From OpenLLM France 🇫🇷 to OpenLLM Europe 🇪🇺

Paving the way to sovereign and open source AI

FOSDEM – Brussels - February 4, 2024
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
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
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

Green: All conditions are met
Yellow: 2 conditions are met
Orange: Only 1 condition is met
Red: No conditions are met
Geographical distribution of LLMs with more than one billion parameters since 2018

<table>
<thead>
<tr>
<th>Language</th>
<th>Percent</th>
<th>Language</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>89.70%</td>
<td>uk</td>
<td>0.07%</td>
</tr>
<tr>
<td>unknown</td>
<td>8.38%</td>
<td>ko</td>
<td>0.06%</td>
</tr>
<tr>
<td>de</td>
<td>0.17%</td>
<td>ca</td>
<td>0.04%</td>
</tr>
<tr>
<td>fr</td>
<td>0.16%</td>
<td>sr</td>
<td>0.04%</td>
</tr>
<tr>
<td>sv</td>
<td>0.15%</td>
<td>id</td>
<td>0.03%</td>
</tr>
<tr>
<td>zh</td>
<td>0.13%</td>
<td>cs</td>
<td>0.03%</td>
</tr>
<tr>
<td>es</td>
<td>0.13%</td>
<td>fi</td>
<td>0.03%</td>
</tr>
<tr>
<td>ru</td>
<td>0.13%</td>
<td>hu</td>
<td>0.03%</td>
</tr>
<tr>
<td>nl</td>
<td>0.12%</td>
<td>no</td>
<td>0.03%</td>
</tr>
<tr>
<td>it</td>
<td>0.11%</td>
<td>ro</td>
<td>0.03%</td>
</tr>
<tr>
<td>ja</td>
<td>0.10%</td>
<td>bg</td>
<td>0.02%</td>
</tr>
<tr>
<td>pl</td>
<td>0.09%</td>
<td>da</td>
<td>0.02%</td>
</tr>
<tr>
<td>pt</td>
<td>0.09%</td>
<td>sl</td>
<td>0.01%</td>
</tr>
<tr>
<td>vi</td>
<td>0.08%</td>
<td>hr</td>
<td>0.01%</td>
</tr>
</tbody>
</table>
Geographical distribution of LLMs with more than one billion parameters since 2018:

<table>
<thead>
<tr>
<th>Language</th>
<th>Percent</th>
<th>Language</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>89.70%</td>
<td>uk</td>
<td>0.07%</td>
</tr>
<tr>
<td>unknown</td>
<td>8.38%</td>
<td>ko</td>
<td>0.06%</td>
</tr>
<tr>
<td>de</td>
<td>0.17%</td>
<td>ca</td>
<td>0.04%</td>
</tr>
<tr>
<td>fr</td>
<td>0.16%</td>
<td>sr</td>
<td>0.04%</td>
</tr>
<tr>
<td>sv</td>
<td>0.15%</td>
<td>id</td>
<td>0.03%</td>
</tr>
<tr>
<td>zh</td>
<td>0.13%</td>
<td>cs</td>
<td>0.03%</td>
</tr>
<tr>
<td>es</td>
<td>0.13%</td>
<td>fi</td>
<td>0.03%</td>
</tr>
<tr>
<td>ru</td>
<td>0.13%</td>
<td>hu</td>
<td>0.03%</td>
</tr>
<tr>
<td>nl</td>
<td>0.12%</td>
<td>no</td>
<td>0.03%</td>
</tr>
<tr>
<td>it</td>
<td>0.11%</td>
<td>ro</td>
<td>0.03%</td>
</tr>
<tr>
<td>ja</td>
<td>0.10%</td>
<td>bg</td>
<td>0.02%</td>
</tr>
<tr>
<td>ja</td>
<td>0.09%</td>
<td>da</td>
<td>0.02%</td>
</tr>
<tr>
<td>ja</td>
<td>0.09%</td>
<td>sl</td>
<td>0.01%</td>
</tr>
<tr>
<td>vi</td>
<td>0.08%</td>
<td>hr</td>
<td>0.01%</td>
</tr>
</tbody>
</table>
WE ADOPT A DATA-FIRST DRIVEN APPROACH

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

Textbooks Are All You Need - [https://arxiv.org/pdf/2306.11644.pdf](https://arxiv.org/pdf/2306.11644.pdf)

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
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 (fluencies, hesitations…)

https://huggingface.co/OpenLLM-France
LUCIE WILL BE AVAILABLE IN APRIL 2024

LUCIE

7B 100 % Open Source model

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

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

Features:
- 100 % open source datasets
- 16k context windows
- Rotary & sliding windows
- Custom tokenizer
SIMILAR INITIATIVES ELSEWHERE IN EUROPE (NOT EXHAUSTIVE)

High Performance Language Technologies

OpenLLM-France

Águila Alpa

Fauno - Italian LLM

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

openGPT-X

LAION

Large-scale Artificial Intelligence Open Network
TRULY OPEN AI. 100% NON-PROFIT. 100% FREE.
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