Claire LLM

OpenLLM France:
Developing open language models in France 🇫🇷

UTIAS – April 5, 2024

WHO AM I - OUR MISSION WITH LINAGORA SINCE 2000

Julie Hunter
NLP/Linguistics

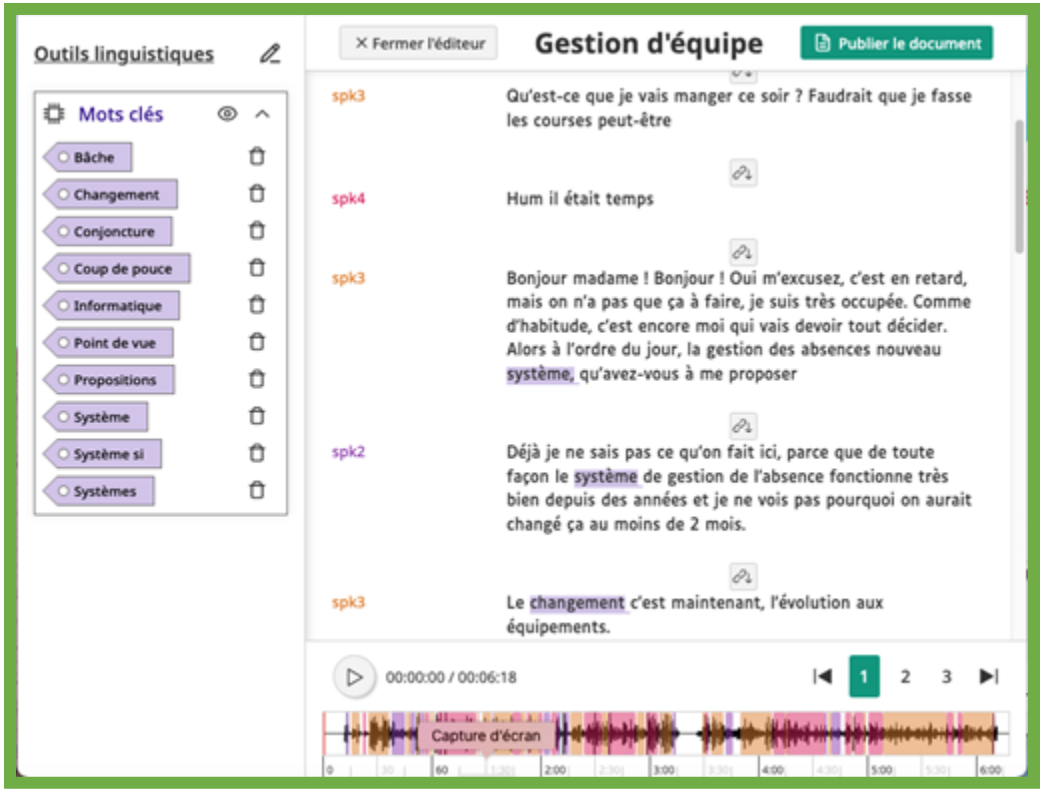
✉ jhunter@linagora.com

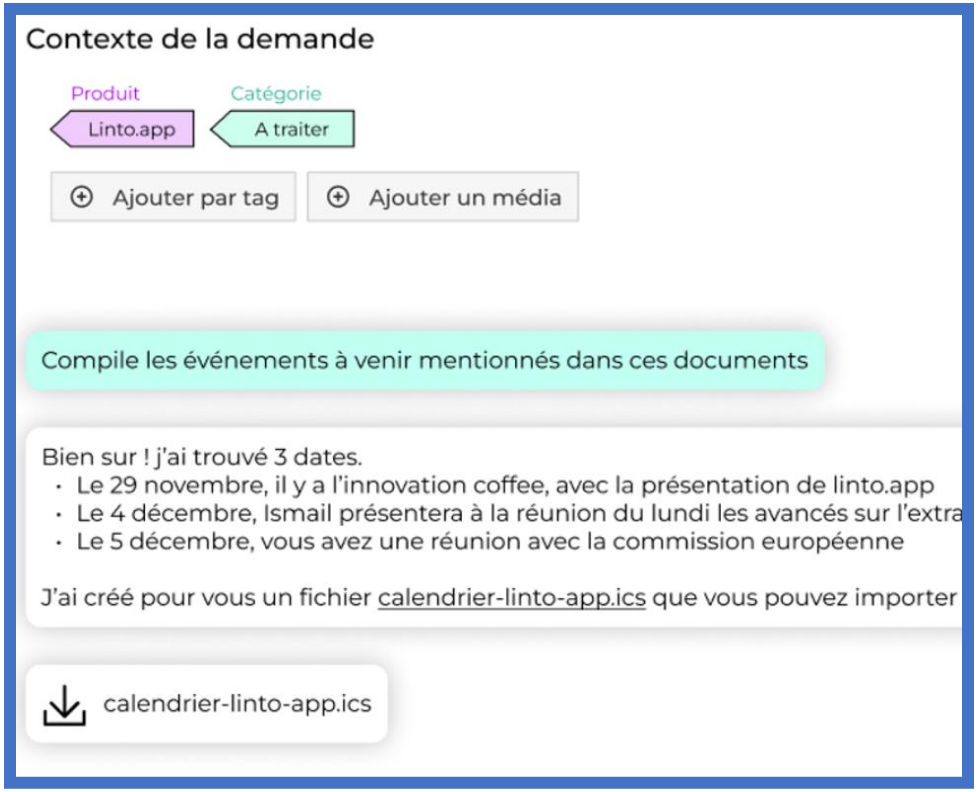
#GoodTechForGood
fantastic-enterprise.com

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

LINAGORA

LINTO: PLATE-FORME VOCALE





Transcription + correction and annotation

Query / response, summarization

AMBITION OF THE OPENLLM FRANCE FR INITIATIVE

1

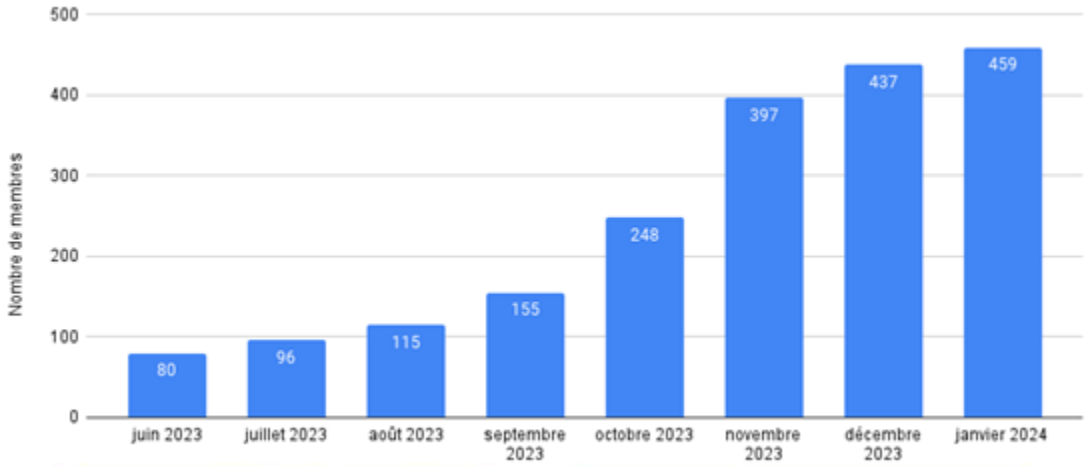
Build trusted, sovereign and
REAL Open Source AI models & technologies

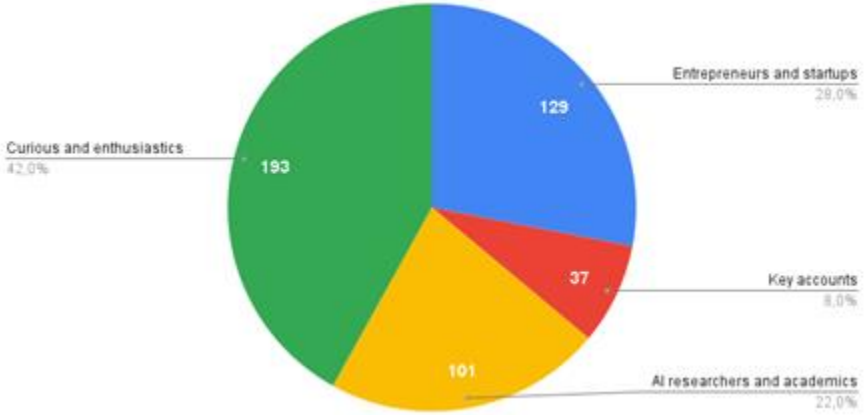
2

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 · IDRIS · WIKIMEDIA · Loria · list · ALMAnaCH · LIX · Class'Code · PANTHÉON SORBONNE · LE VOICE LAB · HUB FRANCE IA · ANITI

Corporates

LINAGORA.AI · opsci · THALES Together • Safer • Everywhere · FARADAY · TALKR.ai · 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

(...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.

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

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

EleutherAI

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

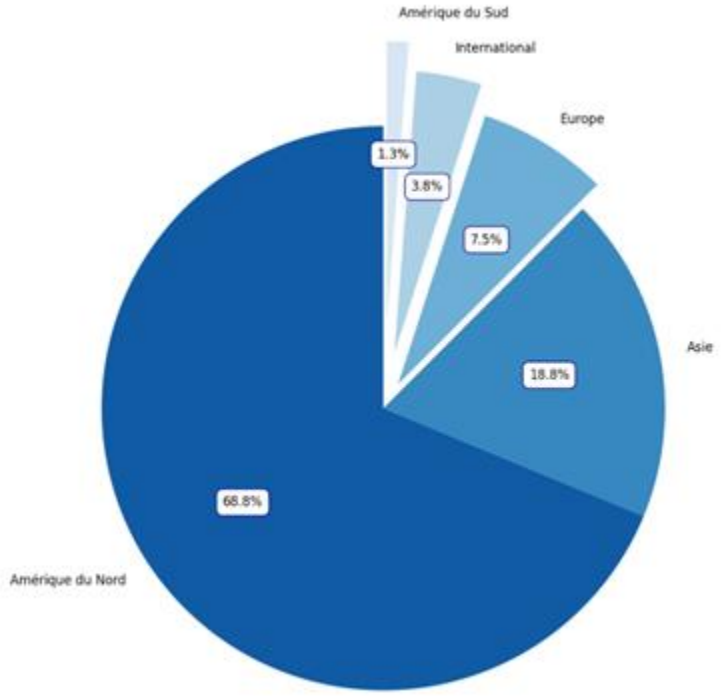
- not just a choice
- not just about the language

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

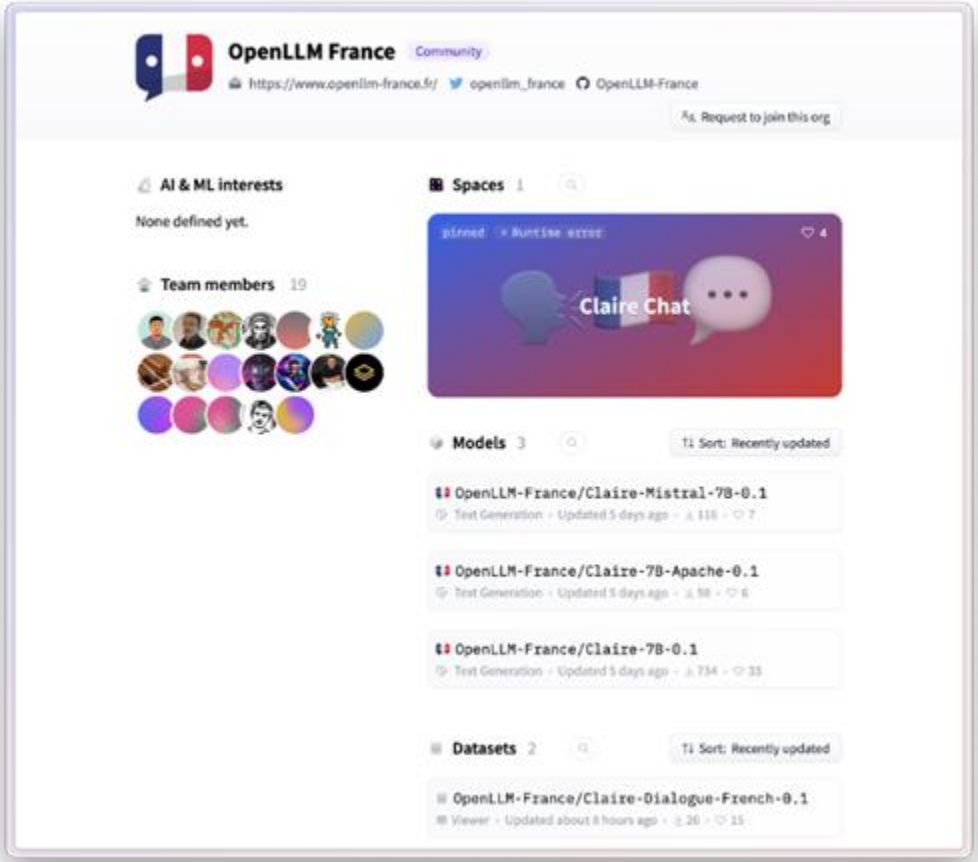
CLAIRE FOUNDATION MODEL (OCTOBER 2023)

CLAIRE

Continual pre-training of Falcon 7B
1000H GPU (8xGPU) on
Jean ZAY supercomputer
25 kWh and 1.5kg of C02 emission

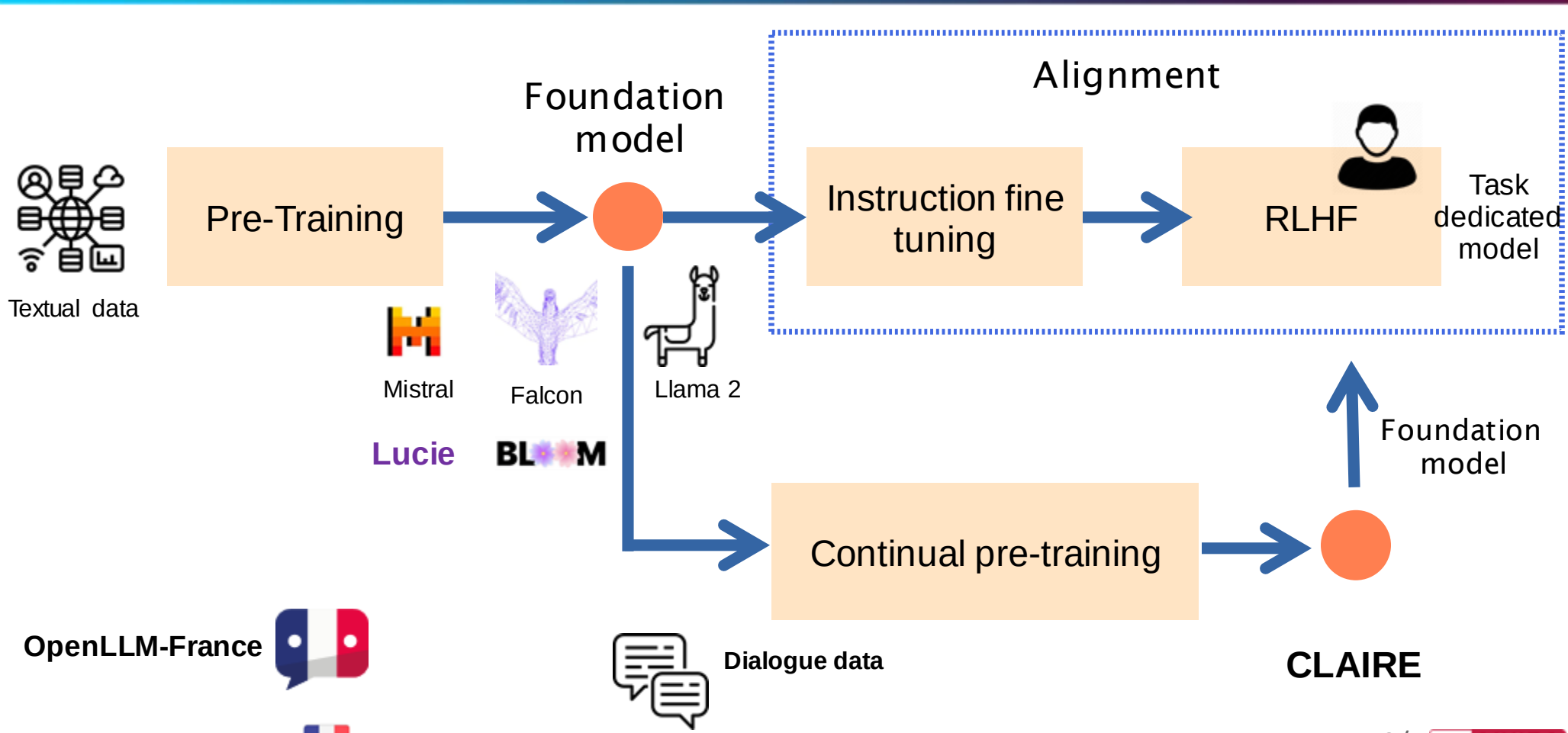
Data: 138M FR conversational data (drama show, literature and real-life meetings transcriptions). **No web-crawled data**

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

Loria Laboratoire lorrain de recherche en informatique et ses applications

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

LINAGORA

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

Feel free to join the team 

OpenLLM-Europe

on Discord

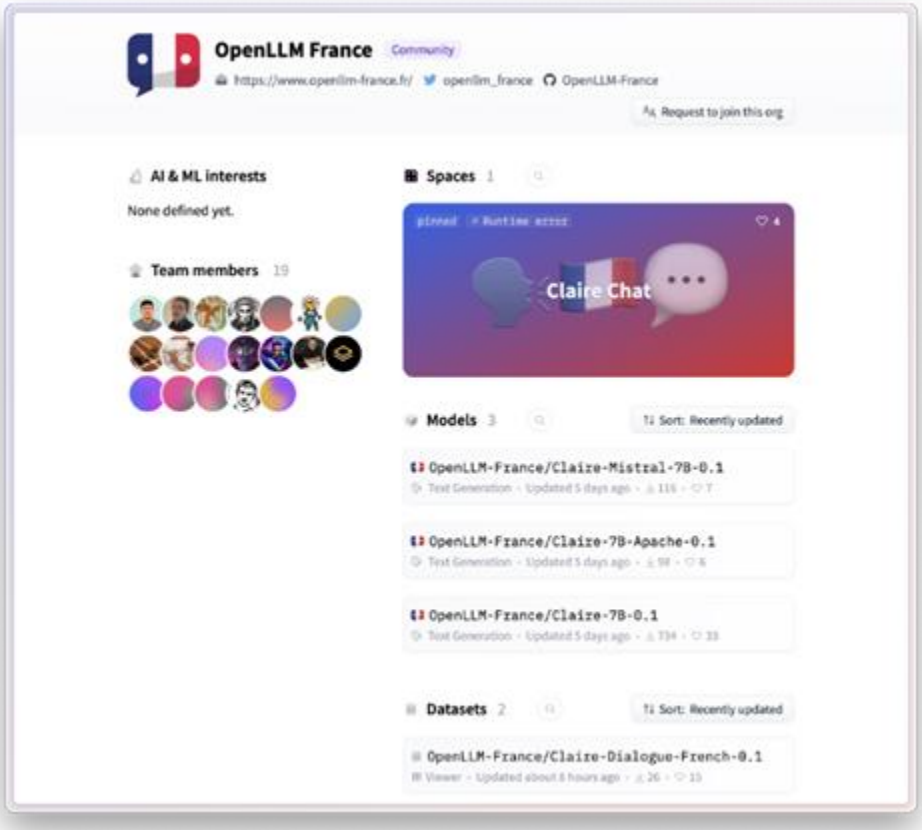
https://discord.gg/3KvntCm6

THANKS FOR YOUR ATTENTION

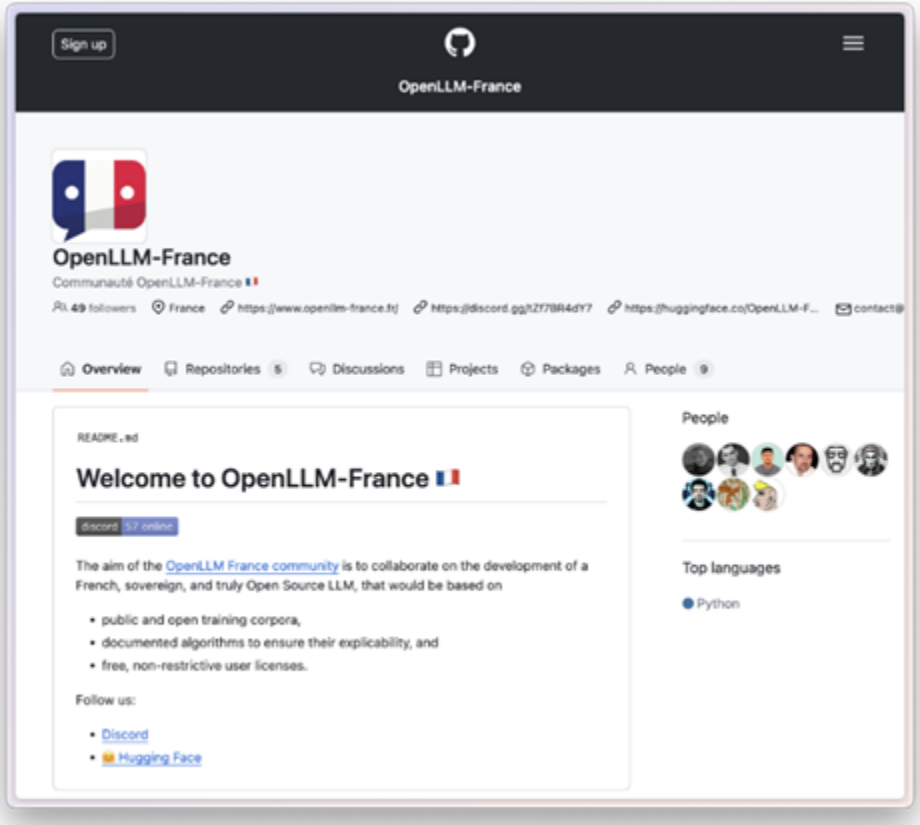
@linagora

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

OPENLLM FRANCE



https://huggingface.co/OpenLLM-France



https://github.com/OpenLLM-France