EUROPEAN LANGUAGE EQUALITY

META-NET
META-FORUM 2023

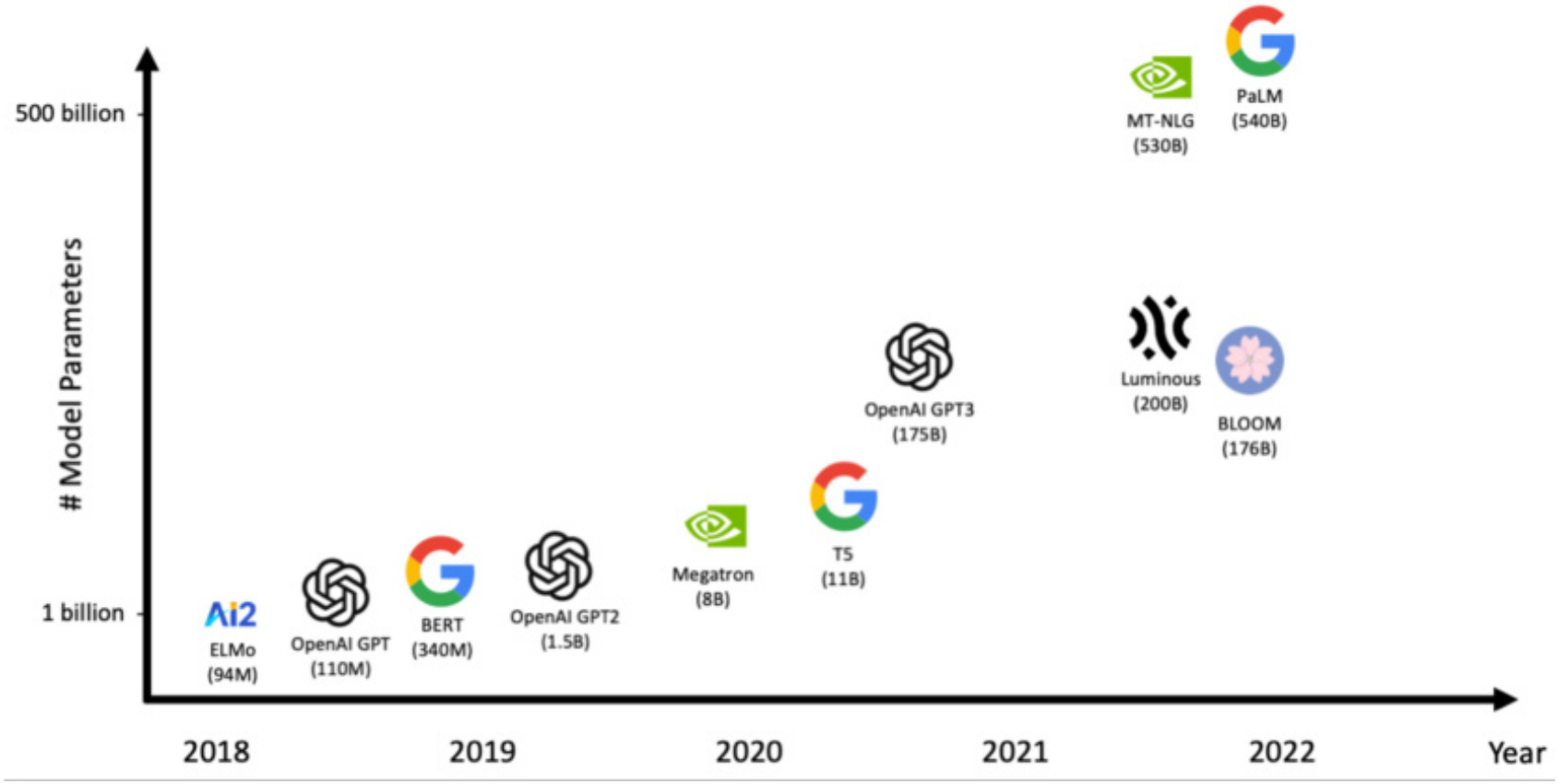
Developing Multilingual LLMs

Pedro Ortiz Suarez (DFKI, Germany)
pedro.ortiz@dfki.de

27-06-2023 META-FORUM 2023 – Digital Language Equality for a Multilingual Europe
http://european-language-equality.eu

Large Language Models

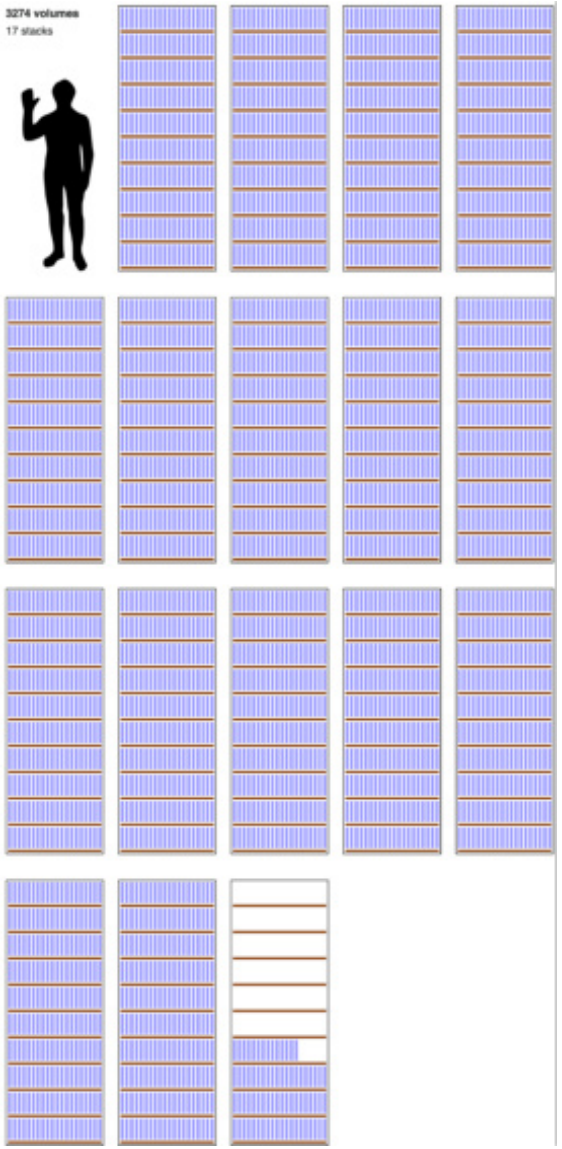
Language models are getting ever larger
(more data, more parameters, more compute).



Parameters	Words
400 Millions	8.0 Billion
1 Billion	20.2 Billion
10 Billion	205.1 Billion
67 Billion	1.5 Trillion
175 Billion	3.7 Trillion
280 Billion	5.9 Trillion
520 Billion	11.0 Trillion
1 Trillion	21.2 Trillion
10 Trillion	216.2 Trillion

Pre-training LLMs on Large and Diverse Datasets

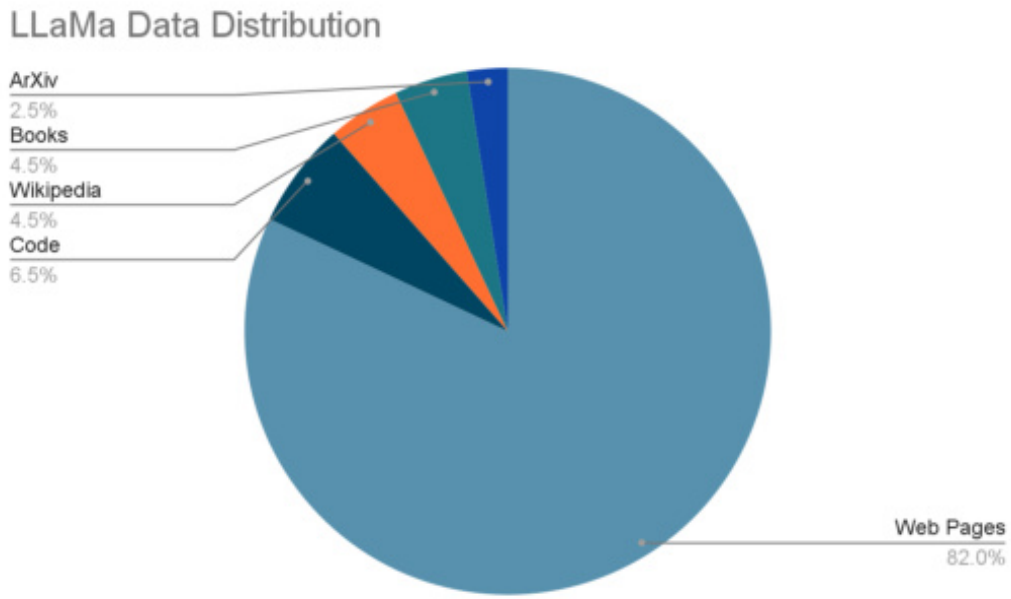
- Training with language modelling objective (next token prediction) on large amounts of text data (+1 trillion tokens, Terabyte to Petabyte scale).
- In general the size of the dataset will always be orders of magnitude bigger than the number of parameters.
- The more you scale the models, the more pre-training data you will need.
- The training data is so large, that we don’t have a way of manually auditing it.
- Pre-training is expensive: Dataset should be **diverse** and **balanced** to produce a general-purpose model.



English Wikipedia: 4,362,297,570 words (21.23 GB)

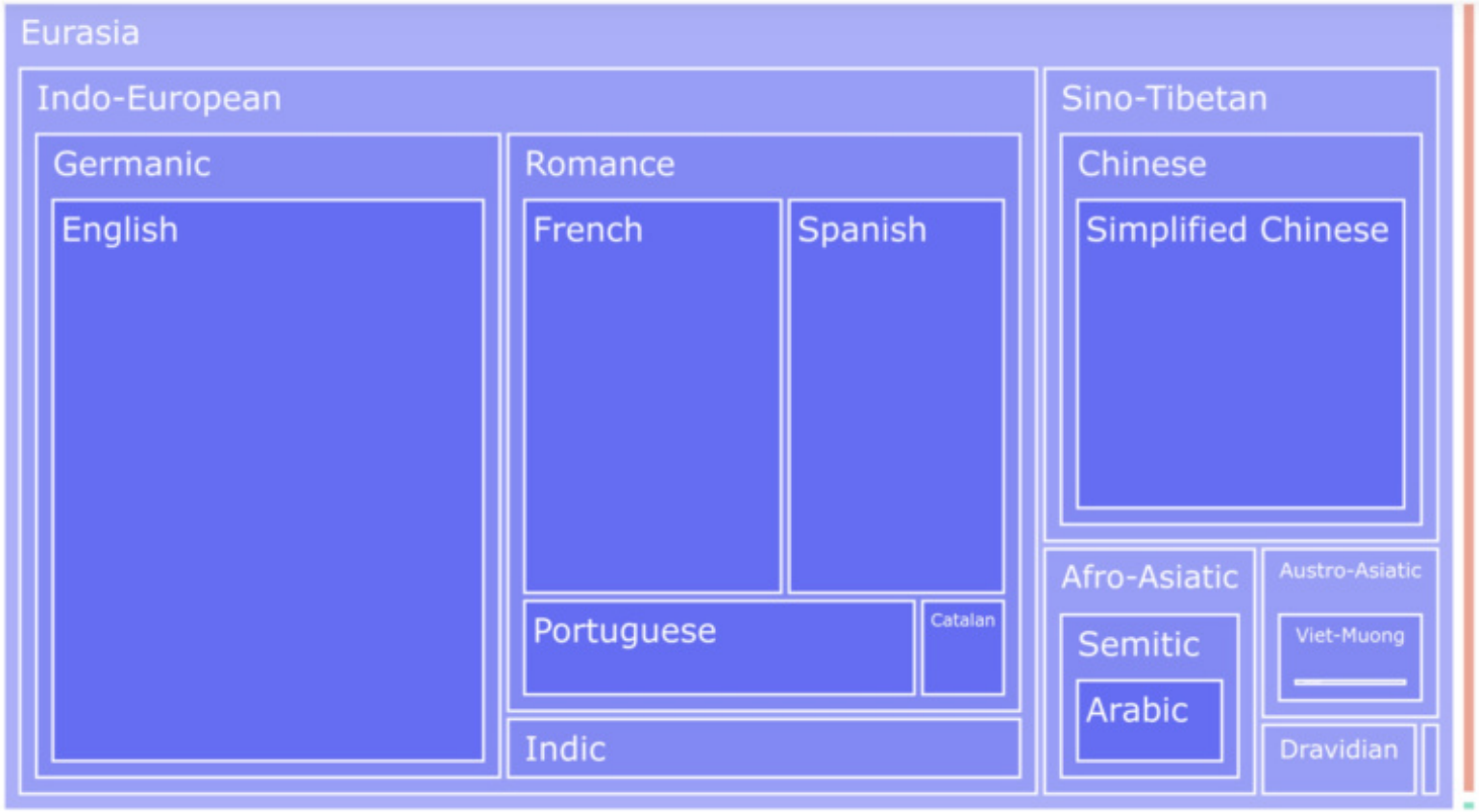
Data Distribution for LLMs

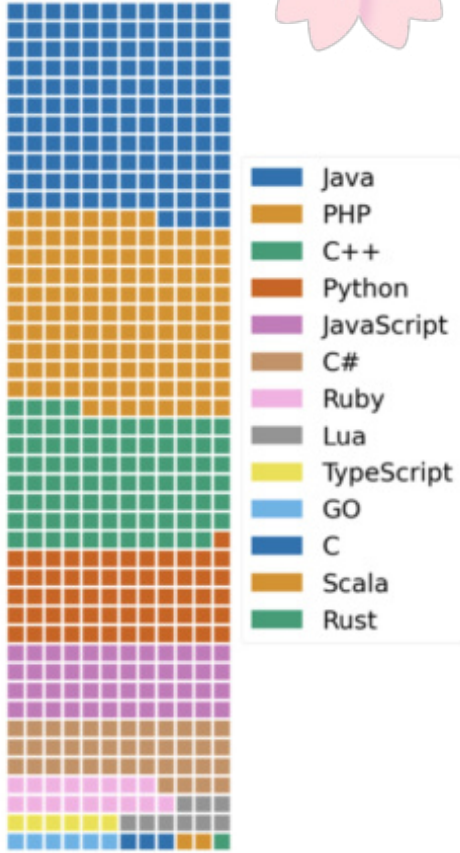
- In practice datasets are not balanced
- Web data is always the cheapest and easiest to get
- Web data is diverse, but definitely not balanced or representative of all the language range.
- Web data always contains unwanted content (fiction, bias, propaganda).
- Programming language code (source code) is becoming ubiquitous in the data mix
- Public domain books and encyclopaedic data is common, but availability varies greatly between languages.



LLaMA Model (Facebook)	
Data Source	Proportion
Web Pages	82%
Code	6.5%
Wikipedia	4.5%
Books	4.5%
Scientific Articles	2.5%

Is Balancing Multilingual Data Even Possible?





Data mix of BLOOM, a multilingual dataset trained by BigScience. The diagram shows the disparities in data availability between languages.

Taken from: The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset, Laurençon et al. 2019. CC BY 4.0

OpenGPT-X: Open European Large Language Models

openGPT-X

Gefördert durch:

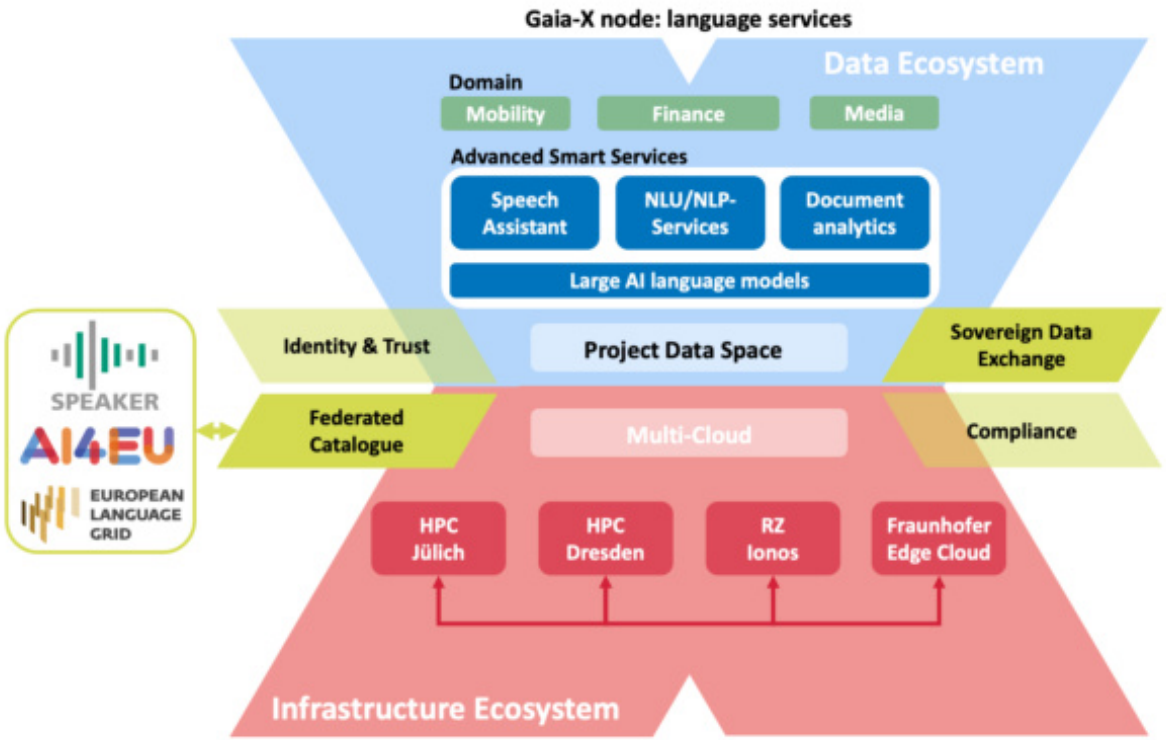
Bundesministerium für Wirtschaft und Klimaschutz

aufgrund eines Beschlusses des Deutschen Bundestages

- Large-scale language models are a key technology.
  - Yet, the most successful AI language models come from the USA and China.
  - These are often not fully available to the free market and available only in English and Chinese.
  - Rapidly growing importance of LLMs calls for Europe to:
    - Ensure technology and data independence.
    - Innovation and competitiveness for Europe.
- OpenGPT-X builds and trains open and European large language models.
  - According to the highest European data protection standards.
  - Fosters innovation and strengthen Europe’s ability to compete with LLMs Made in Europe.
  - OpenGPT-X is funded by the German Federal Ministry of Economics and Climate Protection (BMWK) from January 2022 to December 2024.

OpenGPT-X Consortium

- **Data Ecosystem & Federated Services**
  - Sovereign Data Exchange: An exchange of large data sets (Gaia-X data ecosystem) for the training of large AI language models
  - Federated Catalogue: interoperable catalogue for AI language services
- **Infrastructure Ecosystem HPC Multi-Cloud**
  - Usage of JUWELS-Booster HPC system (FZ Jülich) using
  - 3700 A100-GPUs
  - Utilizing the HPC center of TU-Dresden (ScaDS.AI) with 460 GPUs
  - GPU infrastructure partner IONOS / IPCEI-Initiative, Fraunhofer Edge Cloud



Cross-lingual & Progressive Transfer Learning

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

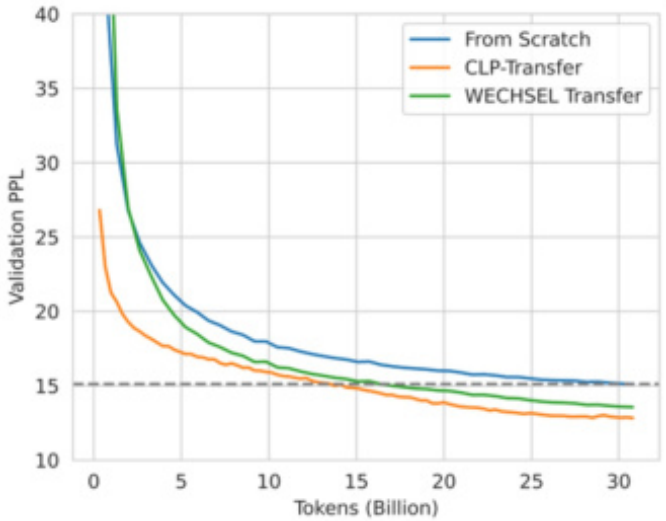
- The "**Open**" in OpenGPT-X stands also for accessibility in terms of data and compute requirements for training LLMs.
- More LLMs are made publicly available (BLOOM, OPT, ...) that we can exploit to train our own LLM **more resource-efficiently**.
- Goal: Train a *large* model in a *target* language (e.g., a large German model).
- **CLP transfer learning.** Instead of training a model from scratch with randomly initialized weights, we recycle weights from pretrained models:
  - Cross-lingual: Transfer a *large* model in a *source* language (e.g., English) to our *target* language.
  - Progressive: Transfer a *small* model in our *target* language to the *large* model size (can be trained with fewer resources or is publicly available).

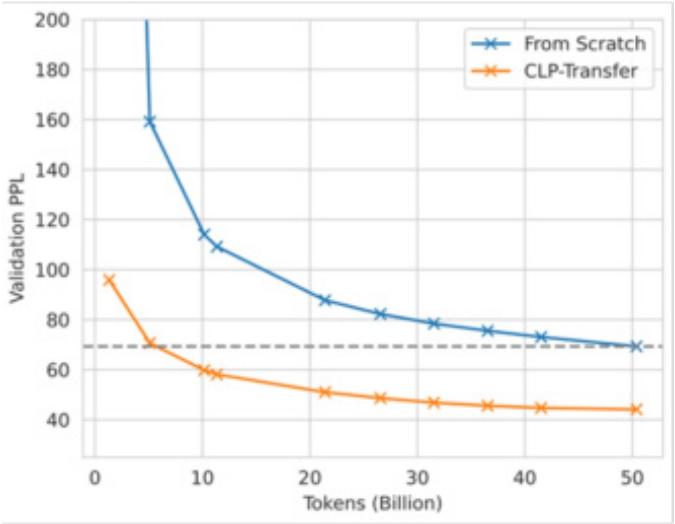
Taken from: M. Ostendorff and G. Rehm. "Efficient Language Model Training through Cross-Lingual and Progressive Transfer Learning." In G. A. Tadesse, E. Bekele, W. Saib, L. Oala, and A. Alaagib (eds.): Practical Machine Learning for Developing Countries Workshop (PML4DC@ICLR 2023), 05 May 2023

Cross-lingual & Progressive Transfer Learning

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

- We train two monolingual German language with CLP:
  - GPT2-XL (1.5B parameters, English)
  - BLOOM (7.1B parameters, multilingual – no German)
- CLP outperforms sole cross-lingual transfer (WECHSEL) and reduces the training effort compared to from scratch by up to **80% for BLOOM** (50% for GPT2-XL).



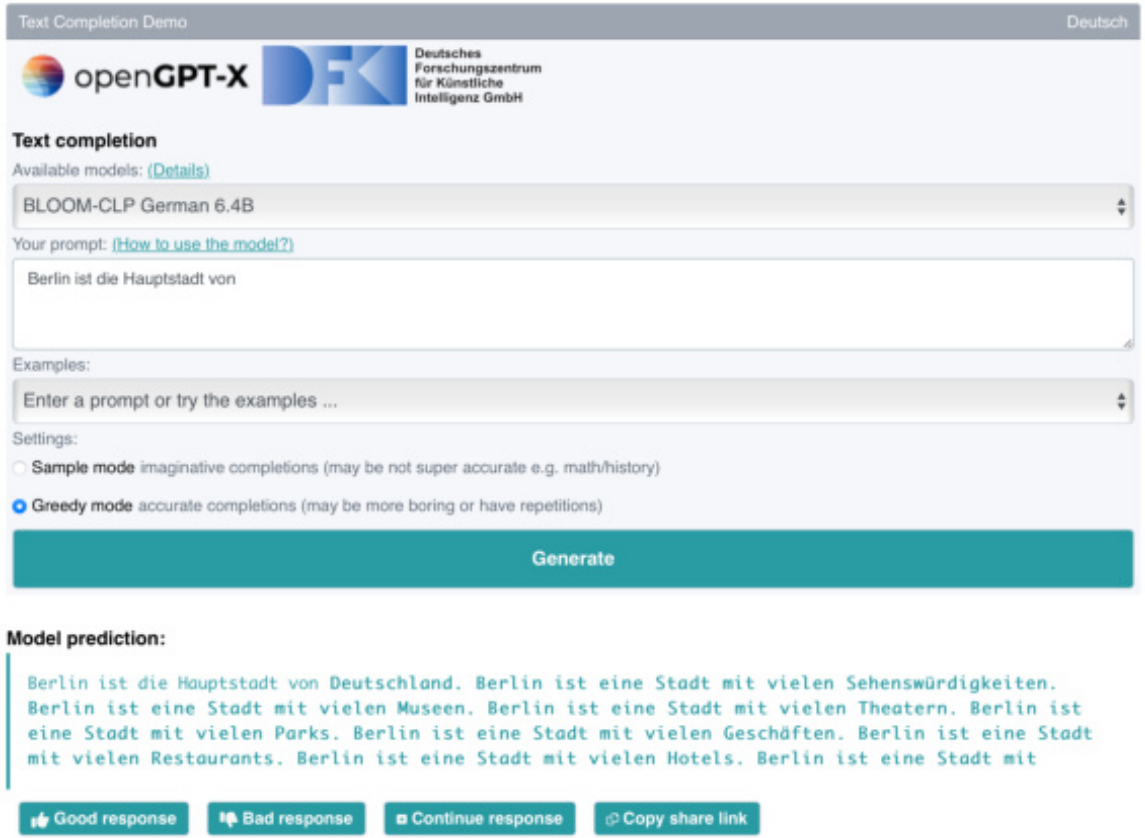


Taken from: M. Ostendorff and G. Rehm. "Efficient Language Model Training through Cross-Lingual and Progressive Transfer Learning." In G. A. Tadesse, E. Bekele, W. Saib, L. Oala, and A. Alaagib (eds.): Practical Machine Learning for Developing Countries Workshop (PML4DC@ICLR 2023), 05 May 2023

First German OpenGPT-X Models

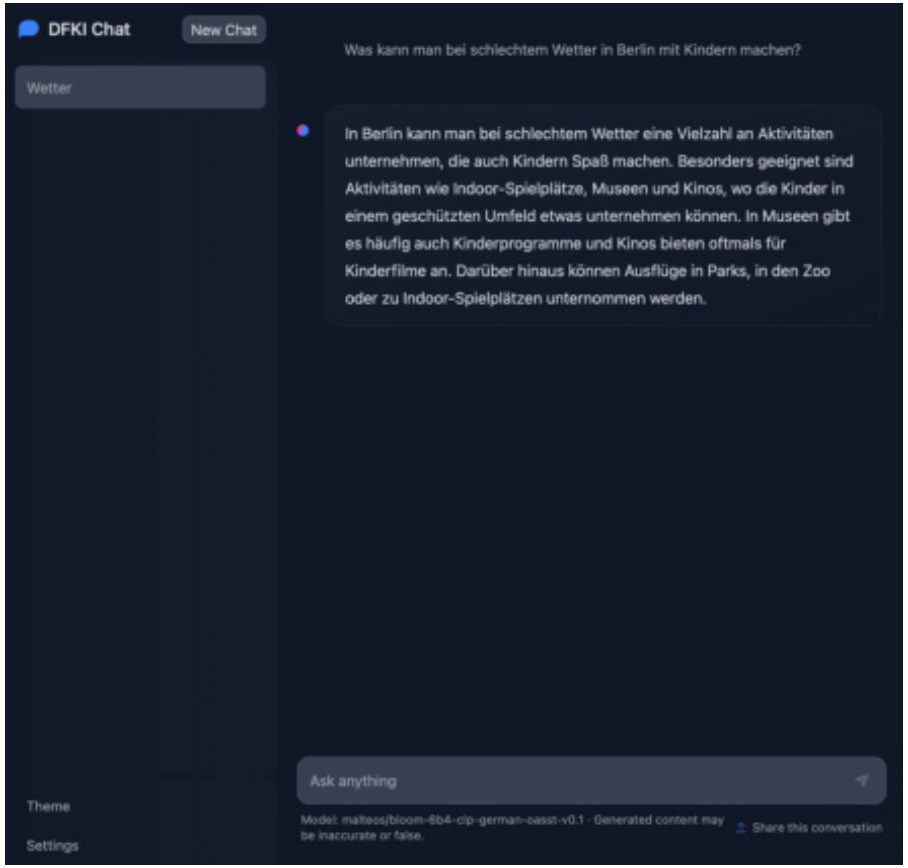
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Base German Model



https://opengptx.dfki.de

German Chat Model

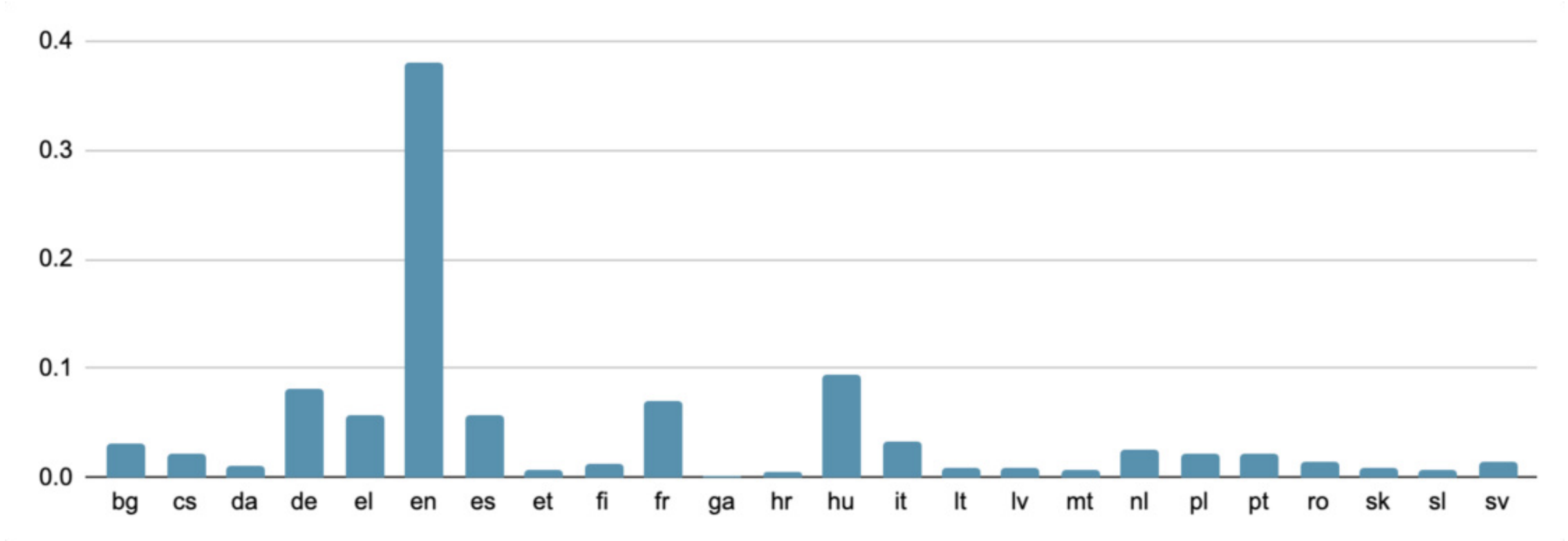


https://opengptx.dfki.de/chat/

Web-crawls: OSCAR

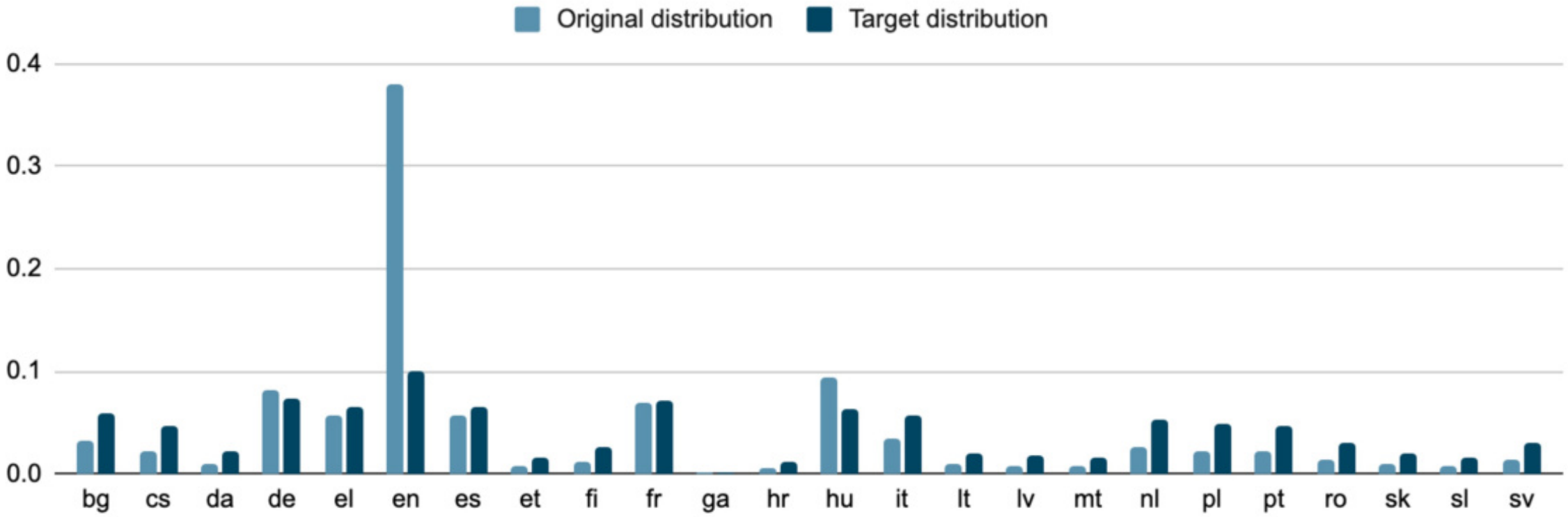
- The majority of the training data will be Web-crawled data.
- Generally lower quality but the only data source available in large quantities.
- Target size: 800B tokens (80% of all data)
- Approach: OSCAR Corpus – https://oscar-corpus.com
  - Preprocessed and annotated version of CommonCrawl
  - Language identification
  - Removal of low quality content
    (adult content, noise, tiny documents)

OSCAR: Language distribution



Available data by language based on OSCAR v23.01

Targeted language distribution



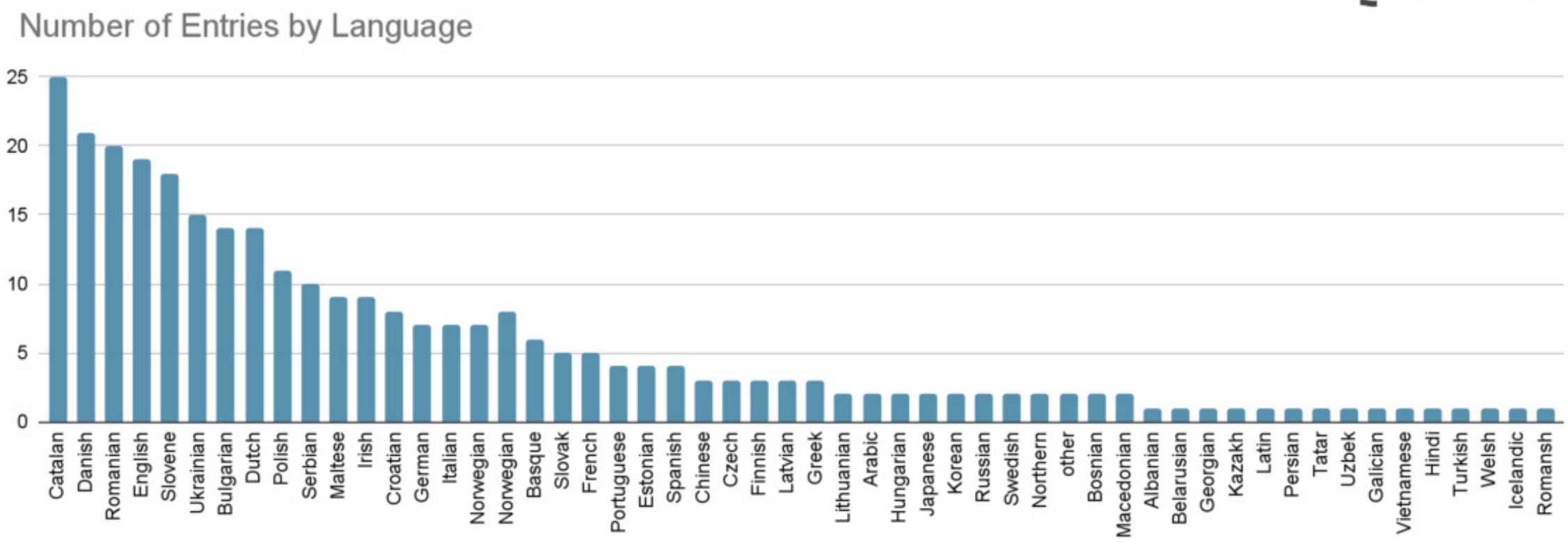
Target distribution vs. available data by language based on OSCAR v2301

Target distribution should be much more balanced!

Curated datasets

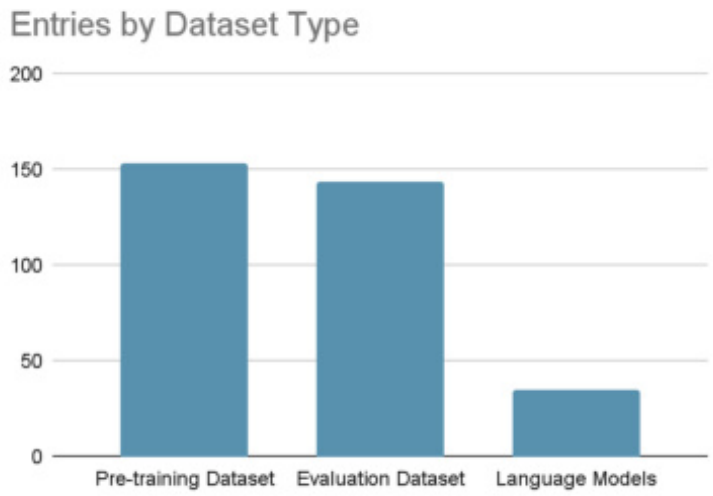
- **High quality data is crucial** for model performance **but only exists in small quantities**, especially for languages other than English.
- **Curated datasets**: Manually curated collections of presumably high quality content (content type and language is known, e.g., legal documents from EUR-Lex).
- Target size: **100B tokens** (10% of all data)
- **Language distribution should be balanced** – but most likely heavily skewed towards English due to lack of curated datasets in other languages.
- **Community effort**:
  - Native speakers know the best data sources for their languages.
  - Recently started effort: ELE initiative → 42 unique contributors so far

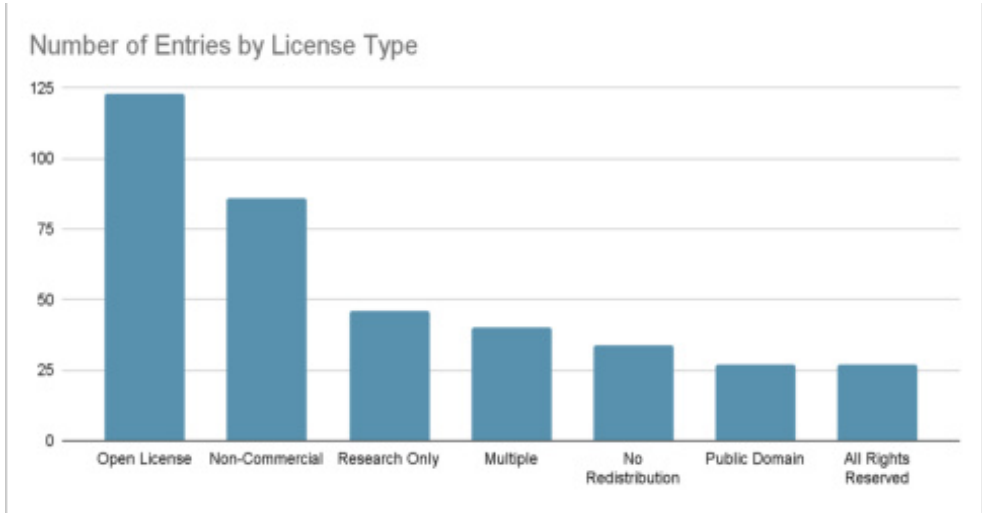
ELE – OpenGPT-X Catalogue

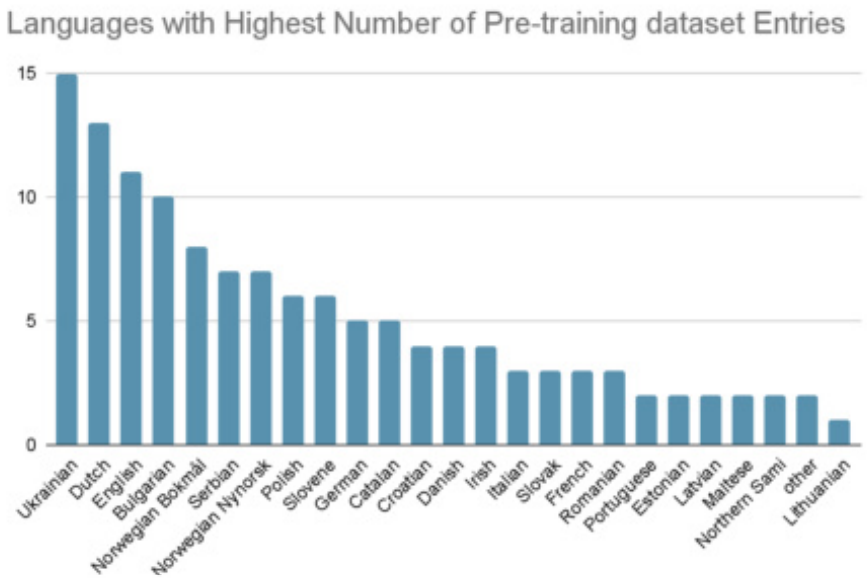


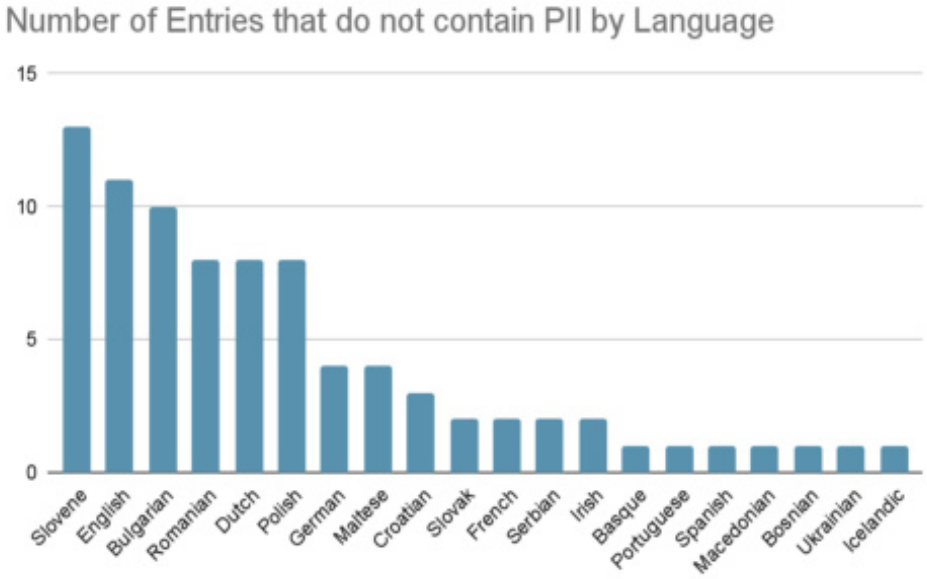
Number of entries in the ELE – OpenGPT-X catalogue by language

ELE – OpenGPT-X Catalogue





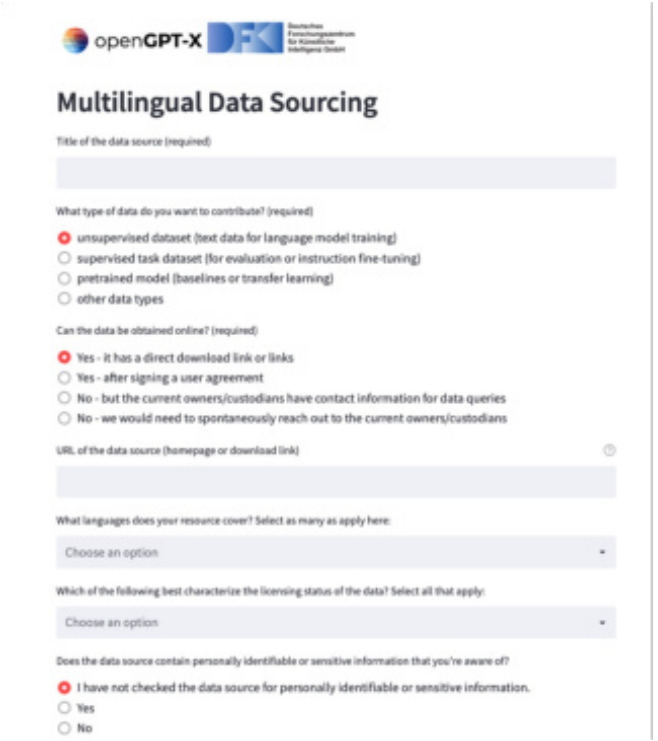




Contributions Welcome!

openGPT-X

EUROPEAN LANGUAGE EQUALITY



openGPT-X DFKI

Multilingual Data Sourcing

Title of the data source (required)

What type of data do you want to contribute? (required)

- unsupervised dataset (text data for language model training)
- supervised task dataset (for evaluation or instruction fine-tuning)
- pretrained model (baselines or transfer learning)
- other data types

Can the data be obtained online? (required)

- Yes - it has a direct download link or links
- Yes - after signing a user agreement
- No - but the current owners/custodians have contact information for data queries
- No - we would need to spontaneously reach out to the current owners/custodians

URL of the data source (homepage or download link)

What languages does your resource cover? Select as many as apply here:

Choose an option

Which of the following best characterize the licensing status of the data? Select all that apply:

Choose an option

Does the data source contain personally identifiable or sensitive information that you're aware of?

- I have not checked the data source for personally identifiable or sensitive information.
- Yes
- No

https://opengptx.dfki.de/data-sourcing

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Thank you!

The European Language Equality project has received funding from the European Union under grant agreements № LC-01641480 – 101018166 (ELE) and № LC-01884166 – 101075356 (ELE2).

Pedro Ortiz Suarez (DFKI, Germany)
pedro.ortiz@dfki.de

27-06-2023 META-FORUM 2023 – Digital Language Equality for a Multilingual Europe
http://european-language-equality.eu