# Introduction to the Language Data Space: Context and Overview

## The European Language Data Space (LDS)

## Workshop in Poland

*Mateusz SZOTURMA*

*Legal and Policy Officer*

*European Commission – DG CONNECT – G3*

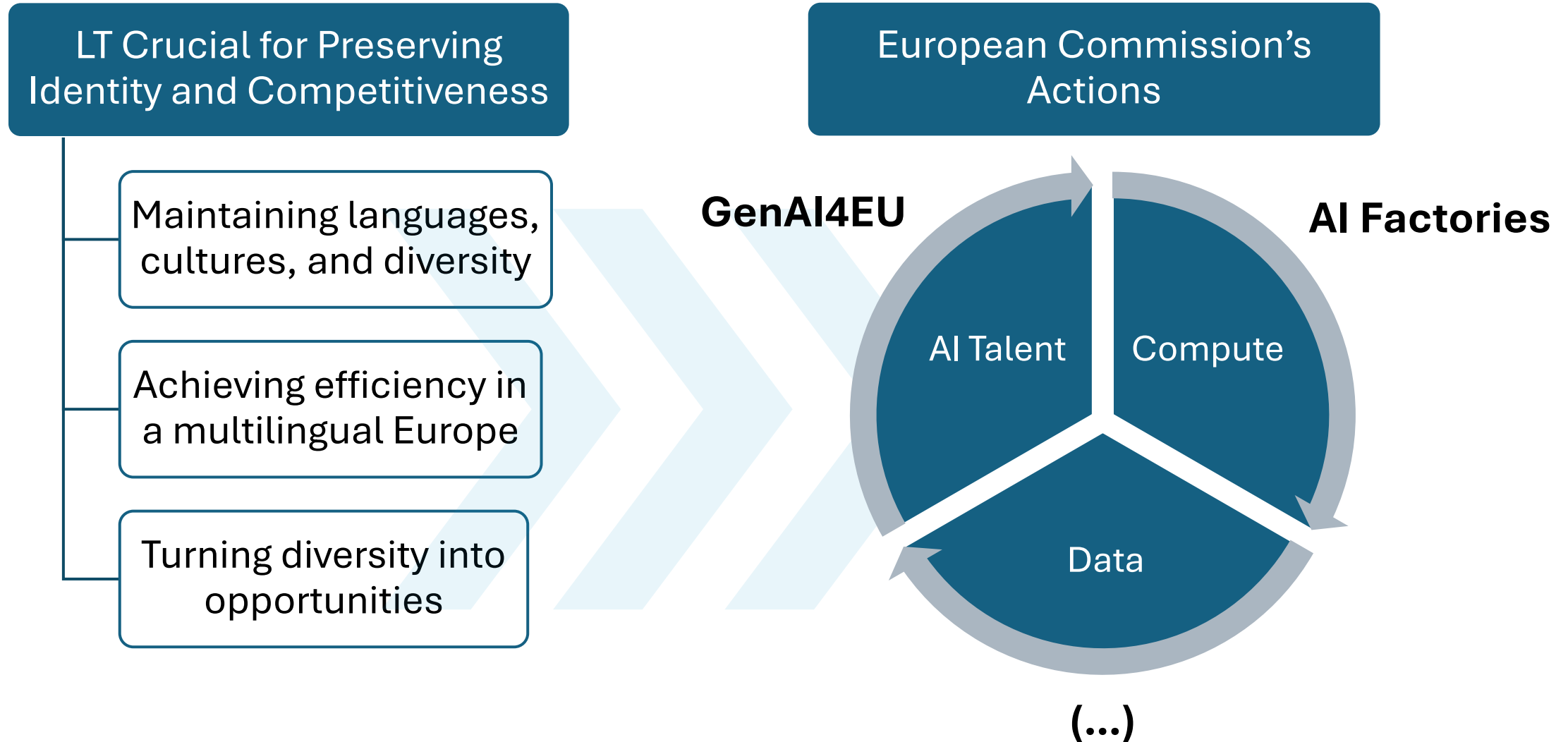# Landscape – the importance of language technologies

**LT Crucial for Preserving Identity and Competitiveness**
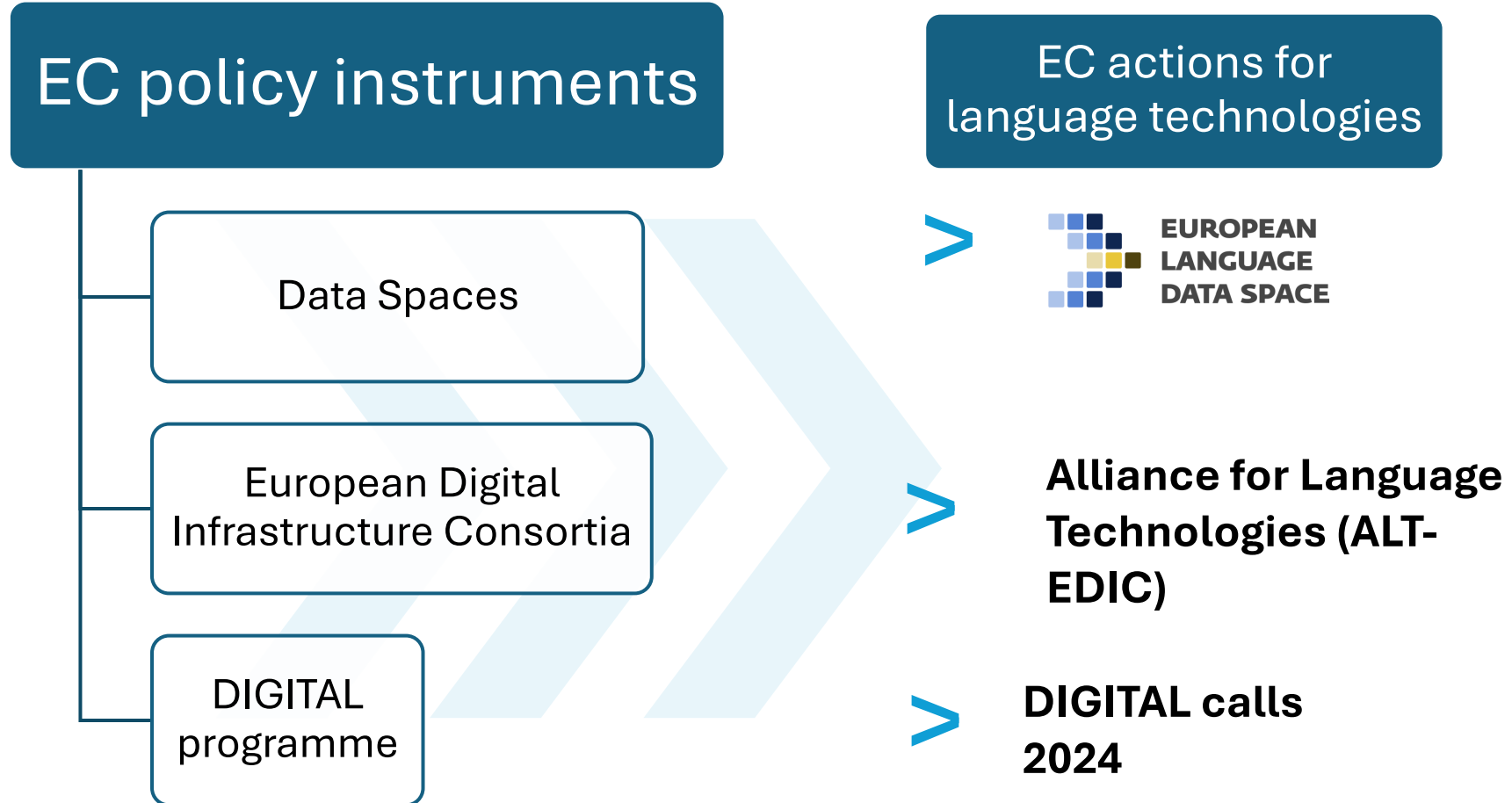
- Maintaining languages, cultures, and diversity
- Achieving efficiency in a multilingual Europe
- Turning diversity into opportunities

# Landscape – the importance of data for language technologies

**LT Crucial for Preserving Identity and Competitiveness**

- Maintaining languages, cultures, and diversity
- Achieving efficiency in a multilingual Europe
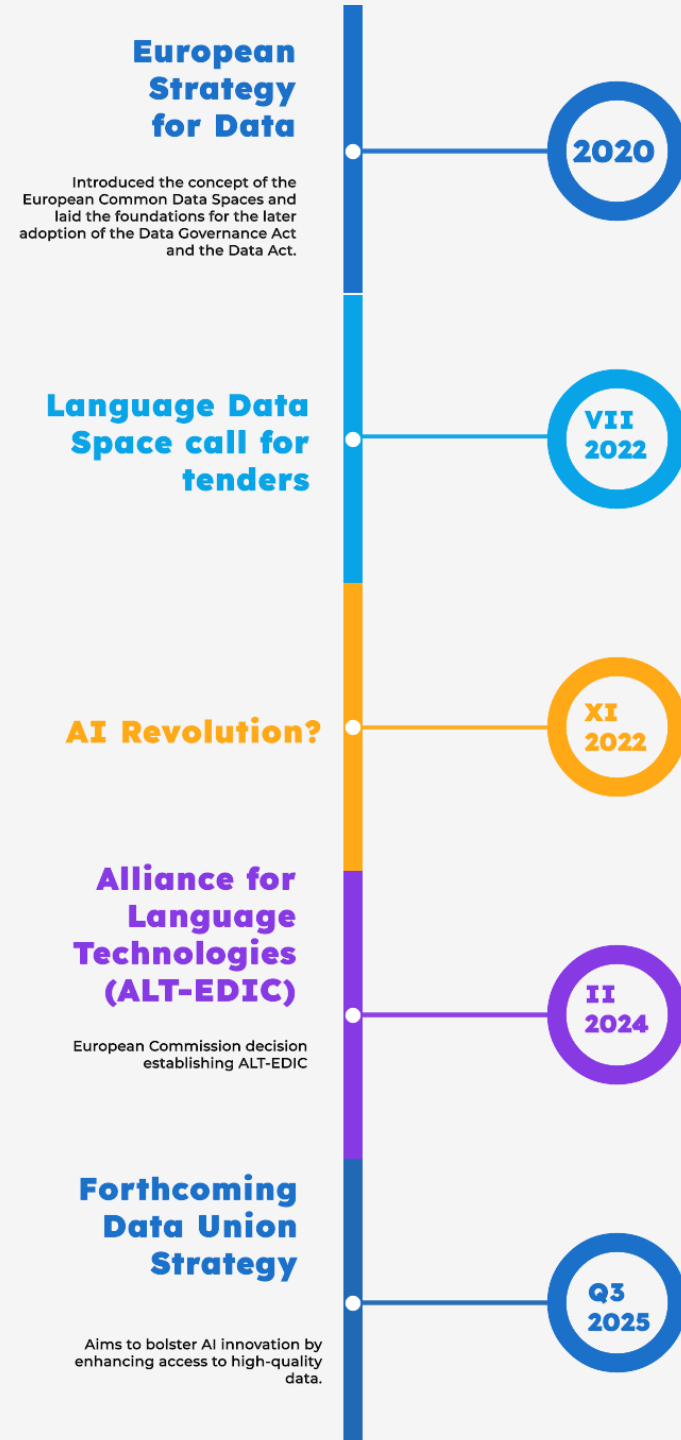- Turning diversity into opportunities

**GenAI4EU**

**European Commission's Actions**

**AI Factories**

- AI Talent
- Compute
- Data

(...)

# Landscape – EC actions for language data and technologies

**EC policy instruments**

- Data Spaces
- European Digital Infrastructure Consortia
- DIGITAL programme

**EC actions for language technologies**

> EUROPEAN LANGUAGE DATA SPACE

> **Alliance for Language Technologies (ALT-EDIC)**

> **DIGITAL calls 2024**

# Language Data Space in the context of EC data policy development

## European Strategy for Data

Introduced the concept of the European Common Data Spaces and laid the foundations for the later adoption of the Data Governance Act and the Data Act.

**2020**

## Language Data Space call for tenders

**VII 2022**

## AI Revolution?

**XI 2022**

## Alliance for Language Technologies (ALT-EDIC)

European Commission decision establishing ALT-EDIC

**II 2024**

## Forthcoming Data Union Strategy

Aims to bolster AI innovation by enhancing access to high-quality data.

**Q3 2025**

# Language Data Space – Overview

**EUROPEAN LANGUAGE DATA SPACE**

## What

**Platform** and **marketplace** for the collection, creation, sharing and re-use of multilingual and multimodal language data

6 Mio € **Procurement** Contract from the **DIGITAL** Work Programme 2021-2022

## Why

Increase **availability** of clean, high-quality, compliant language data

**Support** the development of state-of-the-art European language technologies

## Who

**Consortium**: DFKI, ELDA, TILDE, Athena Institute

**Governance**: Member States representatives and Industry User group

**Stakeholders**: Industry, Public Administration, Research, etc.

## How

- Technical Architecture and Infrastructure        - Openness

- Governance Framework        - Respect of EU Rules and Values
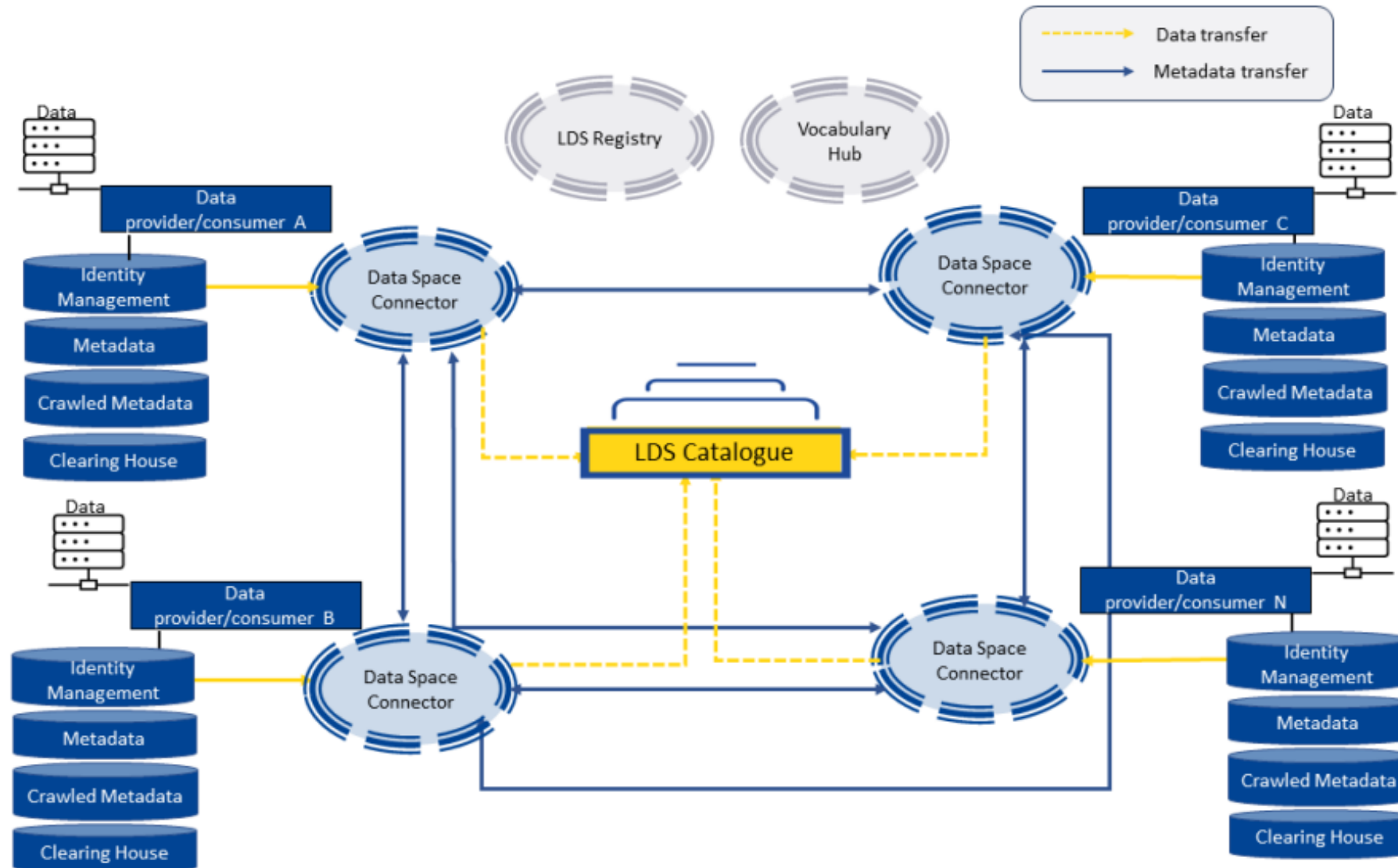
- Promotion        - Overcoming Legal Barriers (IPR, GDPR, etc.)

## When

Started in **January 2023** for Three years – until end 2025 (Possible contract renewal for another year).

Beta version in **Q1 2025**

# Language Data Space – Architecture

# alt-edic

**Alliance for Language Technologies European Digital Infrastructure Consortium**

## Objectives

- **Preserve linguistic and cultural diversity in Europe**
- **Technological leadership and strategic autonomy**
- **Respect European rules and values**
- **Cooperation**
- **Raising awareness**

## Participation

### Members [18]

**FR**, BG, CZ, DK, ES, FI, GR, HR, HU, IE, IT, LT, LV, LU, NL, PL, SI, Flanders

### Observers [8]

AT, BE, CY, EE, MT, PT, RO, SK

### Interested [2]

DE, SE

## Budget

| Projects (Costs) | |
|---|---|
| - LLMs4EU | € 40 Mio |
| - ALT-EDIC4EU | € 4 Mio |
| - OpenEuroLLM | € 40 Mio |
| - LLM-BRIDGE | € 6,8 Mio |

## Actions Plan

**Federate MS and EU language data resources**

Seed fund for new language models

Fine-tuning existing language models

Ease HPC access

Support SMEs take up

**European Ecosystem**

# Thank you

![CC BY license logo]