OpenLLM France: Building transparent and open AI with a French twist

Julie Hunter

Commons AI

FOST 2025

About LINAGORA

Leader in open source for 25 years

Through its software and services, LINAGORA helps large public and private organizations develop technological independence.

Twake Workplace

Collaborative Suite

The only truly Open Source workplace based on major Internet standards

LinShare

Secure file sharing

Private and secure file sharing and cloud storage solution

LinTO

Voice transcription

VoiceTech technology: record, edit and transcribe your meetings

LINAGORA.AI

AI: OpenLLM, OpenRAG

Innovative approach to generative AI, combining Open Source, specialised models and secure deployment

Today

1. Why make open source LLMs?
2. Introduction to the OpenLLM France initiative
3. The OpenLLM models

Why make open source LLMs?

LLMs everywhere!

Chatbot assistance: travel planning, tutoring assistance, document querying, ...

Document generation: summaries, quizes, reformulations, ...

Numerous open weights models!

Do we need more?

Open-weights models permit fine-tuning

But:

- Fine-tuning can’t solve all problems
- We inherit bias (and lack of data transparency) from the base model
- Research is limited to the final steps of training
- Training know-how is left in the hands of a select few*

*problem exacerbated by the extreme cost (computational, data-related) of pretraining

OpenLLM France

OpenLLM France

Project funded by the BPI (09.2024 – 08.2026) and supported by GENCI

Aim to build truly open, ethical, efficient and sovereign generative AI, with a focus on the French language.

LINAGORA

opsci

TALKR.ai

Class'Code

cea

cnrs IDRIS

Loria

LX

UNIVERSITÉ PARIS 1 PANTHÉON SORBONNE

Associated Partners include

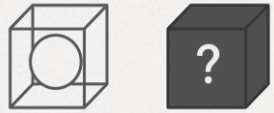
ANITI

SAINT EXUPÉRY

MENS DATA l'esprit des données

https://www.openllm-france.fr/

Truly open source



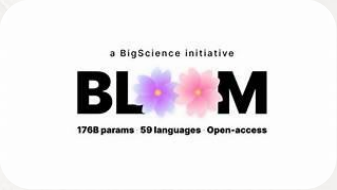
Three essential conditions:

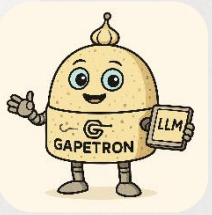
1. License places no restriction on model usage
2. Complete transparency of training methods
3. Training data made available under an open license

A few open initiatives

English focused

Multilingual/focus on French





Anglocentric training

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%

The situation is improving, but anglo-centricity remains the default at least for American models (keeping in mind that we have very few details on dataset proportions)

A question of language AND culture

- History
- Politics
- Art
- Cooking
- Ethics and law

The OpenLLM models

Lucie 7B: a pre-trained model

(Lucie-7B-Instruct-V1.1 – light instruction)

« Lucie » comes from « lux »

Lucie sheds light on the construction of generative AI

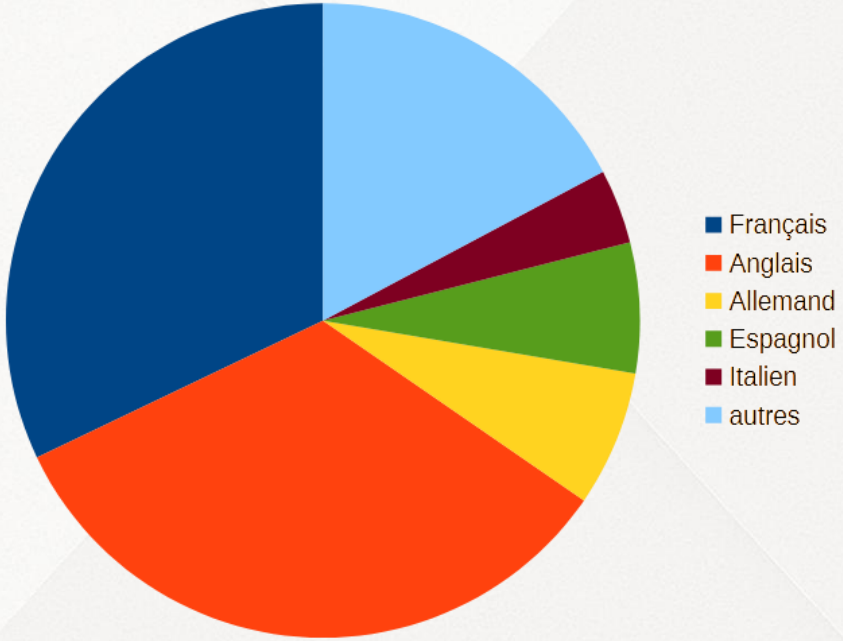
Open Source LLM

data, code, weights (final and intermediate checkpoints)

Focus on the French language

GDPR and AI-Act compliant

Representation of French



3 trillion tokens (2.3 trillion unique)

Custom tokenizer to balance representation of the five natural languages in our training data + code

New models

Still 100% open, still centered on French (30%), still GDPR and AI Act compliant

Three sizes

- 1B: edge cases, tests
- 8B (mamba-transformer hybrid): long context
- 23B: RAG and tasks that require more reasoning

Nvidia NeMo (vs Megatron-Deepspeed)

FP8 training (8B, 23B)

5T tokens (vs 3T)

New languages: pt, nl, ar

Multi-phase training: math and reasoning

Adaptation for educational use-cases

Challenge 1: web data

Very difficult to avoid – the vast majority of pretraining data

Quality: deduplication, heuristic filtering (line length, word repetition, ...), classifier-based filtering (trained on LLM annotations)

Intellectual property: filtrage by

- robots.txt : imperfect, requires retroactive application (cf. Common Crawl)? What impact for open source?
- license (Common Pile), trusted domains (Common Corpus) : severe restriction on token quantity

Challenge 2: toxicity and bias

Web data and public domain data

Argument that LLMs should see « bad » data so that they can recognize them when confronted with them by users (preference optimization)

But still…

- Blacklists, no-go words
- Classifiers trained on LLM annotations (Hugging Face, AI2, …)

Challenge 3 : French data

Difficulty of open licenses (cf. enormous effort of Common Pile)

Data in the public domain (cf. intense effort of Common Corpus) : old documents (bias!), OCR (!!!), government data (good but limited)

Very limited quantity: After selecting the highest quality web data that we can (still less good than their English counterparts), we get to:

- ~500 billion tokens in web
- ~225 billion tokens non-web (almost entirely OCR)

Difficult to imagine a « long horizon » training with a significant proportion of French data

Challenge 4: post-training data

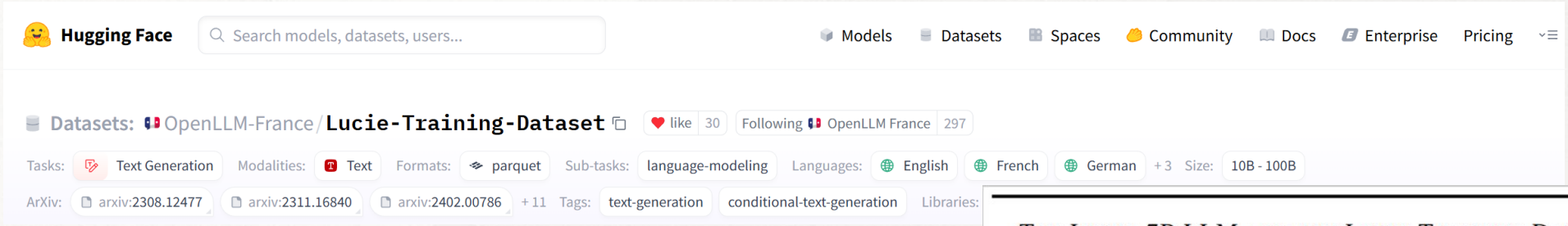
Even less open, even less French

Not easier to acquire: need for less quantity, but higher quality and specific formats human annotation is costly and LLMs bring baggage

- Bias
- Questions of openness and sovereinty: not simple

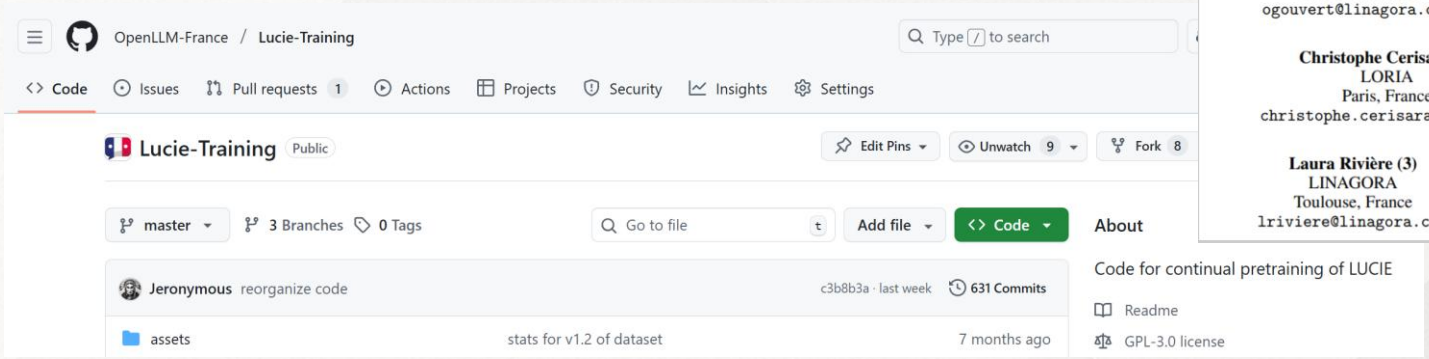
A subject under development!

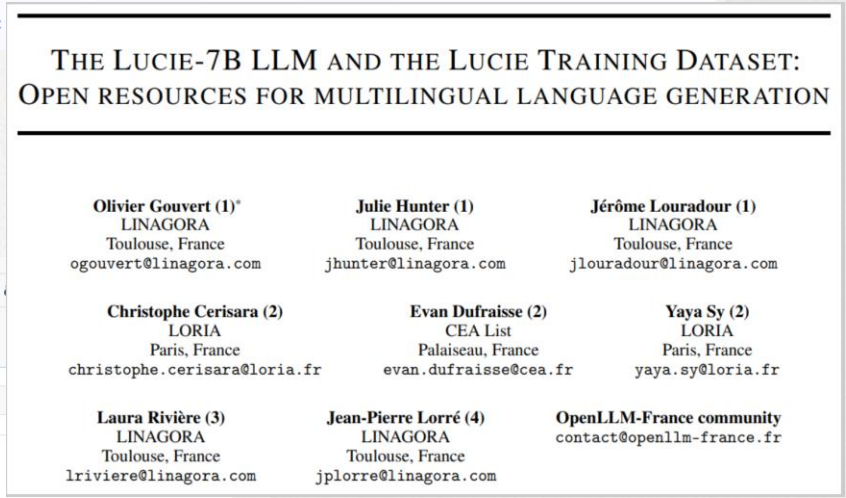
Publication of resources



https://huggingface.co/OpenLLM-France

https://github.com/orgs/OpenLLM-France/repositories





THE LUCIE-7B LLM AND THE LUCIE TRAINING DATASET: OPEN RESOURCES FOR MULTILINGUAL LANGUAGE GENERATION

Olivier Gouvert (1)*
LINAGORA
Toulouse, France
ogouvert@linagora.com

Julie Hunter (1)
LINAGORA
Toulouse, France
jhunter@linagora.com

Jérôme Louradour (1)
LINAGORA
Toulouse, France
jlouradour@linagora.com

Christophe Cerisara (2)
LORIA
Paris, France
christophe.cerisara@loria.fr

Evan Dufraisse (2)
CEA List
Palaiseau, France
evan.dufraisse@cea.fr

Yaya Sy (2)
LORIA
Paris, France
yaya.sy@loria.fr

Laura Rivière (3)
LINAGORA
Toulouse, France
lriviere@linagora.com

Jean-Pierre Lorré (4)
LINAGORA
Toulouse, France
jplorre@linagora.com

OpenLLM-France community
contact@openllm-france.fr

https://arxiv.org/abs/2503.12294

Q&A

Thank you!

Tokenization

The minimal units to give to your LLM

Words: `J’ | adore | la | chocolaterie`

Charcters : `J | ’ | a | d | o | r | e | l | a | c | h | o | c | o | l | a | t | e | r | i | e`

Sub-words : `J ' | ad | ore | la | ch | ocol | ater | ie`

Good tokenization means better performance and lower cost!

Trained on a subset of the data used to train the LLM