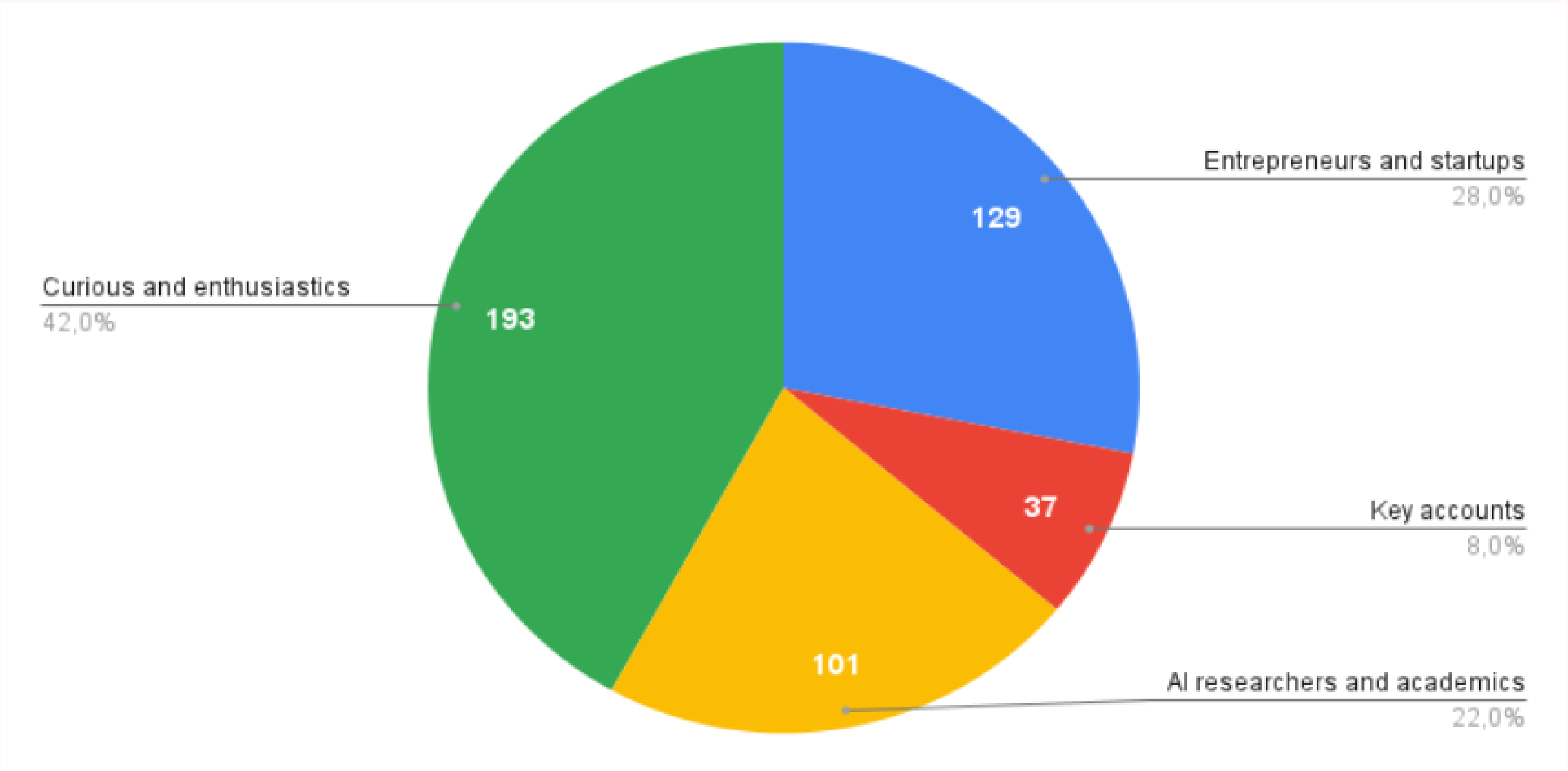
Build inclusive, efficient & ethical LLMs
for low-resource European languages

An end-to-end approach to Open Source:
Data, code, weights

A COMMUNITY OF 500 MEMBERS



A COMMUNITY OF 500 MEMBERS



OUR PARTNERS

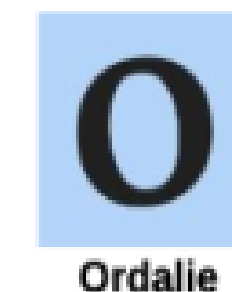
Academics / Public research



Corporates



Giskard

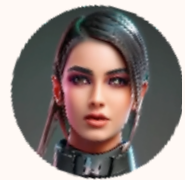


FUNDED BY THE FRENCH GOVERNMENT

DIGITAL COMMONS FOR GENERATIVE AI



OUR MODELS



CLAIRe (2023)

Fine tuning of Falcon 7B

1000H GPU (8xGPU) on
Jean ZAY supercomputer

25 kWh and 1.5kg of CO₂ emission

Data: 138m of conversational data
(drama show, literature and real-life
meetings transcriptions)

Features:

Understand dialogues with diarization
Generation of human-like
conversations (disfluencies,
hesitations...)



LUCIE (2024)

7B 100% Open Source model from scratch

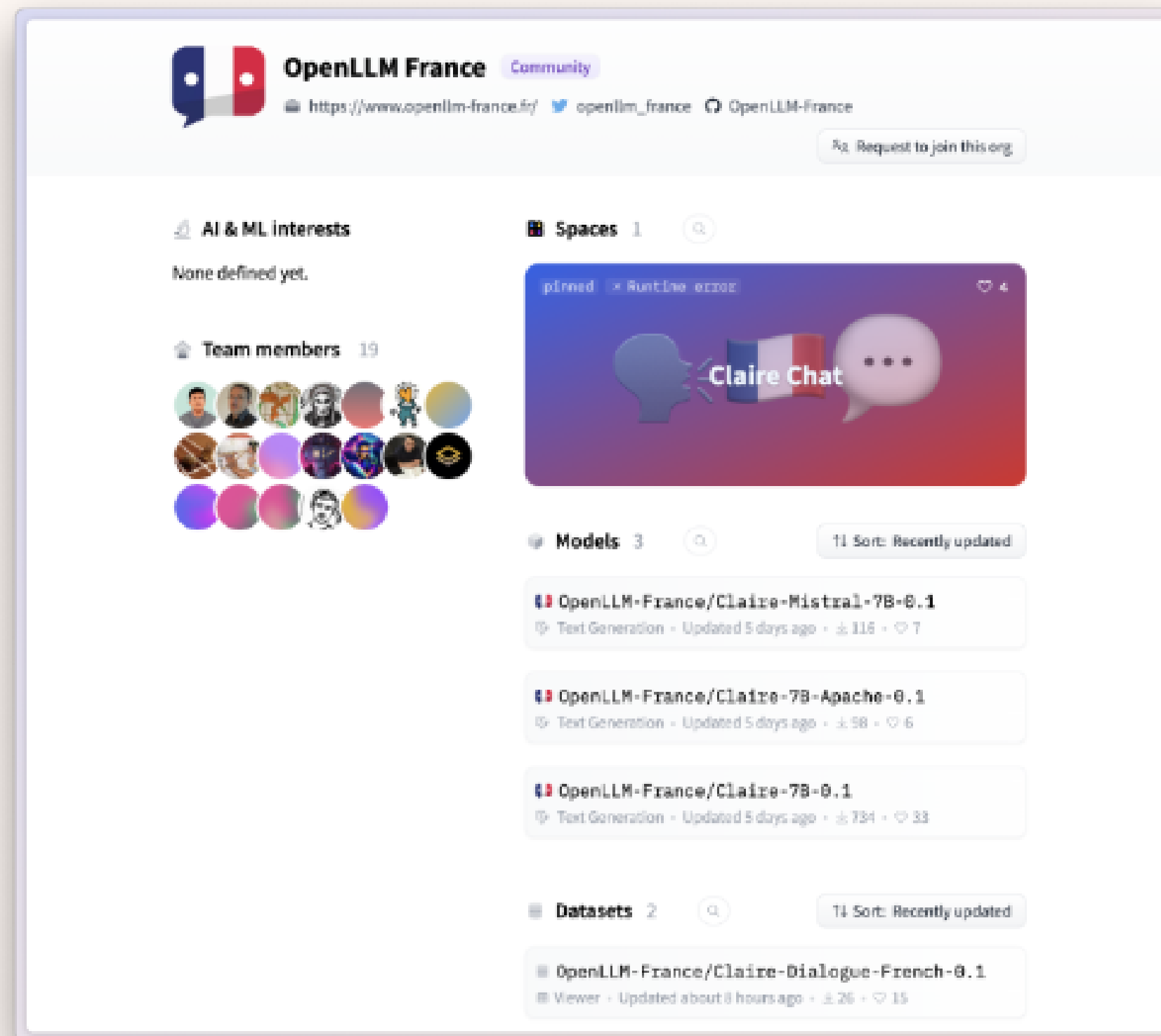
200 000H GPU (96xGPU) on
Jean ZAY supercomputer

Data: 140B of high quality FR data
(Gallica, Hal, Europarl, Wikipedia,
CLAIRe Datasets...), EN (45B, peS20),
GE (3B), ES (3B), Code (180B)

Features:

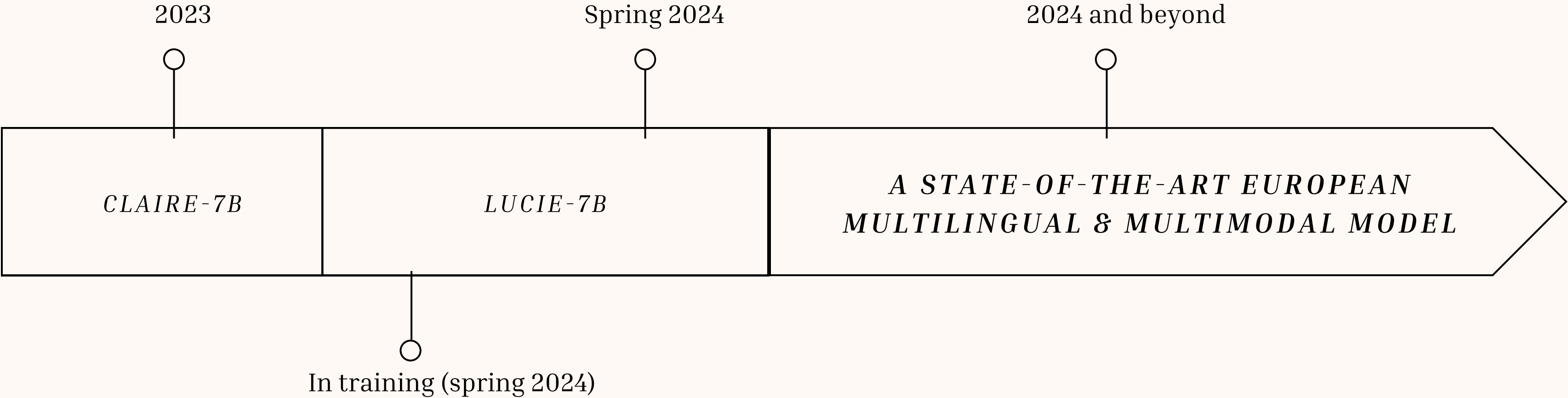
100 % open source datasets
16k context windows
Rotary & sliding windows
Custom tokenizer

OUR MODELS



Our models on Hugging Face

ROADMAP



OUR GOAL: A EUROPEAN MOOSHOT

Build a multilingual LLM embracing Europe's cultural diversity:

- A key asset for **a more competitive Europe**
- A general foundation model available to European SMEs for fine-tuning
- 100 % customizable & auditable (End-to-end Open Source)
- High quality, trustable & ethical training data



AN INDEX OF LOCAL EUROPEAN LLMS

Bulgarian initiatives 🇧🇬 :

- [Insat](#) - Contact: bggpt@insait.ai

Croatian initiatives 🇭🇷 :

- [CroAI](#) - https://www.linkedin.com/posts/croai_large-language-models-have-demonstrated-impressive-activity-7167796231417520128-AIDs/

Czech initiatives 🇨🇪/🇪🇺 :

- [HPLT – High performance languages technologies \(consortium\)](#)
- Czech Republic / European Association 🇨🇪🇪🇺 - <https://cordis.europa.eu/project/id/101070350>

Danish initiatives 🇩🇰 :

- [Danish foundation models](#) - <https://www.linkedin.com/in/saatstrupdan/>

Dutch initiatives 🇳🇱 :

- [Open Future Foundation](#) - Contact: hello@openfuture.eu

[Add your project to the list here!](#)

CONTACT US!



MICHEL-MARIE MAUDET

GENERAL MANAGER @LINAGORA
FOUNDER OF OPEN LLM EUROPE

mmaudet@linagora.com



HADRIEN GAUTROT

EXECUTIVE MANAGER, OPEN LLM EUROPE

hgautrot@linagora.com




[Join OpenLLM on Discord!](#)

hgautrot@linagora.com


Dr. Héctor Allende-Cid, Dr. Najmeh Mousavinezhad

Fraunhofer IAIS, Natural Language Understanding Team

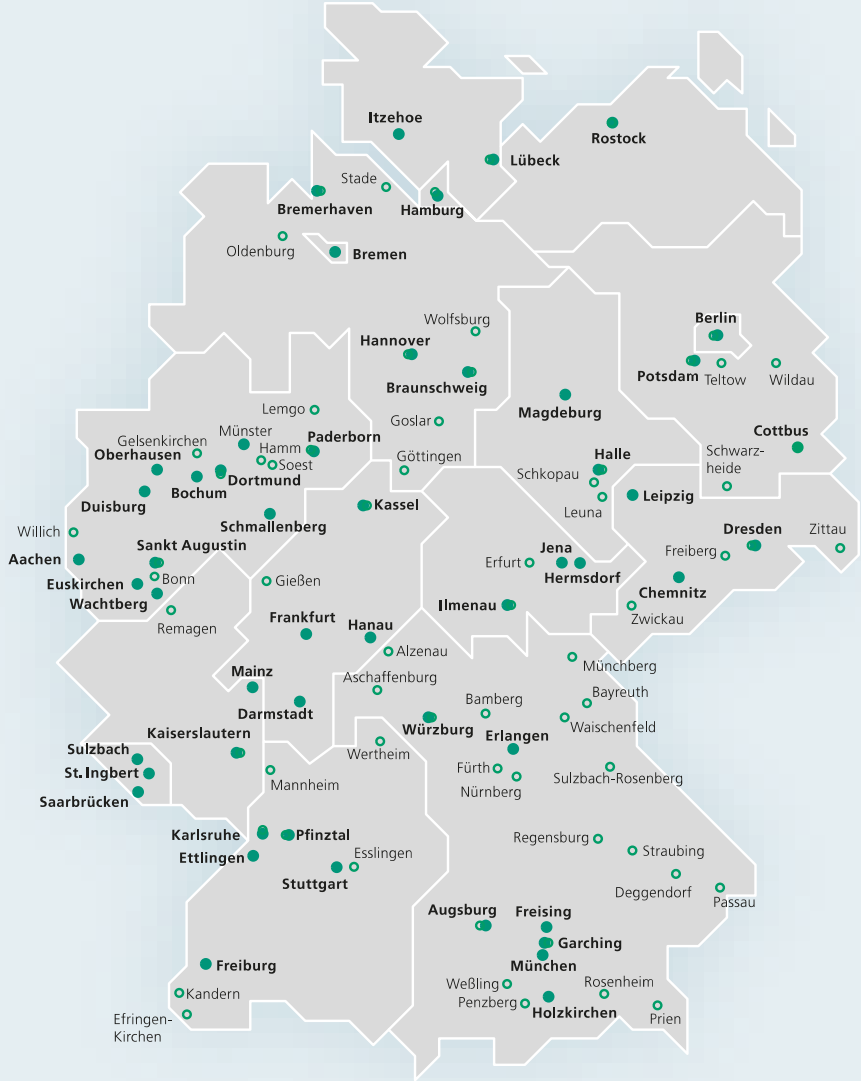
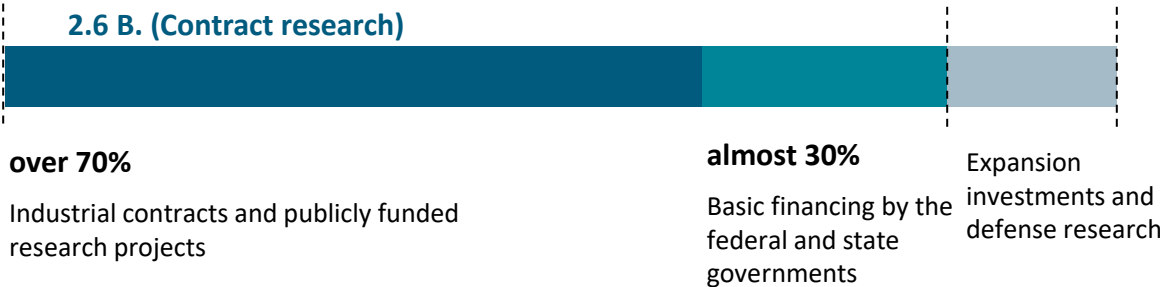
The Fraunhofer-Gesellschaft at a glance

 Application-oriented
Research for the direct benefit of the economy and for the benefit of society

 **30 000+** Employees

 **76** Institutes and research facilities

 **3,0 B. €** Financial volume



Stand: 31.12.2022

Fraunhofer IAIS - Intelligent Systems that Work!

Artificial intelligence, machine learning and big data from Bonn and Dresden



Innovation leader with a comprehensive technology portfolio for customized solutions



Excellence research at the Lamarr Institute and in partnership with the University of Bonn and HBRS

- Natural Language Understanding Group
 - 19 Researchers
 - 2 Senior Researchers, 4 Post-Docs, 6 PhD students, 7 Data Scientists
- Working hand in hand with following groups within IAIS
 - **MLOps**
 - **Health-Care Analytics**
 - **Trustworthiness AI**

350+

Scientists

180+

Research and industry projects

20+

Years experience

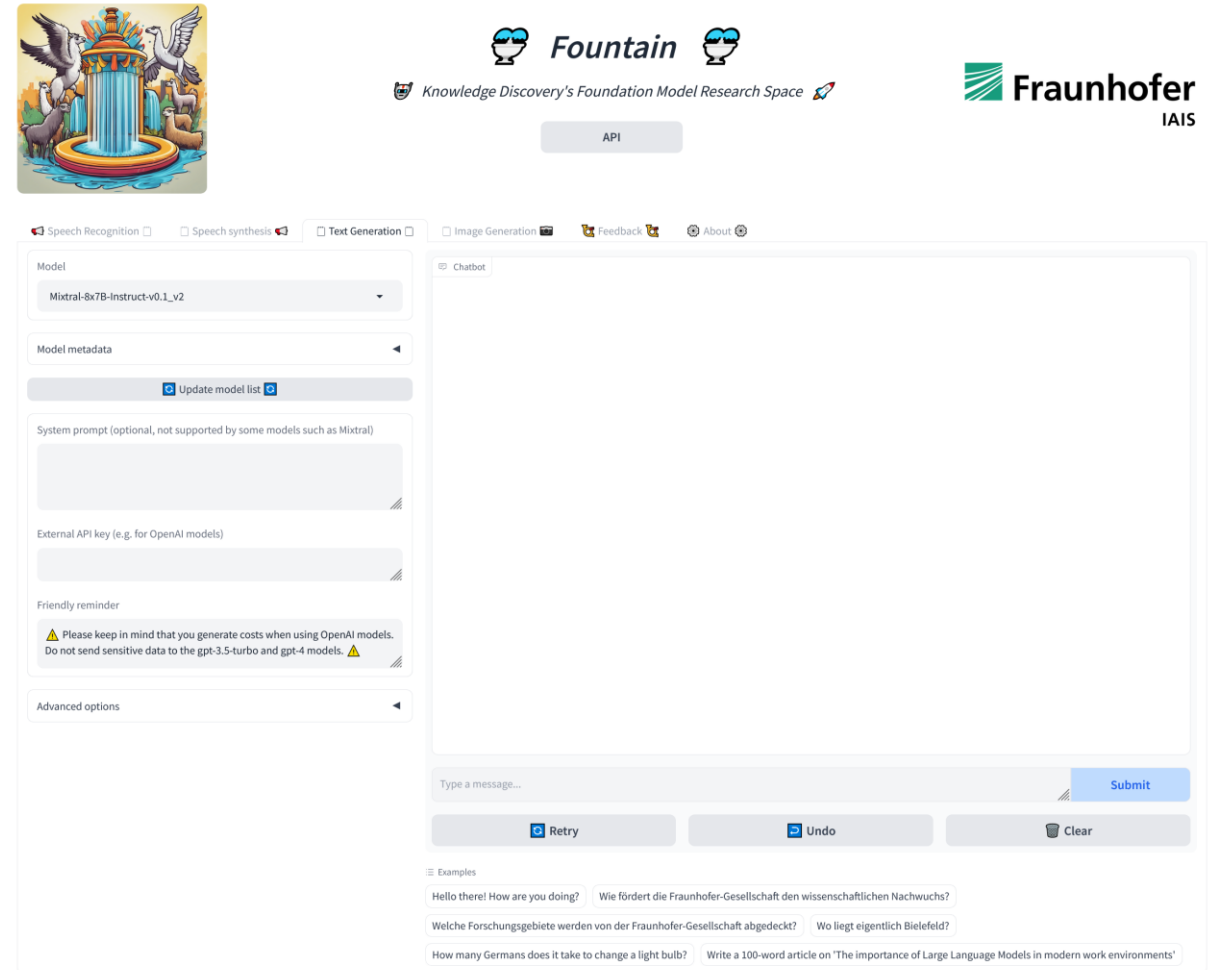


Fountain: Multipurpose AI platform

Platform devoted to perform the following tasks:

- Speech Recognition
- Speech Synthesis
- Text Generation
- Image Generation

Frontend, API server and the model backends are all hosted on servers located inside the IAIS network.



OpenGPT-X: Development of a Gaia-X node for large AI language models and innovative language application services



Creation of large AI language models based on trustworthy integration of company data by Gaia-X

- **Largest German consortium** for the development of GPT models - Fraunhofer is consortium leader
- General Purpose Model vs. **Enterprise Ready Model** – Focus on company requirements
- Building a factual GPT model
- **Digital sovereignty - data remains within the company; privacy preserving**
- Current LLM: 70 billion parameters (focus on German)



The OpenGPT-X project is creating a competitive open source LLM

Status quo of the German lighthouse project



- Development of large **enterprise-ready language models** in compliance with European values since Jan 2022
- By the end of the year: Model with 70 billion parameters and 1.4 trillion tokens
- Languages: English, German, French, Spanish, Italian and other European languages
- Differentiation from the competition:
 - Focus on European languages, especially German
- Optimized model performance - thus optimized computing power
- Open-source
 - Open access for industry and research
 - Data sovereignty through on-premise use
 - Industry-ready applications
 - Fact fidelity



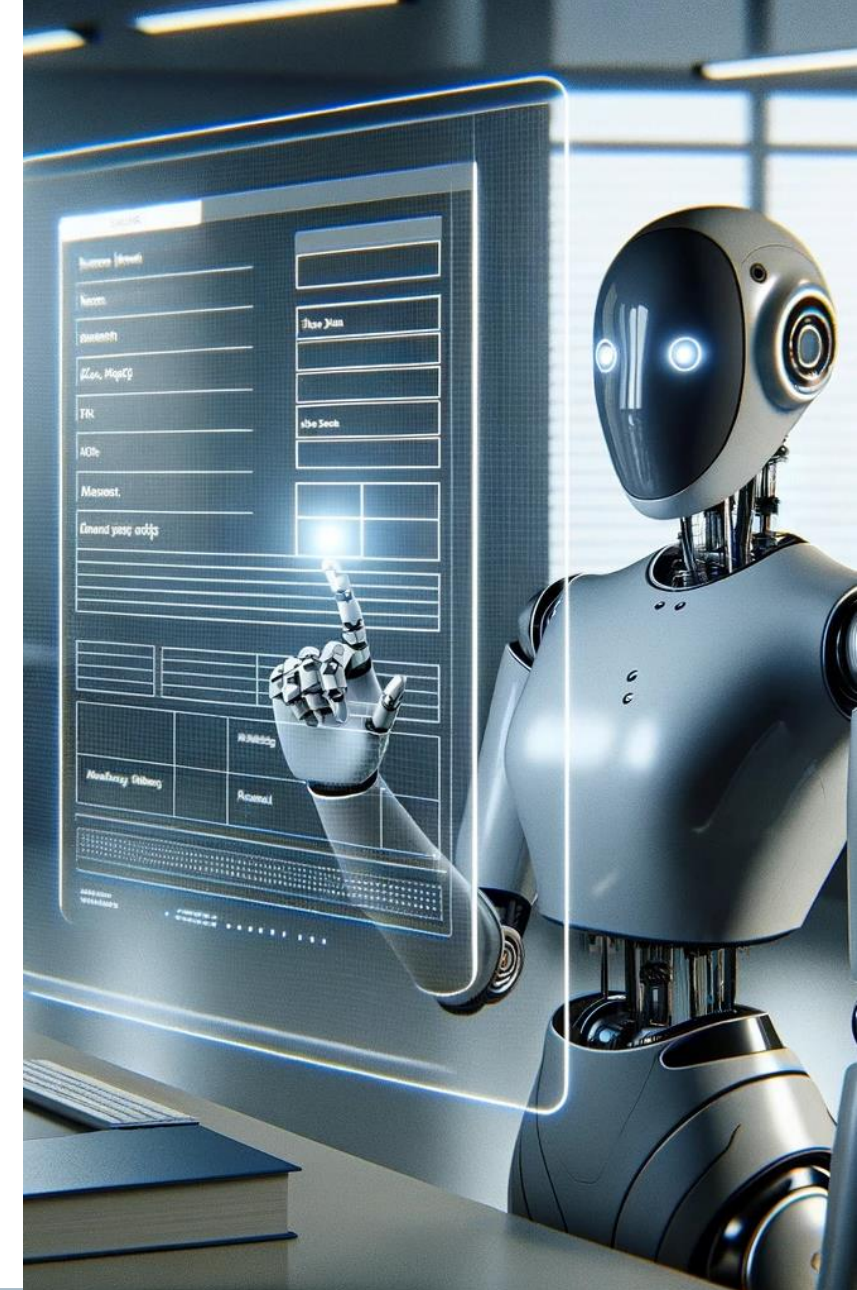
Open Call – Project Proposal

Status quo of the German lighthouse project

- Why are we want to apply to this fund?:
 - Contribution to the Open-Source Community
 - Work with underrepresented languages
 - Focus on specific domains: Industries, Organizations, etc.

Looking for:

- Partners from Research centers, Universities and Companies, that have experience in NLP, Foundation Models and MLOps.
- Continue to enrich our knowledge while working hand in hand with other institutions.



Contact

Dr. Héctor Allende-Cid, Dr. Najmeh Mousavinezhad
Senior Researchers, Natural Language Understanding Group
Tel. +49 2241 14-2220
hector.allende-cid@iais.fraunhofer.de;
najmehsadat.mousavinezhad@iais.fraunhofer.de

Fraunhofer-Institut für Intelligente Analyse-
und Informationssysteme IAIS
Schloss Birlinghoven 1
53757 Sankt Augustin

www.iais.fraunhofer.de