Open Weights models provide control over deployment and the possibility of fine-tuning. However, they leave a lot open to guesswork 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)
- ...

1. Web crawled data
2. English

BigScience

together.ai
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

Ismail Harrando LINAGORA Toulouse, France iharrando@linagora.com

Guokan Shang LINAGORA Paris, France gshang@linagora.com

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

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

November 2023

1 [CS.CL] 28 Nov 2023

Members of our containing discussion:

arxiv:2311.16840

Good Tech for Good – Made in France
SHARING THE DATA, CODE AND MODELS

https://github.com/OpenLLM-France

https://huggingface.co/OpenLLM-France
EVALUATION

Does the model stick to French?

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
• Parliamentry and public addresses

Some complications: OCR
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

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
CLAIRE : 1ÈRE PHASE DE NOTRE DÉMONSTRATION SCIENTIFIQUE
OpenLLM France

- Community
- AI & ML interests
- Team members
- Models
- Datasets

Space: Claire Chat

Welcome to OpenLLM-France

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://huggingface.co/OpenLLM-France

https://github.com/OpenLLM-France