

Claire LLM

*OpenLLM France:
Developing open language
models in France* 🇫🇷



WHO AM I - OUR MISSION WITH LINAGORA SINCE 2000



Julie Hunter
NLP/Linguistics

✉ jhunter@linagora.com



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

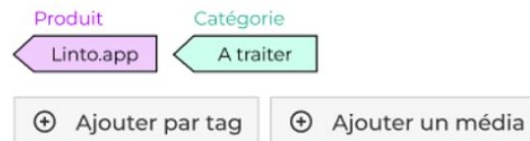




The screenshot shows the LinTO interface with a sidebar of linguistic tools and a main editor area. The sidebar includes 'Outils linguistiques' and 'Mots clés' with categories like Bâche, Changement, Conjoncture, Coup de pouce, Informatique, Point de vue, Propositions, Système, Système si, and Systèmes. The main editor is titled 'Gestion d'équipe' and contains a transcription of a meeting with speaker labels (spk3, spk4, spk3, spk2, spk3) and highlighted words like 'système' and 'changement'. A video player at the bottom shows a waveform and a 'Capture d'écran' label.

Transcription + correction and annotation

Contexte de la demande



The diagram shows 'Produit' as 'Linto.app' and 'Catégorie' as 'A traiter'. Below are two buttons: 'Ajouter par tag' and 'Ajouter un média'.

Compile les événements à venir mentionnés dans ces documents

Bien sur ! j'ai trouvé 3 dates.

- Le 29 novembre, il y a l'innovation coffee, avec la présentation de linto.app
- Le 4 décembre, Ismail présentera à la réunion du lundi les avancés sur l'extra
- Le 5 décembre, vous avez une réunion avec la commission européenne

J'ai créé pour vous un fichier [calendrier-linto-app.ics](#) que vous pouvez importer

 [calendrier-linto-app.ics](#)

Query / response, summarization

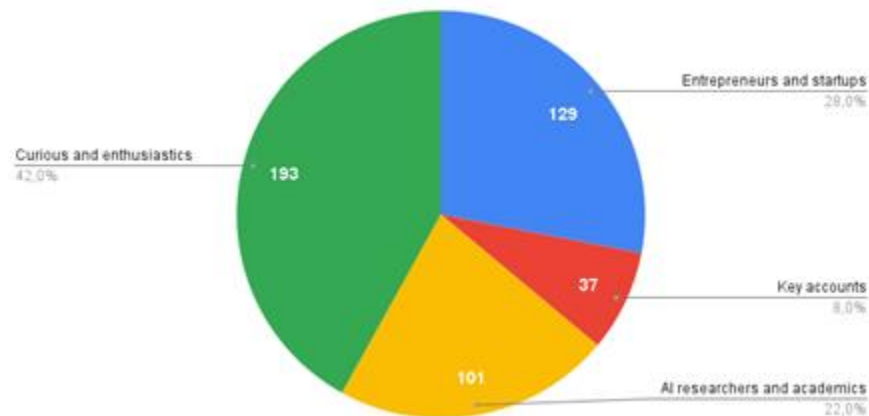
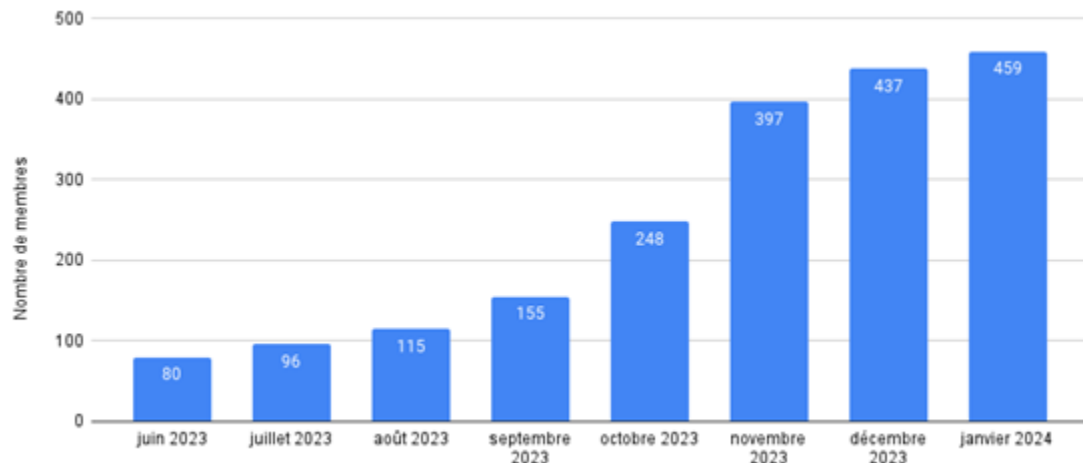
1

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

2

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

WE'VE ALREADY COMPLETED OUR SECOND OBJECTIVE



Academics / Public research



Corporates



July 20, 2023 Stefano Maffulli OSI opinion

Meta's LLaMa 2 license is not Open Source

(...neither is Mistral)



open source
initiative®

What is Open Source AI

An Open Source AI is an AI system made available to the public under terms that grant the freedoms to:

- **Use** the system for any purpose and without having to ask for permission.
- **Study** how the system works and inspect its components.
- **Modify** the system for any purpose, including to change its output.
- **Share** the system for others to use with or without modifications, for any purpose.



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

Model as a derivative work from data

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

THE PUSH FOR OPEN DATA

- **Common Crawl** (web)
- **C4** (web)
- **ArXiv** (academic papers)
- **Wikipedia**
- **StackExchange** (Q/A threads)
- **The Gutenberg project** (Books)
- **The Stack** (GitHub)
- ...



Hugging Face



BigScience



together.ai



1. Web crawled data
2. English

SOME CONCERNS

Web data

- personal information
- toxicity
- quality
- format
- duplication



RefinedWeb, OSCAR, CulturaX
(but can't solve everything)

English

- not just a choice
- not just about the language

```
Pour faire un boeuf  
bourguignon il faut
```

```
---
```

```
commencer par un bon vin.  
To make a beef bourguignon  
you must start with a good  
wine.
```

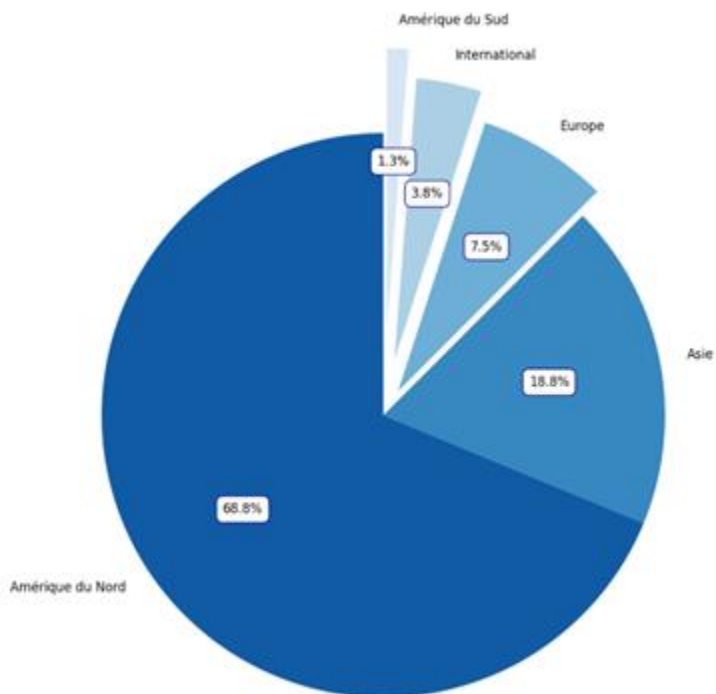
```
- Julia Child
```

```
#####JULIA CHILD
```

```
- - Mistral
```

CREATING LLMS 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%

SMALL IS BEAUTIFUL

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

CLAIRE FOUNDATION MODEL (OCTOBER 2023)

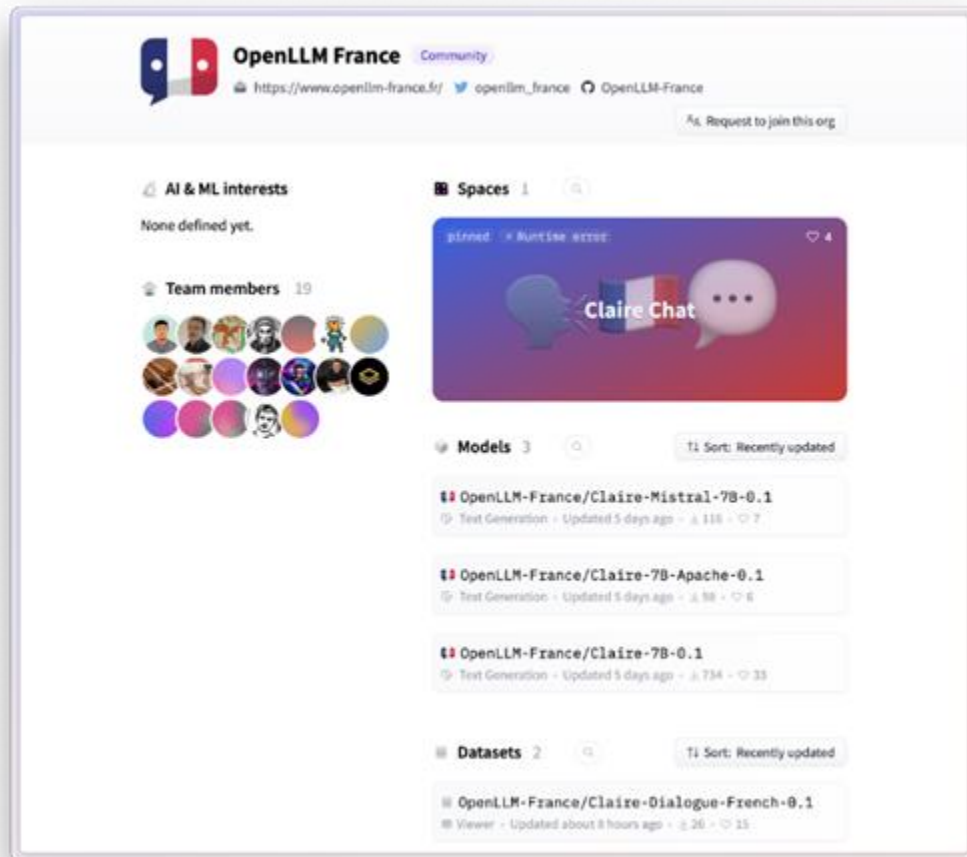
CLAIRE

Continual pre-training of Falcon 7B
1000H GPU (8xGPU) on
Jean ZAY supercomputer
25 kWh and 1.5kg of CO2 emission

Data: 138M FR conversational data
(drama show, literature and real-life
meetings transcriptions). **No web-
crawled data**

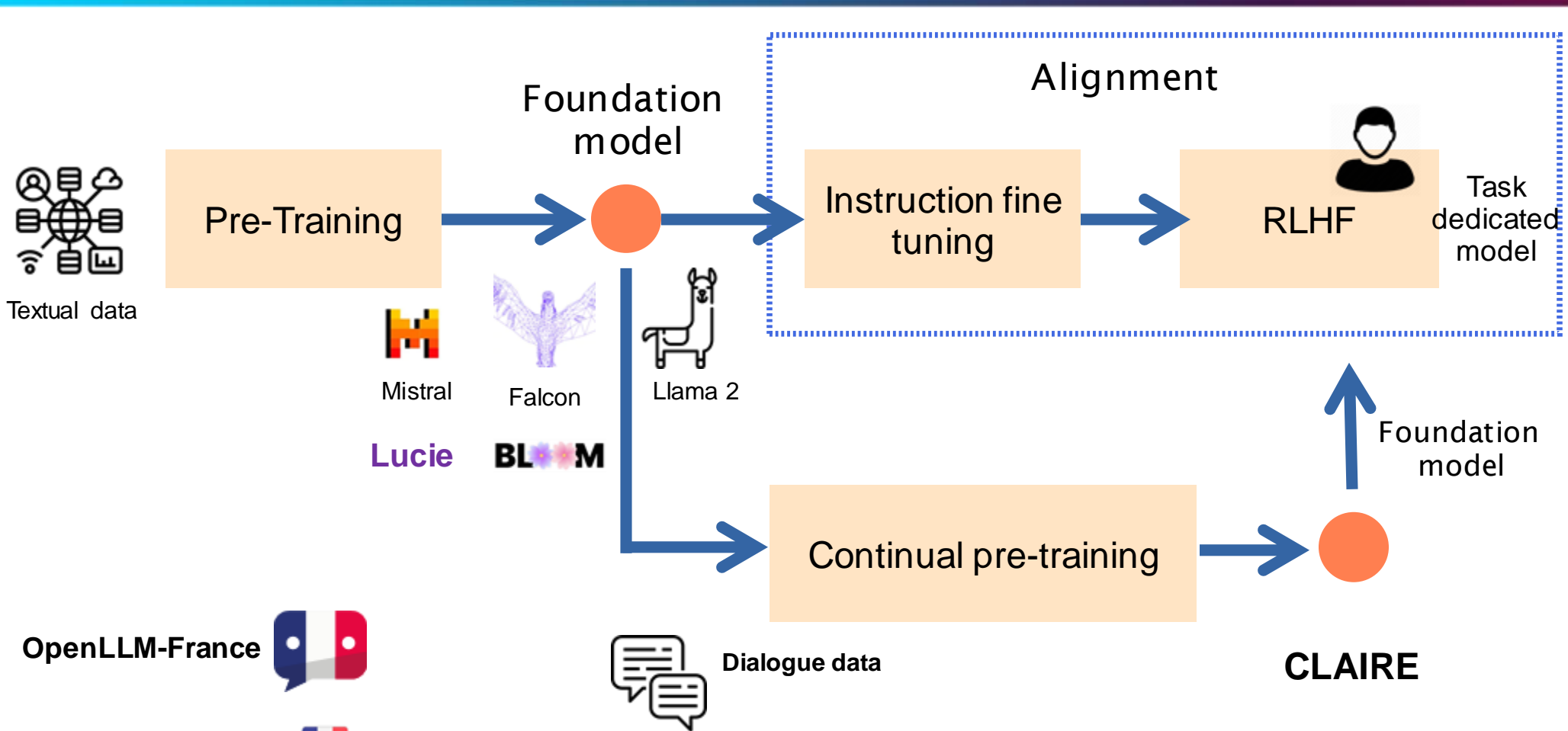
Features:

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



<https://huggingface.co/OpenLLM-France>

CONTINUAL PRETRAINING





Appel à projets Thématiques Spécifiques en Intelligence Artificielle (TSIA) 2023

Research and development of open, voice and text, generative models
for conversation

Applications to business meetings and SAMU emergency call analysis

<https://ia.loria.fr/llm4all>



Hugging Face



LUCIE IS UNDERWAY

LUCIE

7B 100 % Open Source model

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

Data: 120B of high quality FR data
(Gallica, Hal, Europarl, Wikipedia,
CLAIRE Datasets...), EN (60B,
peS2o), GE (3B), ES (3B), IT (2B),
Code (50B)

Features:

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

THE OPEN
SOURCE WAY



IF YOU SHARE OUR VALUES AND OUR AMBITION

Feel free to join the team 

OpenLLM-Europe

on Discord

<https://discord.gg/3KvntCm6>



**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



The screenshot shows the OpenLLM France community page on Hugging Face. At the top, there is a header with the logo and the text "OpenLLM France Community". Below this, there are sections for "AI & ML interests" (None defined yet), "Team members" (19 members), "Spaces" (1 space), "Models" (3 models), and "Datasets" (2 datasets). The "Models" section is expanded, showing three models:

- OpenLLM-France/Claire-Mistral-7B-0.1**: Text Generation - Updated 5 days ago - 116 likes
- OpenLLM-France/Claire-7B-Apache-0.1**: Text Generation - Updated 5 days ago - 98 likes
- OpenLLM-France/Claire-7B-0.1**: Text Generation - Updated 5 days ago - 734 likes

The "Datasets" section shows one dataset:

- OpenLLM-France/Claire-Dialogue-French-0.1**: Viewer - Updated about 8 hours ago - 26 likes

<https://huggingface.co/OpenLLM-France>

The screenshot shows the OpenLLM-France GitHub repository page. At the top, there is a header with the logo and the text "OpenLLM-France". Below this, there is a section for "Welcome to OpenLLM-France" with a link to the community page. The main content area contains the following text:

The aim of the [OpenLLM France community](#) is to collaborate on the development of a French, sovereign, and truly Open Source LLM, that would be based on

- public and open training corpora,
- documented algorithms to ensure their explicability, and
- free, non-restrictive user licenses.

Follow us:

- [Discord](#)
- [Hugging Face](#)

<https://github.com/OpenLLM-France>