Building Open Source Language Models

Julie Hunter

LINAGORA

FOSDEM – Brussels - February 4, 2024

OPEN SOURCE MOMENTUM

Meta

mosaicML

TII Technology Innovation Institute

Open Weights models

control over deployment

possibility of fine-tuning

But leaves a lot open to guess work if you want to know *why* something works (or doesn't)

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

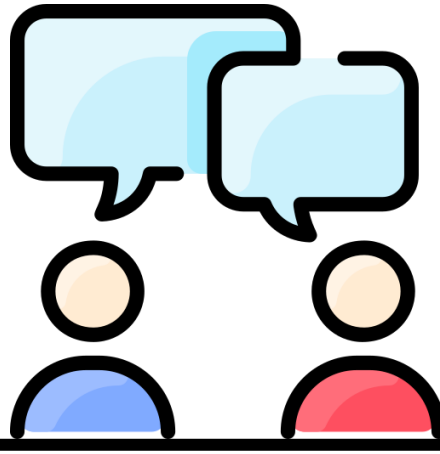
```
Pour faire un boeuf
bourguignon il faut
---
commencer par un bon vin.
To make a beef bourguignon
you must start with a good
wine.
- Julia Child
#####JULIA CHILD
```

- - Mistral

FIRST OBJECTIVES

- Create a **French** dataset with **traceable licenses** that were compatible with LLM training
- Use it to **fine-tune an open weights model** (as a first step) to evaluate impact of dataset

Use case: **spoken dialogue** for understanding and generation



FRENCH DATA DESCRIPTION AND SOURCES

ASSEMBLÉE NATIONALE

Théâtre Classique, Théâtre Gratuit

LINAGORA

Ortolang

1 [cs.CL] 28 Nov 2023

THE CLAIRE FRENCH DIALOGUE DATASET

Julie Hunter[*]
LINAGORA
Toulouse, France
jhunter@linagora.com

Jérôme Louradour[*]
LINAGORA
Toulouse, France
jlouradour@linagora.com

Virgile Rennard
LINAGORA
Ecole Polytechnique
Paris, France
vrennard@linagora.com

Ismaïl Harrando
LINAGORA
Toulouse, France
iharrando@linagora.com

Guokan Shang
LINAGORA
Paris, France
gshang@linagora.com

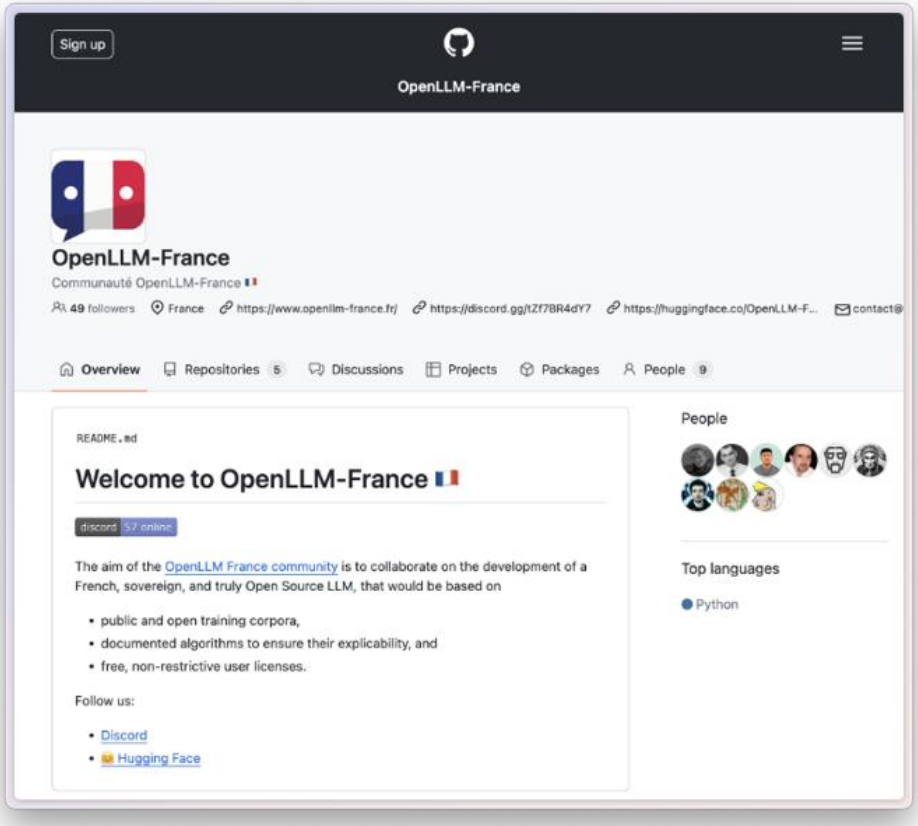
Jean-Pierre Lorré
LINAGORA
Toulouse, France
jplorre@linagora.com

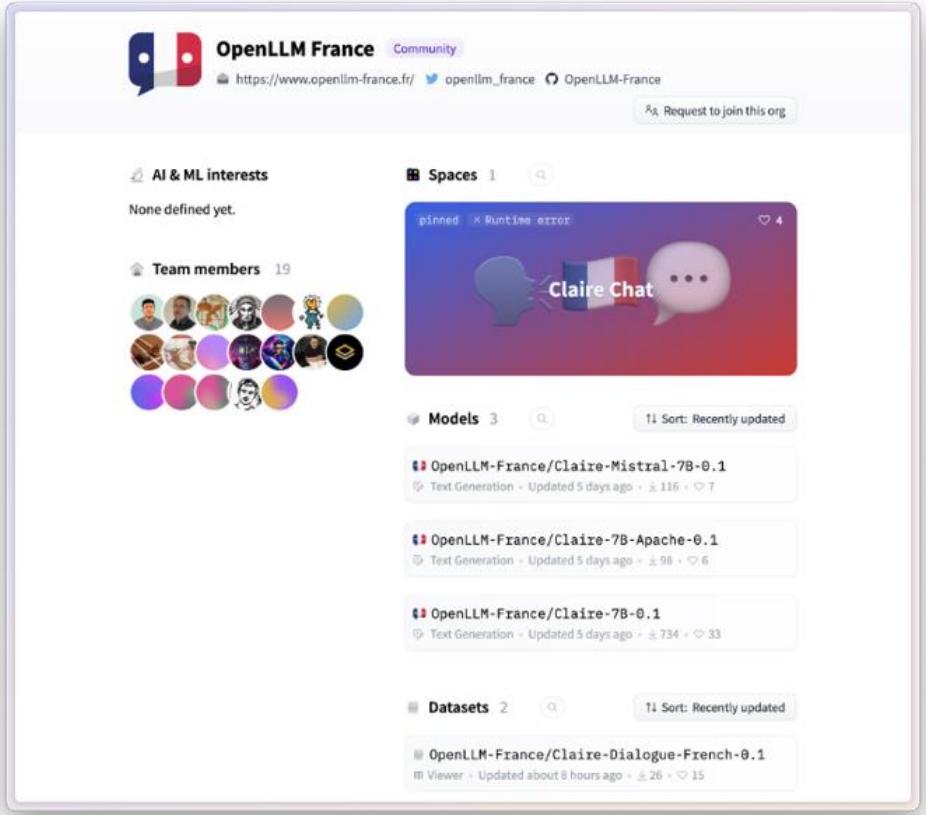
November 2023

We present t … members of LINAGORA … us containing …

arxiv:2311.16840

SHARING THE DATA, CODE AND MODELS





https://github.com/OpenLLM-France

https://huggingface.co/OpenLLM-France

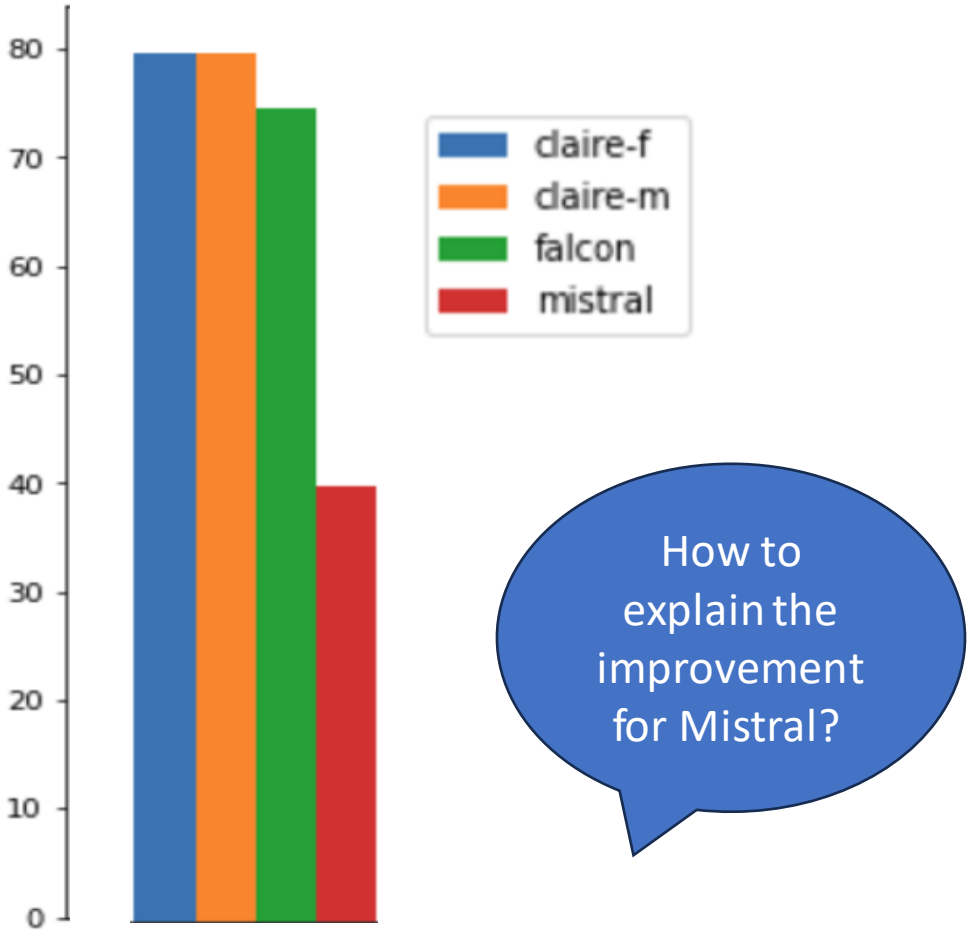
EVALUATION

Does the model stick to French?

```
Je sais pas non plus. Ça veut dire que le
mot est retourné en sens inverse.
[Marc:] Oh oui, comme le verlan.


\\bigskip


\\begin{minipage}{0.55\\textwidth}
\\begin{center}
\\begin{tabular}{r l}
\\toprule
\\textit{Mot} & \\textit{Mot inversé et son sens} \\\\
\\midrule
bécane    ...
```



How to explain the improvement for Mistral?

NEXT STEPS

LUCIE

From scratch
7B parameters
Even mixture of French and English
+ some code
+ some German and Spanish

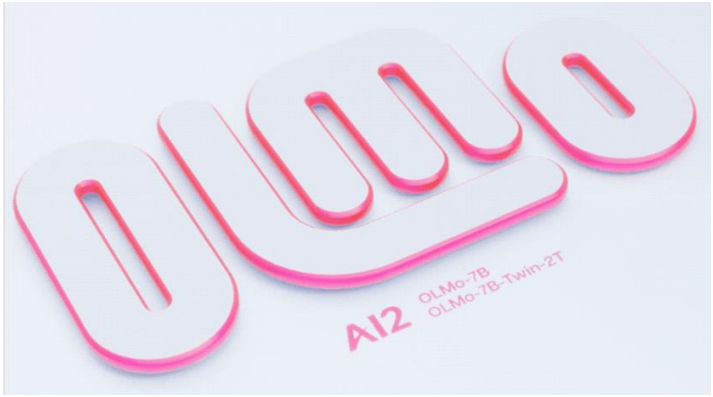
~140B tokens of French

- Gallica (French National Library)
- Wikipedia
- Academic publications with open access
- Parlementary and public addresses

Some complications: OCR

~s. < i A 'rr-oAtt——~-~———-. --, 1 S ..-----j| A TERME. I Derniers cottrt I IDttrétl 1 -T-S.* J Tsox A TERME. cotés «l J^Ptant. Liq. ! : -. précidemmrnt. et à démission Il le, couro Pins liant. fPÎ. bas. |Dera« cours. ^'c^^TTermèr] Dividendes. H22ii autre ■ «T. ■ eu ," t..!. -~ 8 1 g O o/ ! cakq. ,. | 1 Is ? ! S *3 /o .J| l" ianv. 76. ! 67f33 40 45 30 55 60 63 60 ,, fin coar. 67 50 67 75 S 67 45 67 70 67 30 67 37I | jPr.fteîc.] d 2n H S > ! | S I iPr.iinc. d lf,67 77% 68 d 50! 1 il SI Pr. fin c. 68 68 20 d 25 S Pr.Rnp. .d2f Ij |

RELATED (AND VERY RECENT!) PROJECTS



Papers arxiv:2402.00786

CroissantLLM: A Truly Bilingual French-English Language Model

Published on Feb 1 · Featured in Daily Papers on Feb 2

Datasets: PleIAs / French-PD-Newspapers

Tasks: Text Generation Languages: French Tags: ocr

OpenLLM-France

More open data – in more languages –and more open models on the way!

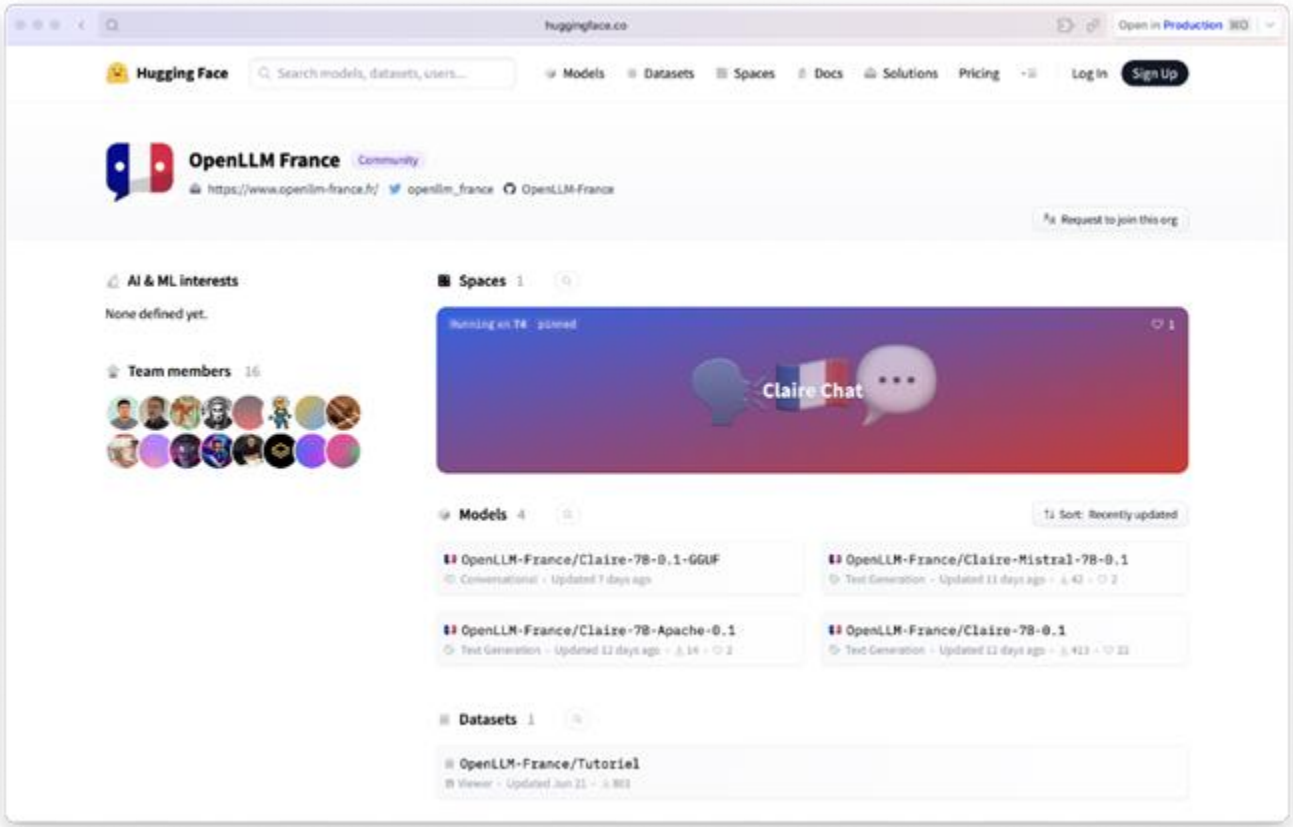
MERCI DE VOTRE ATTENTION

@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

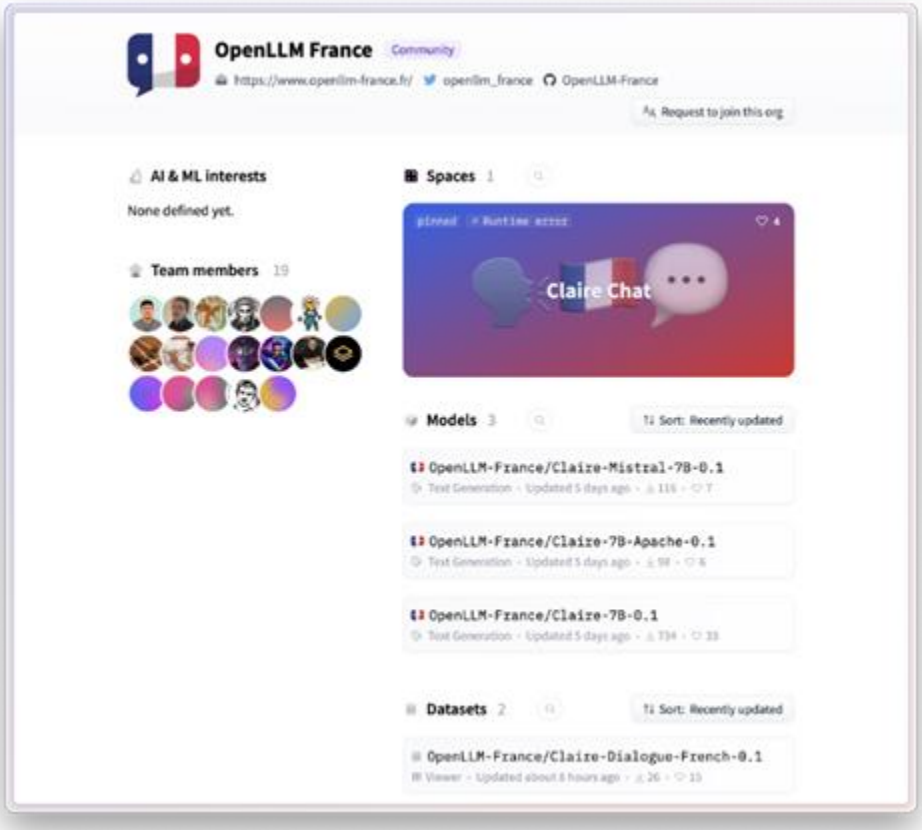
OPENLLM FRANCE

CLAIRE : 1ÈRE PHASE DE NOTRE DÉMONSTRATION SCIENTIFIQUE

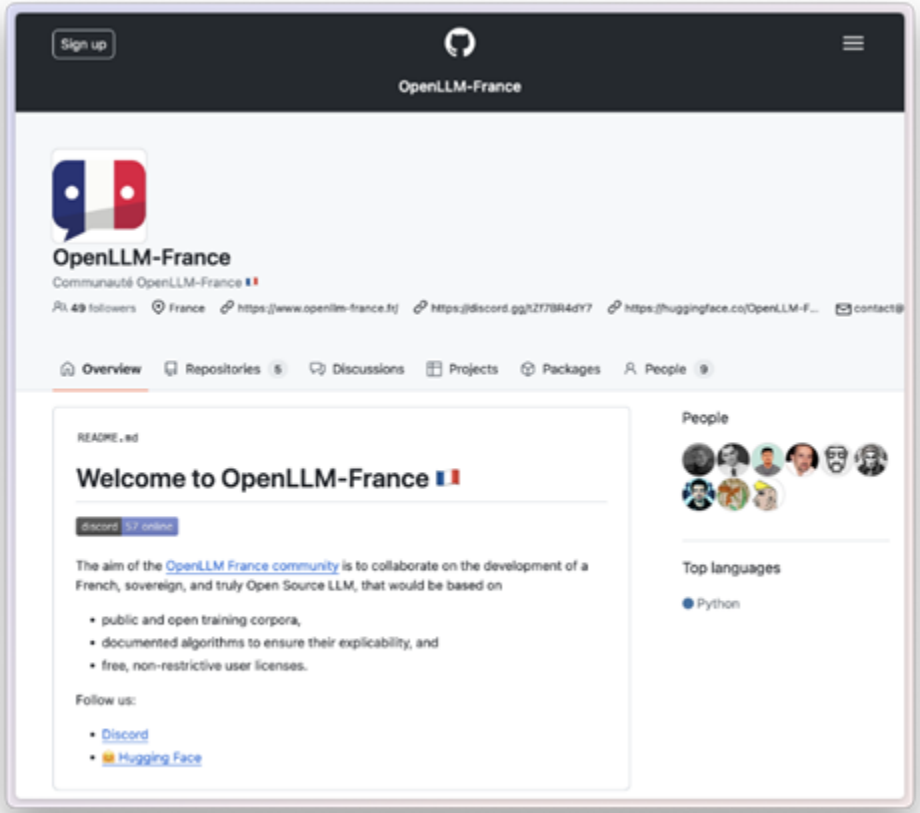


OPENLLM FRANCE

REJOIGNONS-NOUS !



https://huggingface.co/OpenLLM-France



https://github.com/OpenLLM-France