

# Building Open Source Language Models

Julie Hunter

**LIN** AGORA





## 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**



**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

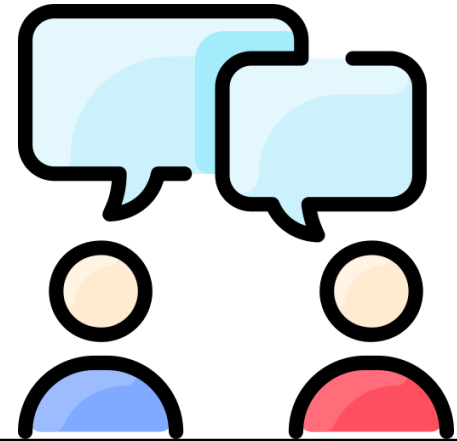
#####JULIA CHILD

- - 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



Théâtre  
Classique,  
Théâtre Gratuit



## 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

**Ismail Harrando**  
LINAGORA  
Toulouse, France  
iharrando@linagora.com

**Guokan Shang**  
LINAGORA  
Paris, France  
gshang@linagora.com

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

November 2023

We present t  
LINAGORA



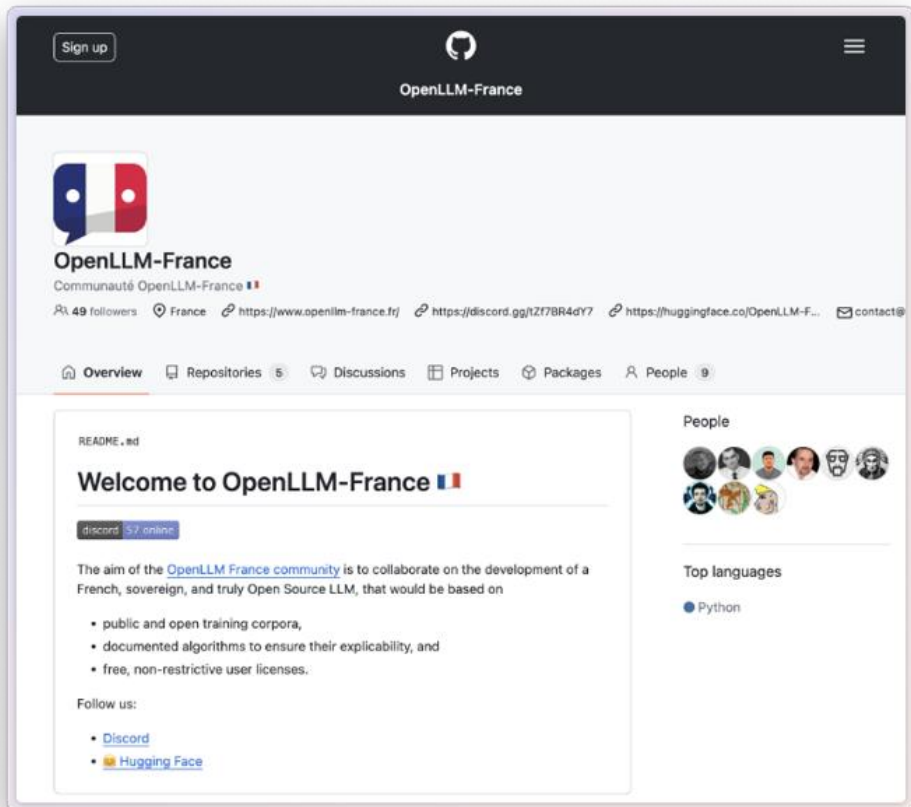
arxiv:2311.16840

members of  
us containing

1 [cs.CL] 28 Nov 2023

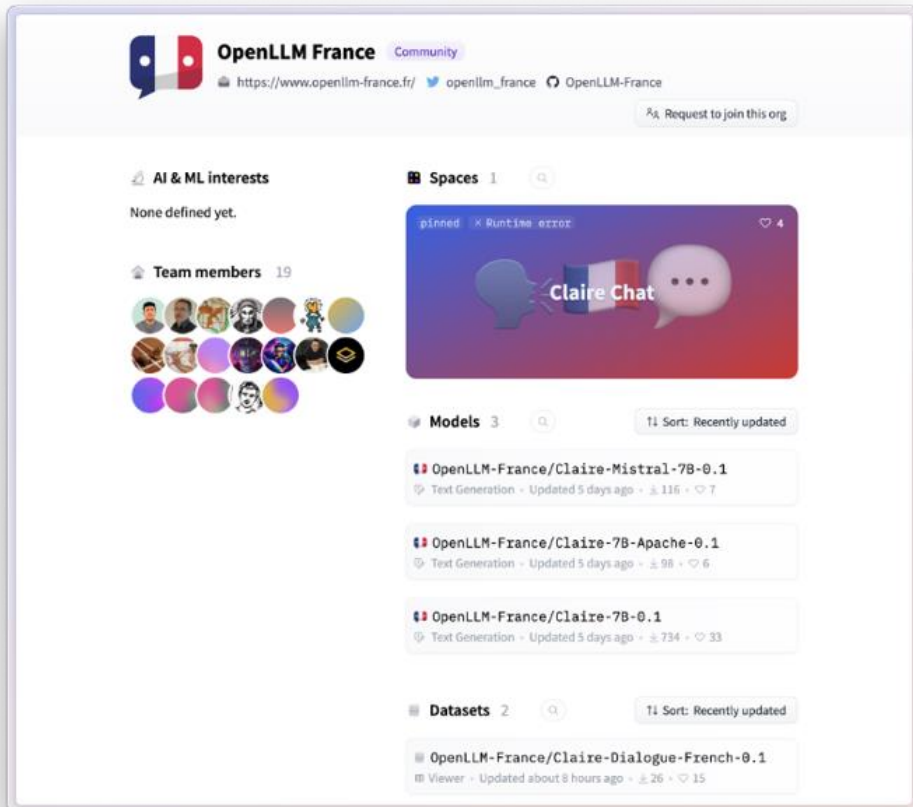


# SHARING THE DATA, CODE AND MODELS



The screenshot shows the GitHub profile for OpenLLM-France. At the top, there is a 'Sign up' button and the OpenLLM-France logo. Below the logo, the name 'OpenLLM-France' is displayed, followed by 'Communauté OpenLLM-France' and a list of links for followers, location, website, Discord, HuggingFace, and contact. The main content area shows a 'README.md' file with a 'Welcome to OpenLLM-France' heading. The text describes the community's aim to collaborate on developing a French, sovereign, and truly Open Source LLM, based on public and open training corpora, documented algorithms, and free, non-restrictive user licenses. It also lists 'Follow us' links for Discord and Hugging Face. On the right side, there are sections for 'People' (19 members) and 'Top languages' (Python).

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



The screenshot shows the HuggingFace profile for OpenLLM-France. At the top, there is the OpenLLM-France logo, the name 'OpenLLM France', and a 'Community' tag. Below this, there are links for the website, Twitter, and GitHub. A 'Request to join this org' button is visible. The main content area is divided into sections: 'AI & ML interests' (None defined yet), 'Team members' (19 members), 'Spaces' (1 space), 'Models' (3 models), and 'Datasets' (2 datasets). The 'Models' section lists three models: 'OpenLLM-France/Claire-Mistral-7B-0.1', 'OpenLLM-France/Claire-7B-Apache-0.1', and 'OpenLLM-France/Claire-7B-0.1'. The 'Datasets' section lists 'OpenLLM-France/Claire-Dialogue-French-0.1'. A 'pinned' space titled 'Runtime error' is also visible, featuring a 'Claire Chat' interface.

<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}
```

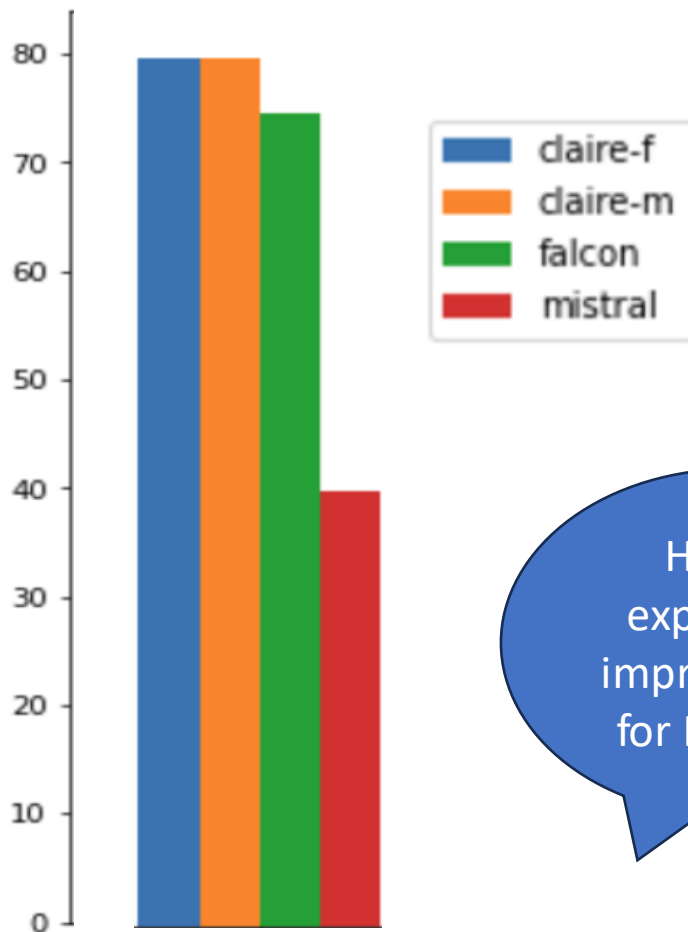
```
\\begin{tabular}{r l}
```

```
\\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
- Parliamentary and public addresses

Some complications: OCR

~s. < i A 'rr-oAtt—~—~———. --, 1 S ..-----j | A TERME. I  
Derniers cotttr l IDttrétl 1 -T-S.\* J Tsox A TERME. cotés «l J^Ptant.  
Liq. ! : -. précidemmnt. 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 ls ? ! 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 37l | jPr.fteîc.] d 2n H S > ! | S l iPr.iinc. d lf,67 77% 68 d 50! 1  
il SI Pr. fin c. 68 68 20 d 25 S Pr.Rnp. .d2f lj |

# 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](#)



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



The screenshot shows the Hugging Face profile for OpenLLM France. The page includes a search bar, navigation links for Models, Datasets, Spaces, Docs, Solutions, and Pricing, and a 'Sign Up' button. The community page features a profile picture, a bio, and a 'Request to join this org' button. Below the profile, there are sections for 'AI & ML interests', 'Team members', and 'Spaces'. The 'Spaces' section is highlighted, showing a space named 'Claire Chat' with a banner image. Below the space, there are sections for 'Models' and 'Datasets'. The 'Models' section lists several models, including 'OpenLLM-France/Claire-7B-0.1-GGUF', 'OpenLLM-France/Claire-Mistral-7B-0.1', 'OpenLLM-France/Claire-7B-Apache-0.1', and 'OpenLLM-France/Claire-7B-0.1'. The 'Datasets' section lists 'OpenLLM-France/Tutoriel'.

The screenshot shows a tweet by Jean-Pierre LORRE, Research director at Linagora. The tweet text reads: "LINAGORA et la communauté #OpenLLM\_France publient aujourd'hui le premier modèle ouvert #LLM CLAIRE sur Hugging Face : il s'agit du modèle Claire-7B-0.1. ...voir plus". Below the text is a video thumbnail with the text "OpenLLM-France /Claire-7B-0.1" and the Hugging Face logo. The video title is "OpenLLM-France/Claire-7B-0.1 · Hugging Face" and the description is "huggingface.co · Lecture de 4 min". Below the video, there is a red box highlighting the engagement information: "Vous et 377 autres personnes · 27 commentaires · 40 republications". At the bottom of the tweet, there are buttons for "Bravo", "Commenter", "Republier", and "Envoyer".



The screenshot shows the GitHub community page for OpenLLM France. At the top, there's 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 titled "Claire Chat"), "Models" (3 models, sorted by "Recently updated"), and "Datasets" (2 datasets, sorted by "Recently updated"). The models listed are:

- OpenLLM-France/Claire-Mistral-7B-0.1 (Text Generation - Updated 5 days ago - 116)
- OpenLLM-France/Claire-7B-Apache-0.1 (Text Generation - Updated 5 days ago - 98)
- OpenLLM-France/Claire-7B-0.1 (Text Generation - Updated 5 days ago - 734)

The datasets listed are:

- OpenLLM-France/Claire-Dialogue-French-0.1 (Viewer - Updated about 8 hours ago - 26)

The screenshot shows the GitHub repository page for OpenLLM-France. The header includes "Sign up" and the repository name "OpenLLM-France". Below the repository name, it says "Communauté OpenLLM-France" and lists 49 followers. The main content area has a "README" section with the title "Welcome to OpenLLM-France" and a "discord" link. The text describes the aim of the community: "to collaborate on the development of a French, sovereign, and truly Open Source LLM, that would be based on" the following principles:

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

Under "Follow us:", there are links to "Discord" and "Hugging Face".

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

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