From OpenLLM France 🇫🇷 to OpenLLM Europe 🇪🇺

Paving the way to sovereign and open source AI

FOSDEM – Brussels - February 4, 2024

WHO AM I - OUR MISSION WITH LINAGORA SINCE 2000

Michel-Marie MAUDET
General Manager

✉ mmaudet@linagora.com
☎ +33 6 60 46 98 52

#GoodTechForGood
fantastic-enterprise.com

We invent and develop
« Good Tech For Good »

« Ethical, responsible, sustainable and Open Source technologies to make the world a better place, with maximum positive impact on people, society and the planet. »

AMBITION OF OPENLLM FRANCE 🇫🇷

1

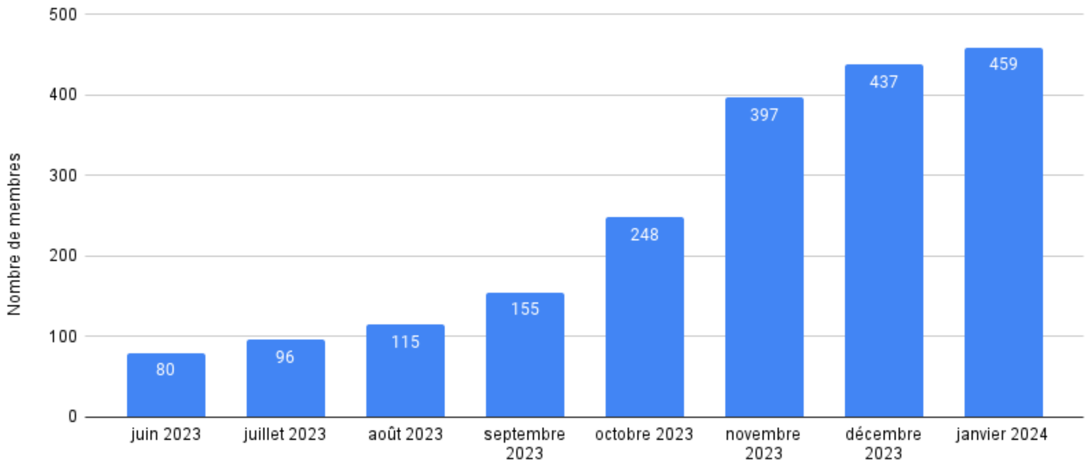
Build trusted, sovereign and **REAL Open Source** AI models & technologies

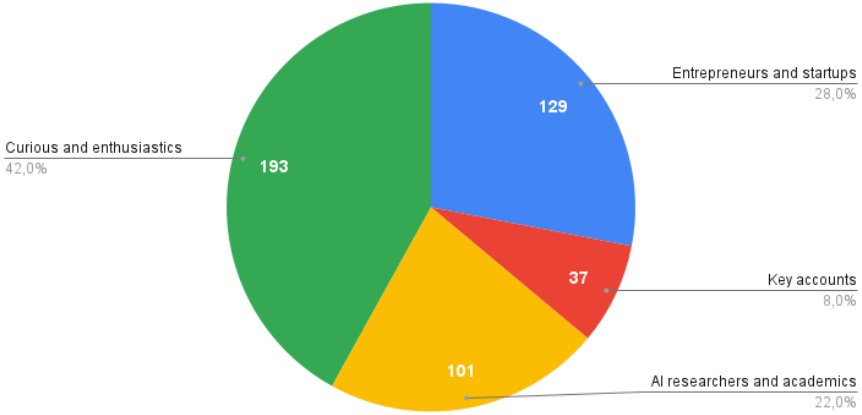
2

Build an **open and transparent collaborative ecosystem** around LLMs and generative AI

https://www.openllm-france.fr/

WE'VE ALREADY COMPLETED OUR SECOND OBJECTIVE





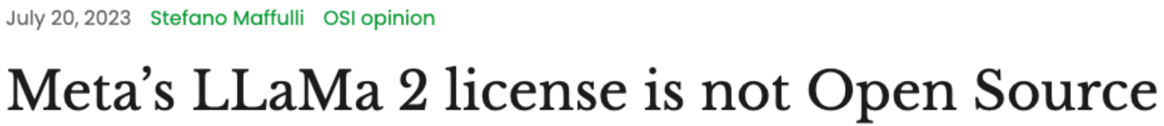
Academics / Public research

GENCI
cnrs IDRIS
WIKIMEDIA FRANCE
Loria
list ceatech
ALMAnaCH Équipe-projet Inria
LIX
Class'Code
UNIVERSITÉ PARIS 1 PANTHÉON SORBONNE
LE VOICE LAB
HUB FRANCE IA

Corporates

LINAGORA.AI
opsci
THALES Together • Safer • Everywhere
FARADAY
TALKR.ai
3DS OUTSCALE
Crédit Mutuel ARKEA
SNCF CONNECT & TECH
Giskard
Ordalie

OPEN WEIGHTS ≠ OPEN SOURCE



July 20, 2023 Stefano Maffulli OSI opinion

Meta’s LLaMa 2 license is not Open Source

OSI’s Open Source Definition

- free redistribution
- source code availability
- derivatives allowed
- no limitations of who may use it or for what
- no additional license in place
- license must not depend on distribution format, technology, presence of other works

open source initiative®

OUR CODE OF CONDUCT → CONTRIBUTIONS TO OSI AI DEFINITION

1 Open Source code

→ *Source code of the code model, pre training tool released under OSI compliant Open Source Licence*

2 Open Model

→ *Licence and user agreement without any restrictions on who may use it and for what*

3 **Open Corpus**

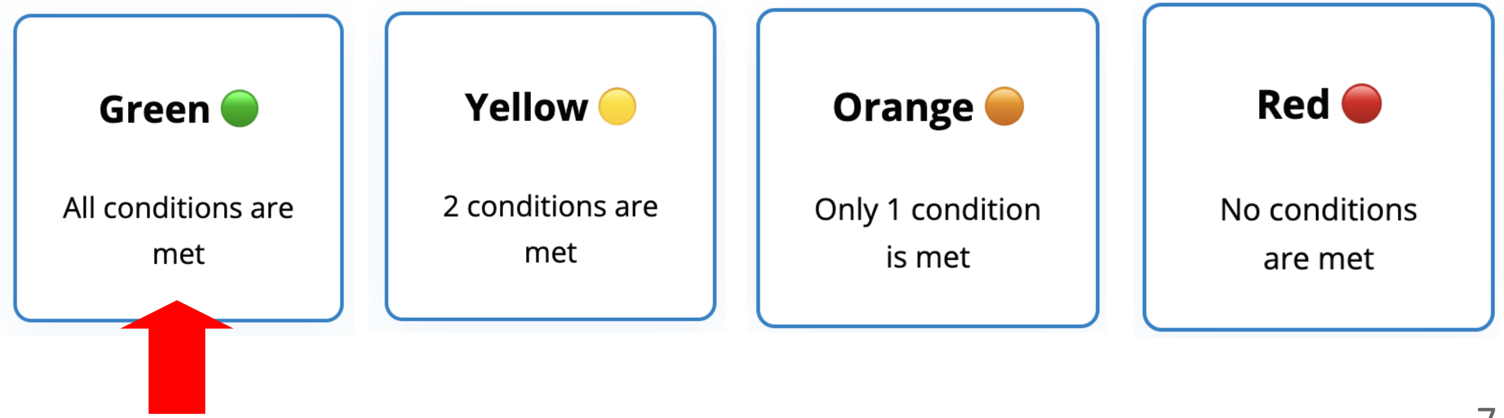
→ *100% of training data must be publicly available in a format that allows for investigation of the model's biases (preferences) and for retraining*

https://github.com/OpenLLM-France/OpenSourceAI-Definition

FULLY COMPLIANT WITH NEXTCLOUD ETHICAL AI RATING

1. Open Source code
2. Open Model
3. **Open Corpus**

Nextcloud
Ethical AI



Green 🟢

All conditions are met

Yellow 🟡

2 conditions are met

Orange 🟠

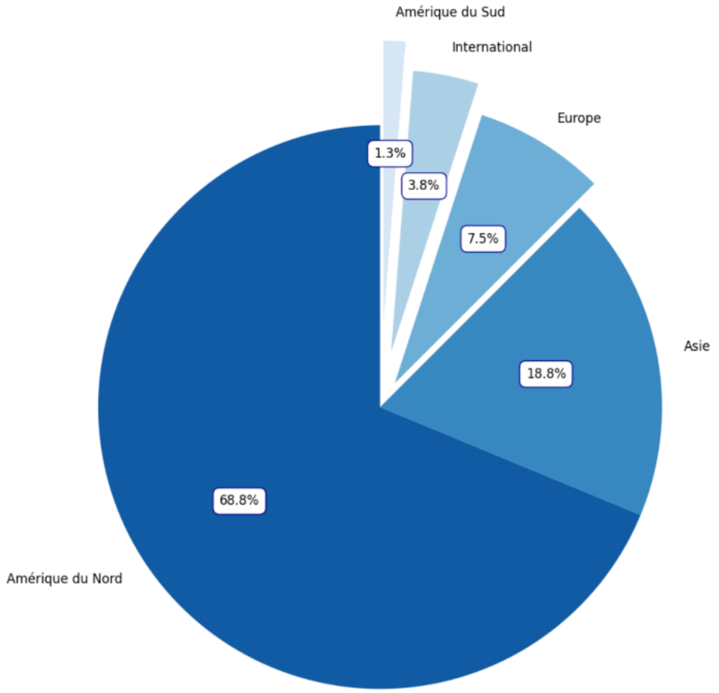
Only 1 condition is met

Red 🔴

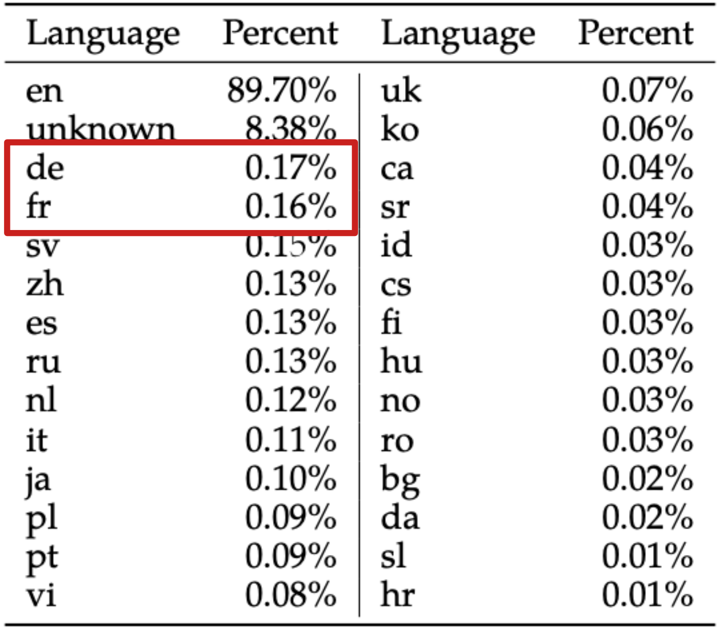
No conditions are met

CREATING AN LLM COMPLIANT WITH OUR CULTURE AND VALUES

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



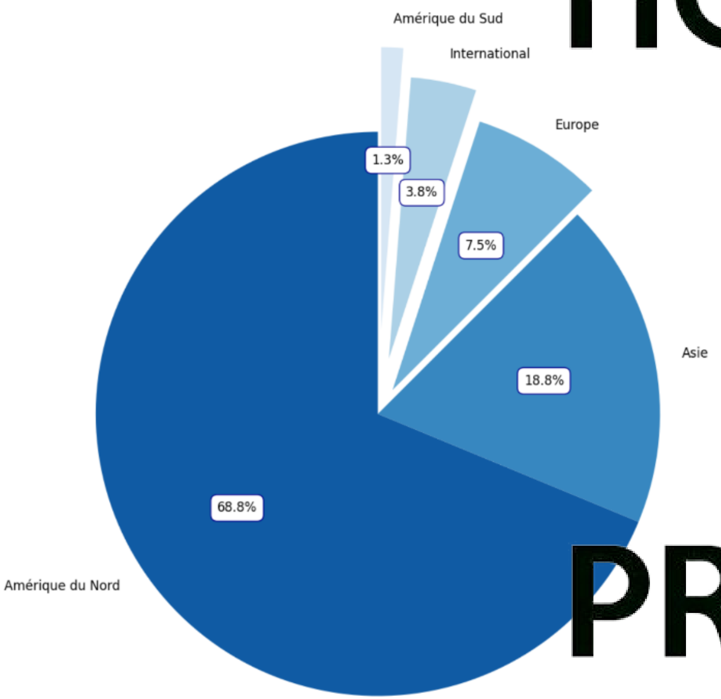
LLAMA V2 : Language distribution in pretraining data with percentage



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%

CREATING AN LLM COMPLIANT WITH OUR CULTURE AND VALUES

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



HOUSTON

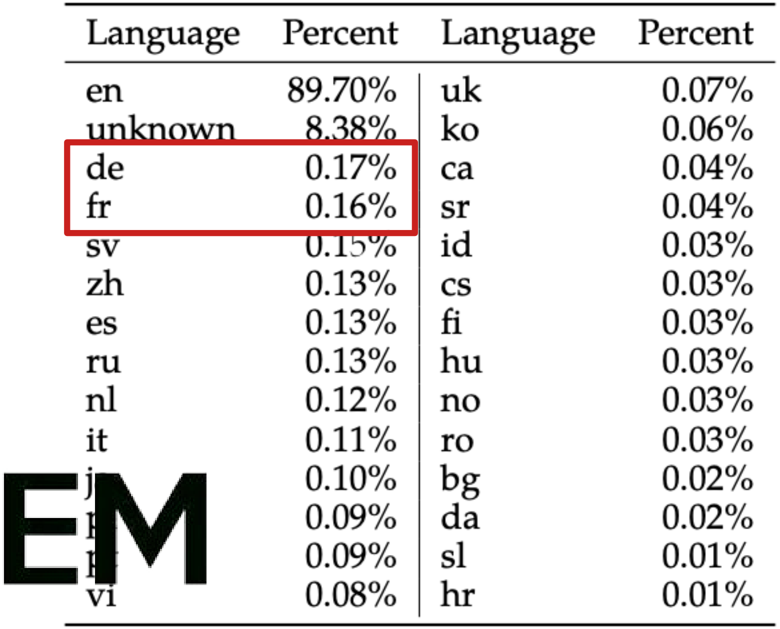
WE

HAVE

A

PROBLEM

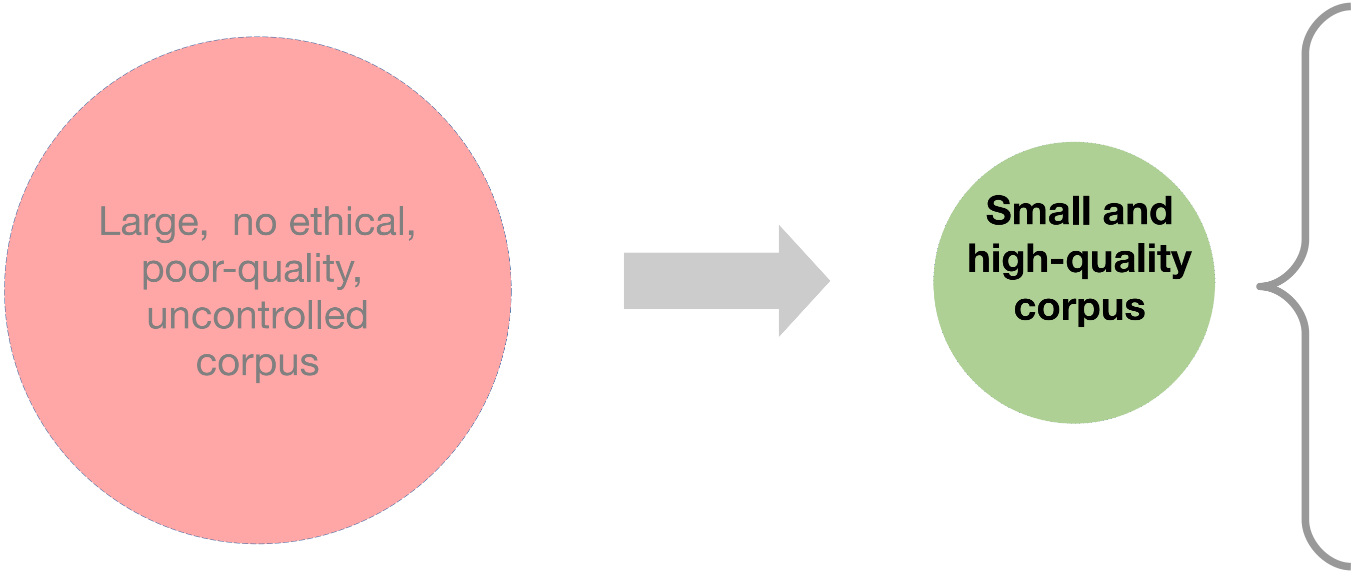
LLAMA V2 : Language distribution in pretraining data with percentage



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%

WE ADOPT A DATA-FIRST DRIVEN APPROACH

Change the bad LLM training practices currently used by all closed (chatGPT) or Open Weights models.



Large, no ethical, poor-quality, uncontrolled corpus

Small and high-quality corpus

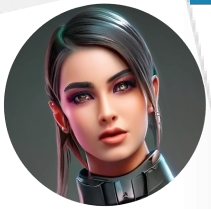
Small is beautiful !

- Compact & specialized models
- Better control → explicability / reliability
- Light training → continuous learning
- Low inference costs

Textbooks Are All You Need - https://arxiv.org/pdf/2306.11644.pdf

FIRST MODEL RELEASE IN OCTOBER 2023

CLAIRE

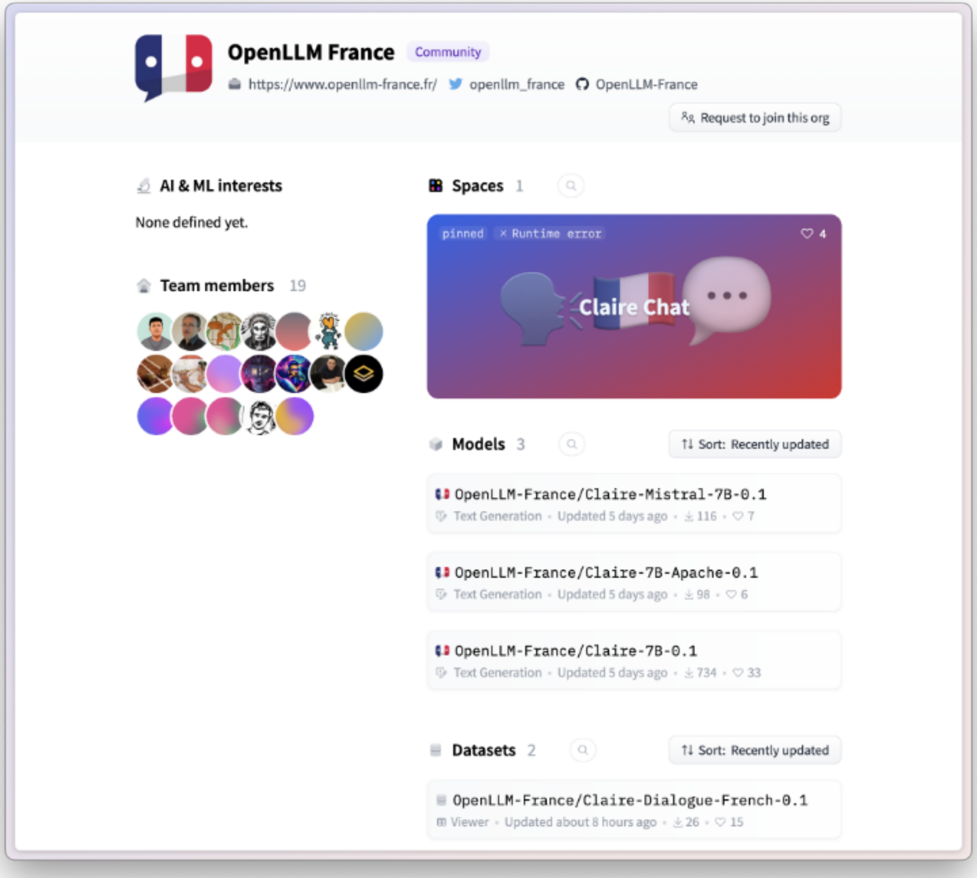


Fine tuning of Falcon 7B 1000H GPU (8xGPU) on Jean ZAY supercomputer 25 kWh and 1.5kg of C02 emission

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

Features:

- Understand dialogues with diarization
- Generation of human-like conversations (difluencies, hesitations…)



OpenLLM France Community

https://www.openllm-france.fr/ openllm_france OpenLLM-France

Request to join this org

AI & ML interests

None defined yet.

Team members 19

Spaces 1

Models 3

Sort: Recently updated

OpenLLM-France/Claire-Mistral-7B-0.1
Text Generation · Updated 5 days ago · 116 · 7

OpenLLM-France/Claire-7B-Apache-0.1
Text Generation · Updated 5 days ago · 98 · 6

OpenLLM-France/Claire-7B-0.1
Text Generation · Updated 5 days ago · 734 · 33

Datasets 2

Sort: Recently updated

OpenLLM-France/Claire-Dialogue-French-0.1
Viewer · Updated about 8 hours ago · 26 · 15

https://huggingface.co/OpenLLM-France

LUCIE WILL BE AVAILABLE IN APRIL 2024

LUCIE

7B <u>100 % Open Source</u> model

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

<u>Data:</u> 140B of high quality FR data (Gallica, Hal, Europarl, Wikipedia, CLAIRE Datasets...), EN (45B, peS2o), GE (3B), ES (3B), IT (??), Code (180B)

<u>Features:</u>

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

THE OPEN SOURCE WAY

SIMILAR INITIATIVES ELSEWHERE IN EUROPE (NOT EXHAUSTIVE)

PORO

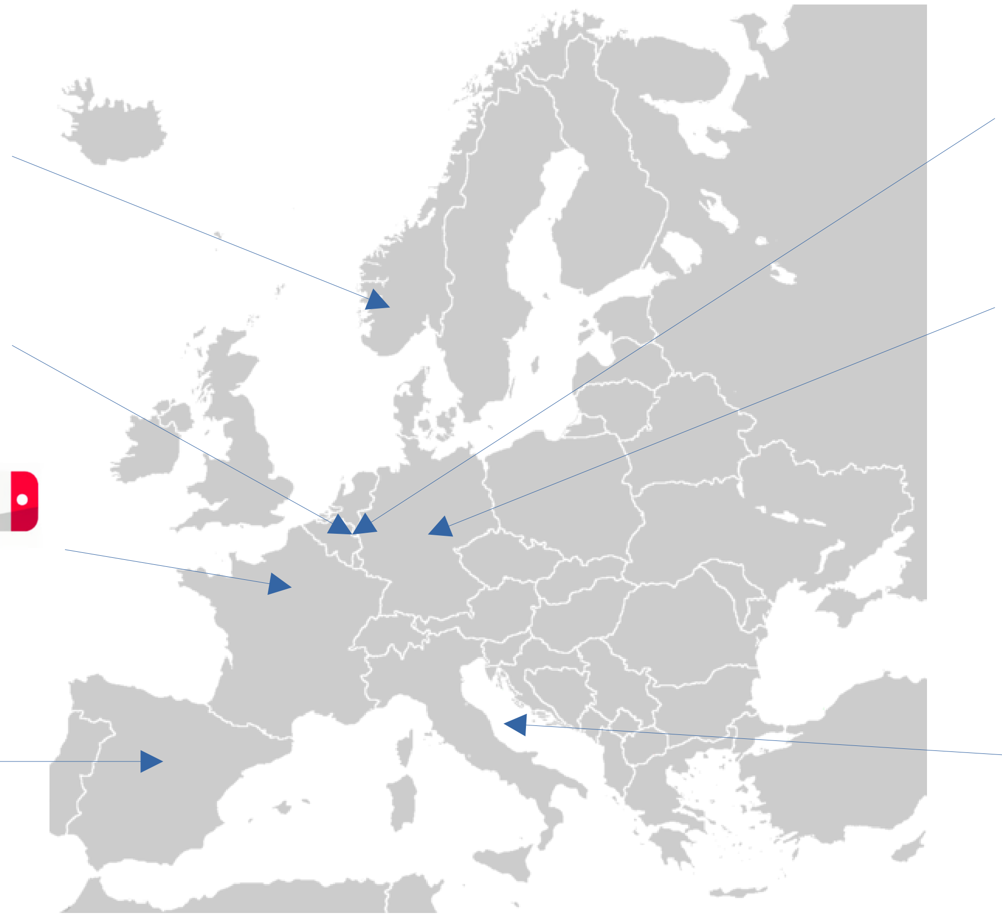
Poro - a family of open models that bring European languages to the frontier

SILOGEN

High Performance Language Technologies

OpenLLM-France

Ăguila Alpa



openGPT-X

LAION

Large-scale Artificial Intelligence Open Network

TRULY OPEN AI. 100% NON-PROFIT. 100% FREE.

Fauno - Italian LLM

IF YOU SHARE OUR VALUES AND OUR AMBITION

Feel free to join now the team 😍

OpenLLM-Europe

https://discord.gg/tZf7BR4dY7

contact@openllm-europe.org

THANKS FOR YOUR ATTENTION

@linagora

Villa Good Tech | 37 Rue Pierre Poli, 92130
Issy-les-Moulineaux, FRANCE
Tél. : +33 (0)1 46 96 63 63 - Fax : +33 (0)1 46 96 63 64