LINAGORA

Open LLM
Europe

Lucie 7B

cnrs GDR Groupement de recherche
TAL Traitement automatique des langues

Jean-Pierre Lorré

13/ 03 / 2025

Introduction

Jean-Pierre Lorré
Directeur de recherche
jplorre@linagora.com

Agenda

- Context: why Lucie 7B ?
- Lucie 7B details: training & evaluation
- Future work

LINAGORA

Lucie-7B LLM Context

LINAGORA

Leader in Open Source

#GOODTECHFORGOOD

100% Open Source

GAFAM-free

Founded in **2000**
160 employees
6 offices worldwide

Research topics: NLP, Speech Recognition

LE VOICE LAB

ANITI

LinTO.ai

LinTO STUDIO

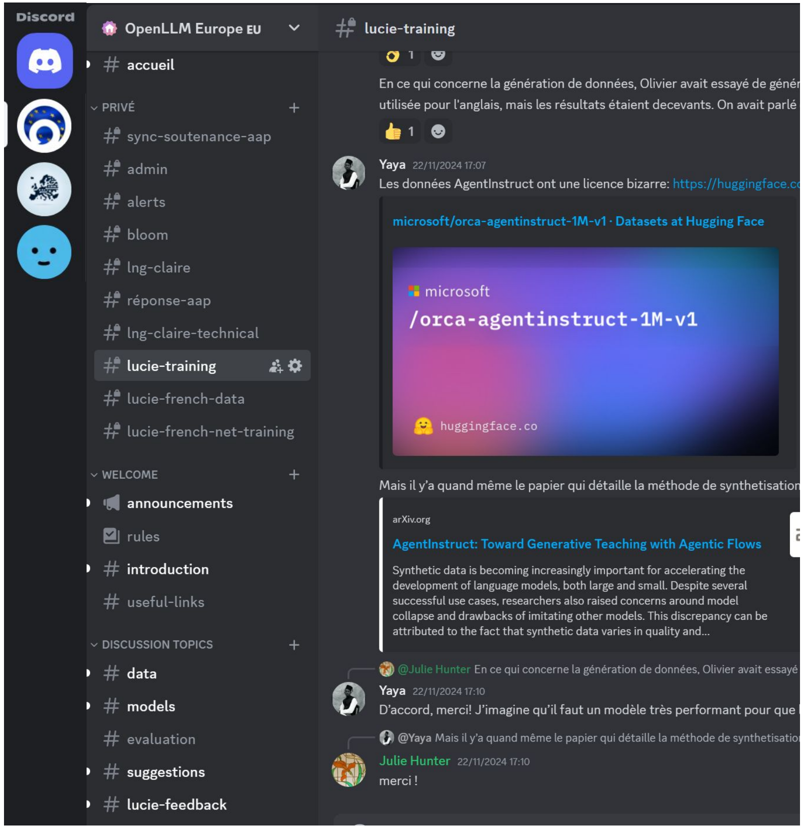
ASR & NLP platform

OpenLLM-Europe

Community for the development of sovereign, and truly Open Source LLM
> 1 100 members

👨‍👩‍👧‍👦 The OpenLLM Community 🌍

Discord (1 100+ members)



En ce qui concerne la génération de données, Olivier avait essayé de génér
utilisée pour l'anglais, mais les résultats étaient decevants. On avait parlé

Yaya 22/11/2024 17:07
Les données AgentInstruct ont une licence bizarre: https://huggingface.c

microsoft/orca-agentinstruct-1M-v1 · Datasets at Hugging Face

microsoft
/orca-agentinstruct-1M-v1
huggingface.co

Mais il y'a quand même le papier qui détaille la méthode de synthetisation

arXiv.org
AgentInstruct: Toward Generative Teaching with Agentic Flows
Synthetic data is becoming increasingly important for accelerating the development of language models, both large and small. Despite several successful use cases, researchers also raised concerns around model collapse and drawbacks of imitating other models. This discrepancy can be attributed to the fact that synthetic data varies in quality and...

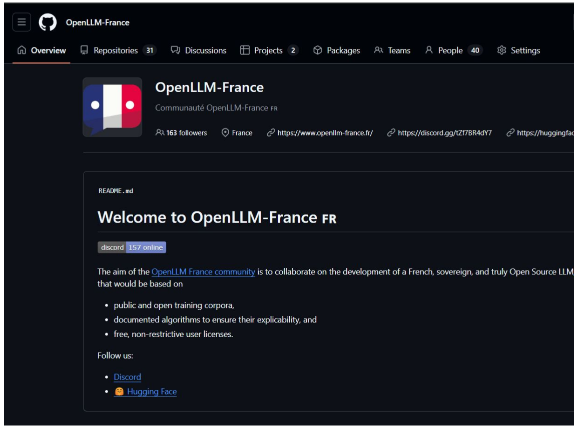
@Julie Hunter En ce qui concerne la génération de données, Olivier avait essayé
Yaya 22/11/2024 17:10
D'accord, merci! J'imagine qu'il faut un modèle très performant pour que

@Yaya Mais il y'a quand même le papier qui détaille la méthode de synthetisatio
Julie Hunter 22/11/2024 17:10
merci !

https://discord.gg/VFJxQnqrEU

Hugging Face (100+ members)

GitHub (40 dev, 100+ followers)



OpenLLM-France
Communauté OpenLLM-France FR
163 followers · France · https://www.openllm-france.fr/ · https://discord.gg/tZf7BR4dY7 · https://huggingfac

README.md

Welcome to OpenLLM-France FR

discord 157 online

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.

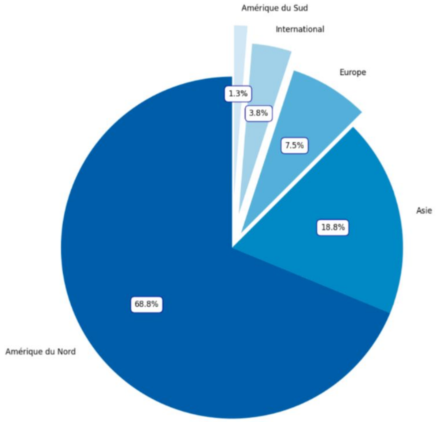
Follow us:

- Discord
- 🤗 Hugging Face

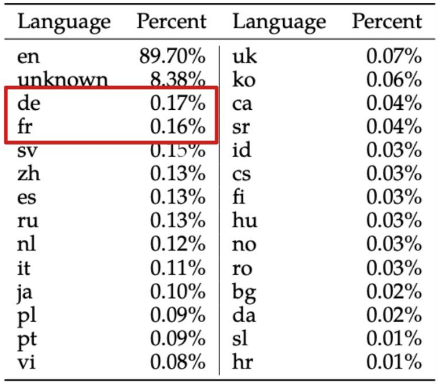
Motivation — Cultural Representation

Better representation of French and French-speaking communities

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%

Not just language:

- History
- Politics
- Art
- Religion
- Social practices
- Cooking …

Llama 3: pretrained on 15T tokens and 5% non-English data

Motivation – Truly Open Source AI

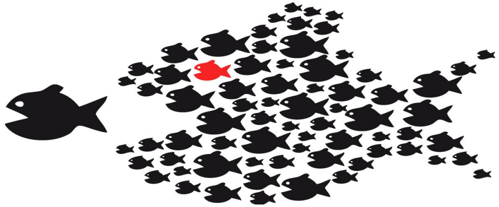
Open Source AI Definition

open source initiative®

4 Freedoms	Open Weights	Open Code	Data
Use Study Modify Share	Model weights and (hyper-) parameters	Source code used to train the system Source code used to create the dataset	The complete list of datasets used to train the system **and** the actual datasets when allowed

A license that allows unrestricted usage

Community impact



Open LLM Europe

Motivation – Small Language Models

Small, specialised models can achieve comparable or even better performance than large, general-purpose models on given tasks

Smaller models are more resource efficient at both training and inference times

Small, specialised models can be combined with other AI or non-AI models to develop complex applications

OpenLLM-France Project

APPEL À PROJETS

COMMUNS NUMÉRIQUES POUR L'INTELLIGENCE ARTIFICIELLE GÉNÉRATIVE

bpifrance FRANCE 2030 GOUVERNEMENT

FRANCE 2030

T0: 01/09/24

2 years

10.5 M€

Develop multimodal, voice and text LLM models that are trusted, controlled and transparent

Focus on the education application domain

Taking account of ethical, legal and environmental aspects

LINAGORA

cea

opsci.ai

TALKR.ai

Class'Code

UNIVERSITÉ PARIS 1 PANTHÉON SORBONNE

cnrs IDRIS

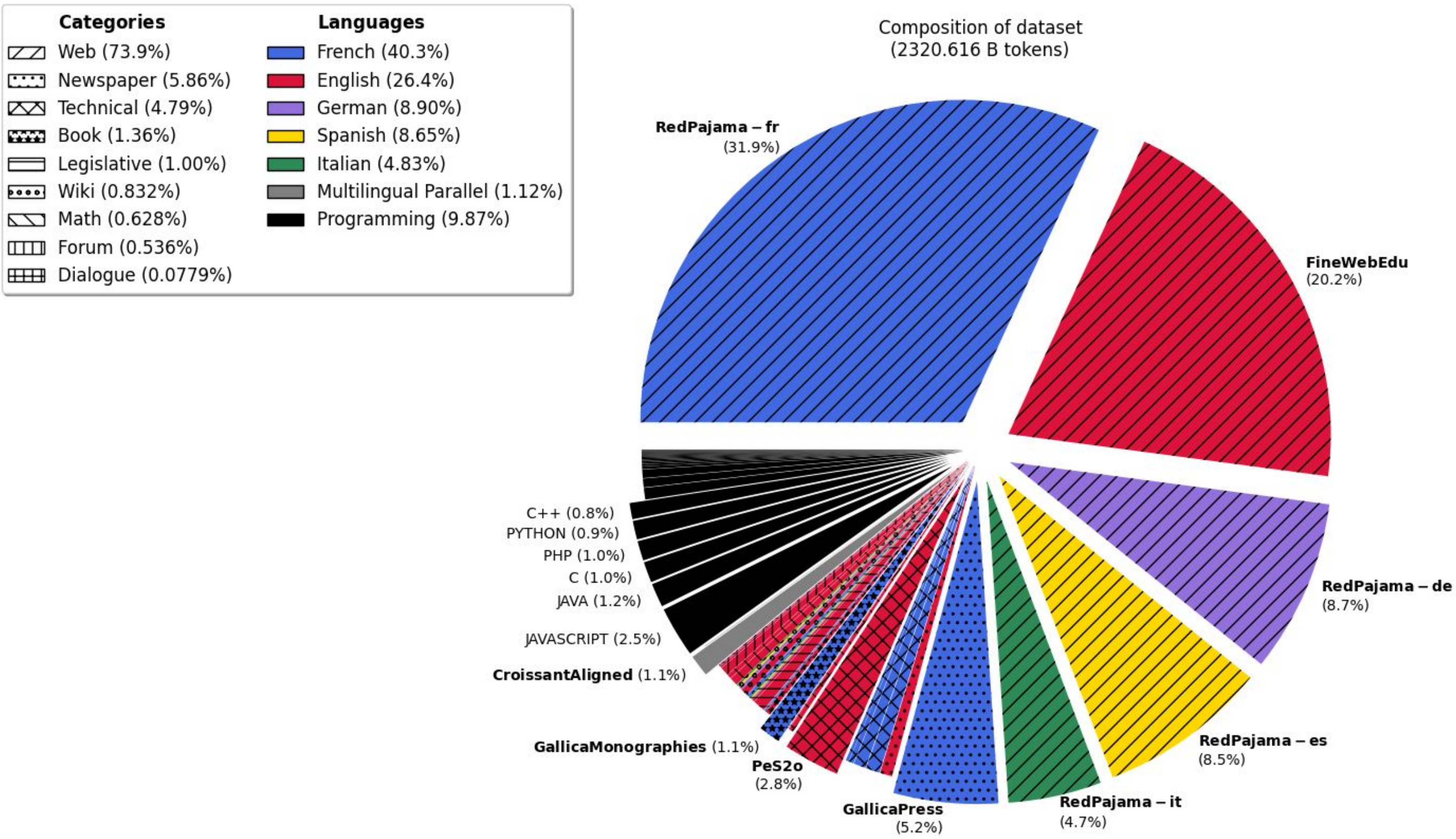
Loria Laboratoire lorrain de recherche en informatique et ses applications

LIX

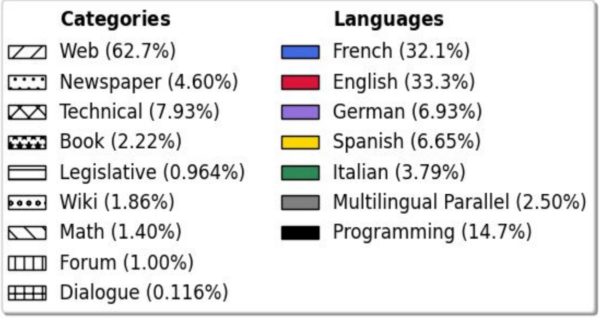
Lucie-7B LLM

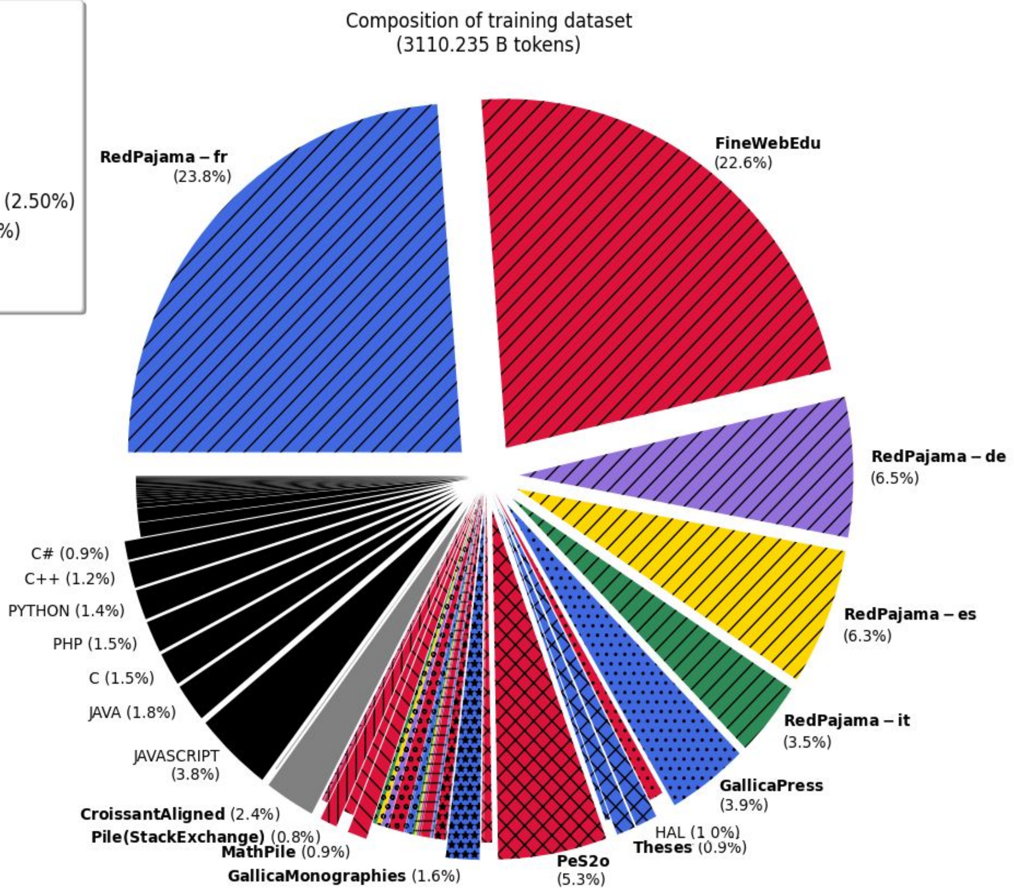
Lucie Training Dataset

OpenLLM-France/Lucie-Training-Dataset



Lucie Training Dataset - Data mix for Lucie pretraining



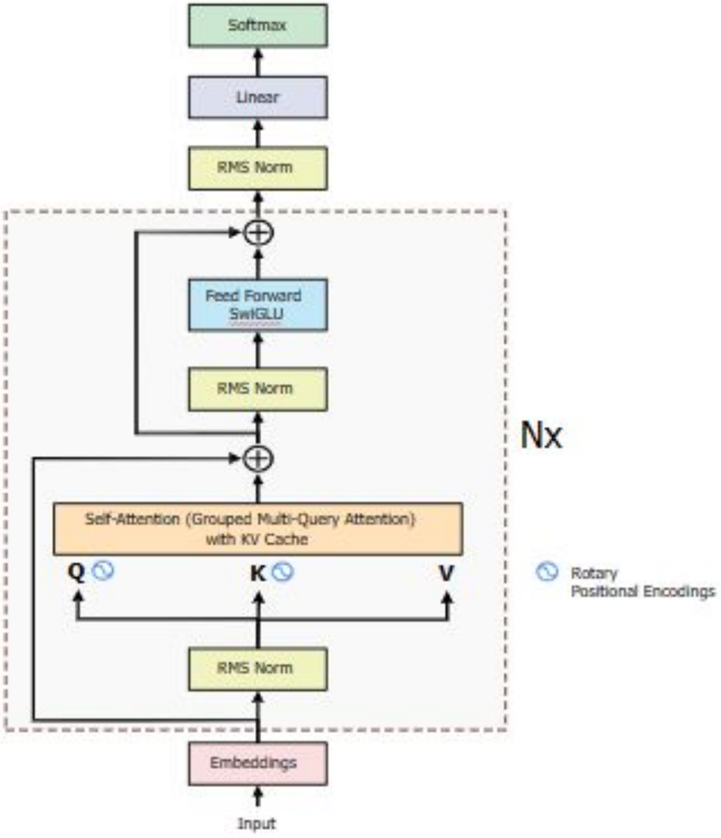


- Upsampling of English and higher quality data sets
- Final proportions:
  - French 40% → **33%**
  - English 26% → **33%**
  - Web data 74% → **63%**
  - Code 10% → **15%**

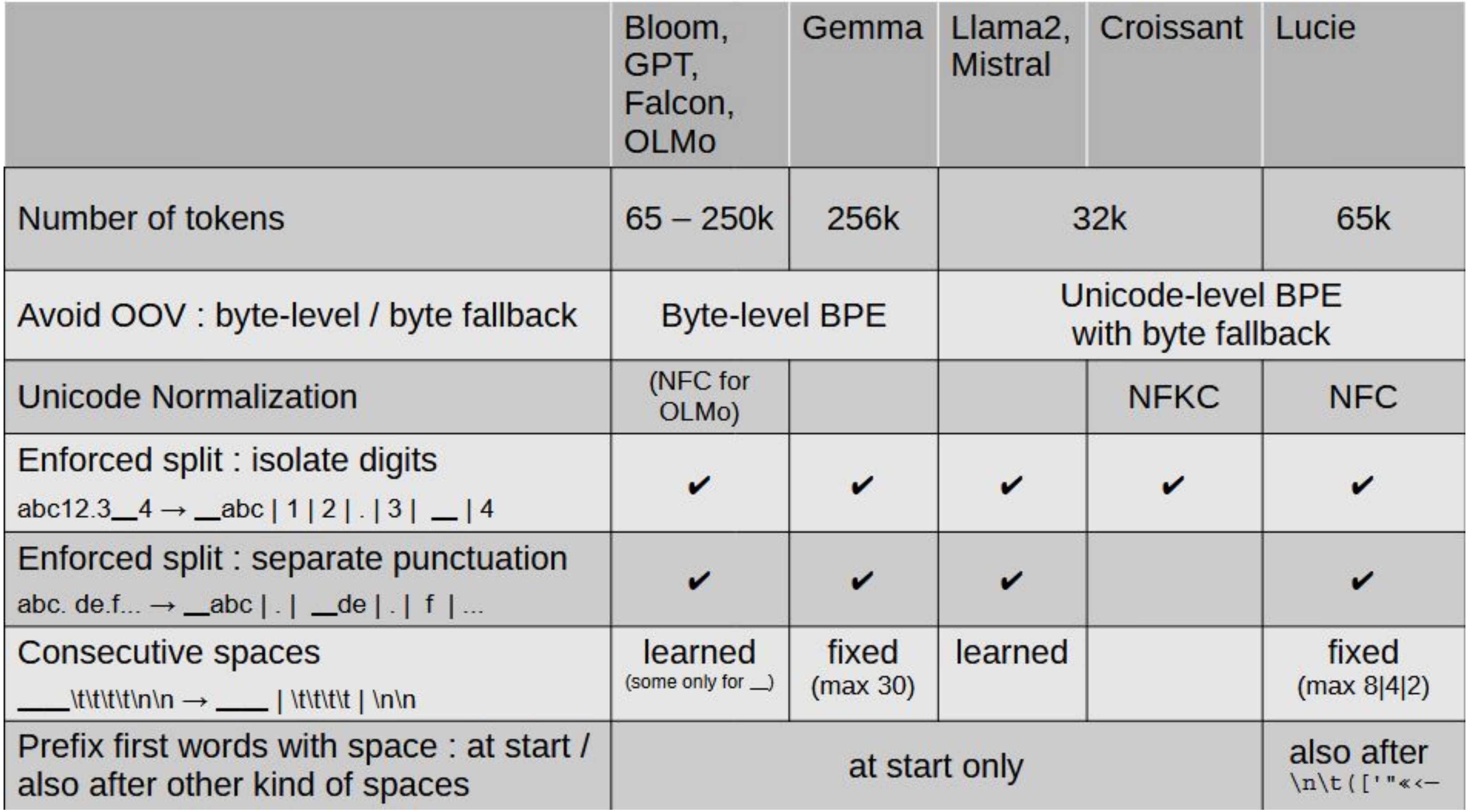
🤖 Architecture

- Causal decoder-only model. Next-token prediction.
- Llama 3 architecture
  - Group Query Attention
  - Rotary Positional Embedding (RoPE)
  - Configuration:
    - Vocabulary size: 65 k tokens
    - Layers: 32
    - Hidden size: 4096
    - Context length: 4096 (étendu ensuite à 32k)
- Most of the weights lie in the Feed-Forward Networks

a.	Embedding:	266 M	(x2)	→	0.53 B
b.	Attention block:	42 M	(x32)	→	1.34 B
c.	FFN block:	151 M	(x32)	→	<u>4.83 B</u>
				Total	6.70 B


Softmax
Linear
RMS Norm
Feed Forward SwiGLU
RMS Norm
Nx
Self-Attention (Grouped Multi-Query Attention) with KV Cache
Q
K
V
Rotary Positional Encodings
RMS Norm
Embeddings
Input

🔢 Tokenization – Text Pre-Processing Options and Constraints on Tokens

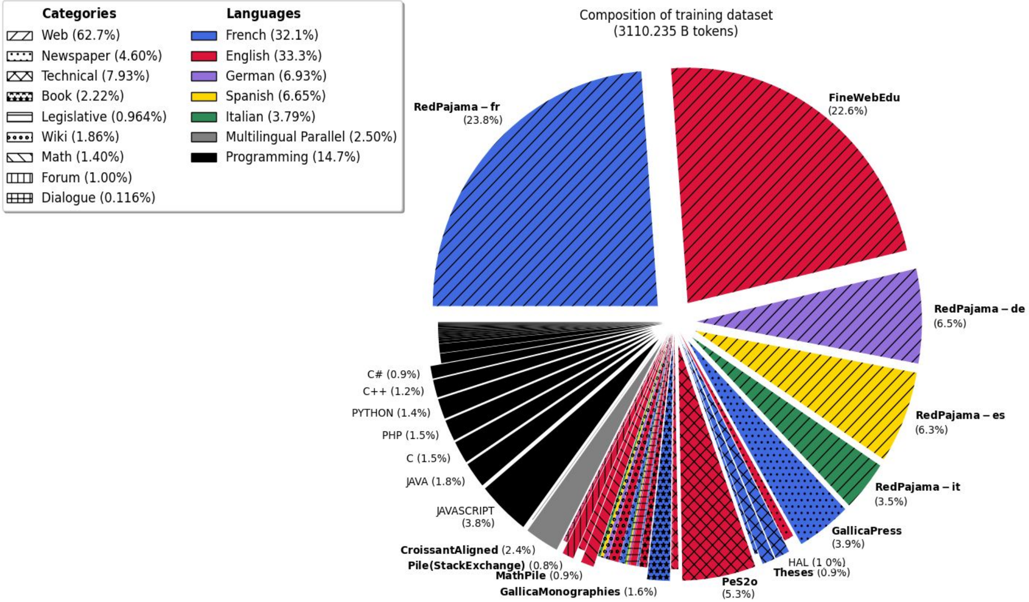


	Bloom, GPT, Falcon, OLMo	Gemma	Llama2, Mistral	Croissant	Lucie
Number of tokens	65 – 250k	256k	32k		65k
Avoid OOV : byte-level / byte fallback	Byte-level BPE		Unicode-level BPE with byte fallback		
Unicode Normalization	(NFC for OLMo)			NFKC	NFC
Enforced split : isolate digits abc12.3_4 → _abc | 1 | 2 | . | 3 | _ | 4	✔	✔	✔	✔	✔
Enforced split : separate punctuation abc. de.f... → _abc | . | _de | . | f | ...	✔	✔	✔		✔
Consecutive spaces ____\t\t\t\t\n\n → ____ | \t\t\t\t | \n\n	learned (some only for _)	fixed (max 30)	learned		fixed (max 8|4|2)
Prefix first words with space : at start / also after other kind of spaces	at start only				also after `\n\t(['"«‹—`

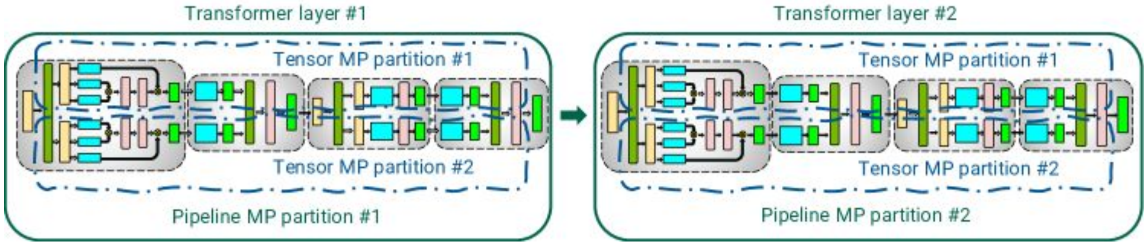
Training Pipeline

Three pre-training phases:

- Main pre-training phase
  - 3.1T tokens
  - Knowledge of the world acquisition
- Context extension phase
  - 5B tokens
  - Extend the context length from 4096 to 32k tokens
- Annealing phase
  - 5B tokens
  - High-quality dataset with a focus on mathematical content



Parallel Training



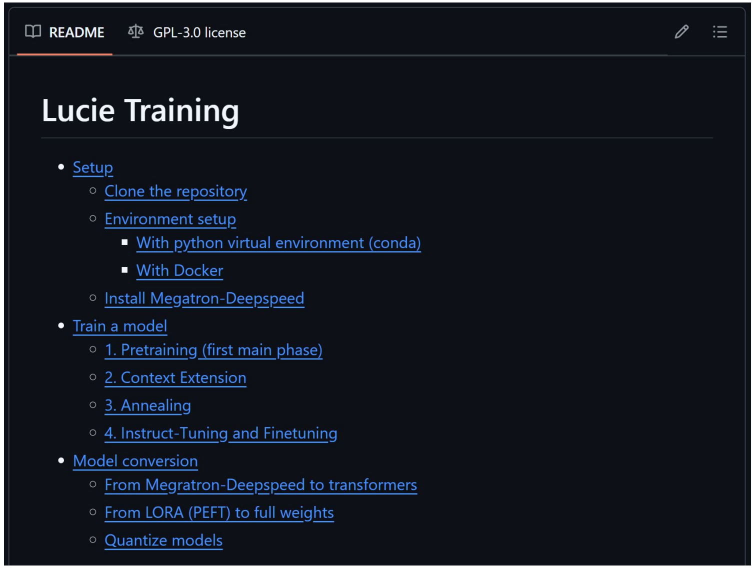
Lucie-7B was pre-trained on:

- 512 80GB-VRAM H100
- for about 500k GPU hours

The training code is based on a fork of Megatron-DeepSpeed

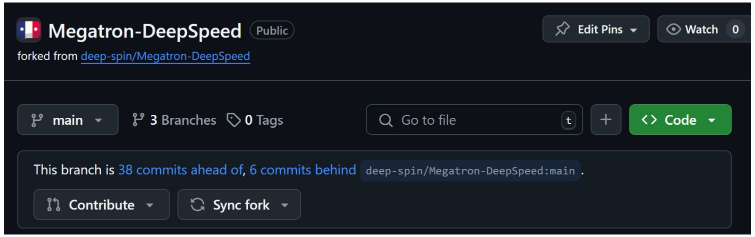
3D Parallelism:

- Data Parallelism	— 32
- Pipeline Parallelism	— 4
- Tensor Parallelism	— 4
- Batch size	~ 4M tokens



Lucie Training

- Setup
  - Clone the repository
  - Environment setup
    - With python virtual environment (conda)
    - With Docker
  - Install Megatron-Deepspeed
- Train a model
  - 1. Pretraining (first main phase)
  - 2. Context Extension
  - 3. Annealing
  - 4. Instruct-Tuning and Finetuning
- Model conversion
  - From Megratron-Deepspeed to transformers
  - From LORA (PEFT) to full weights
  - Quantize models

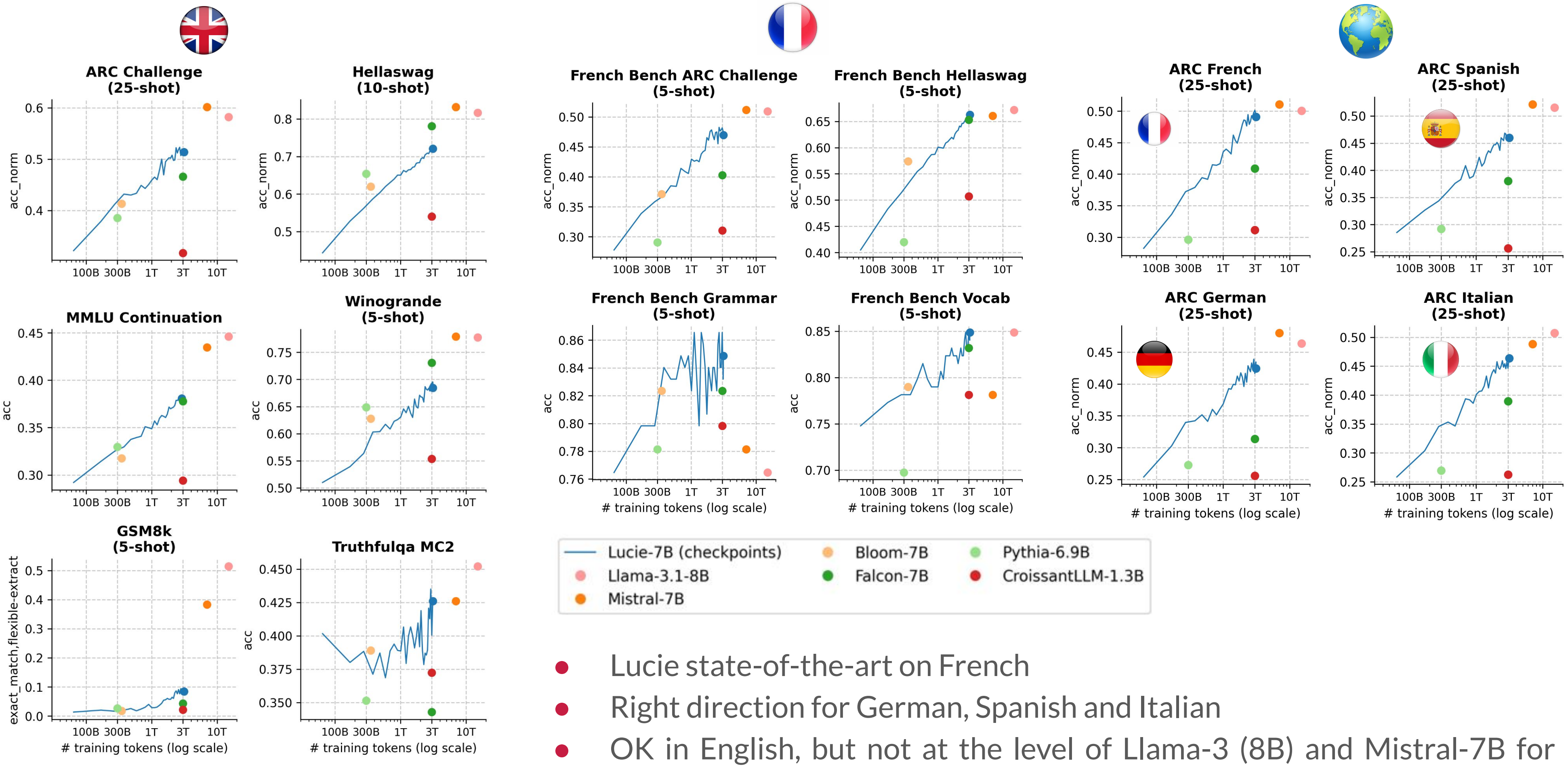


Megatron-DeepSpeed

forked from deep-spin/Megatron-DeepSpeed

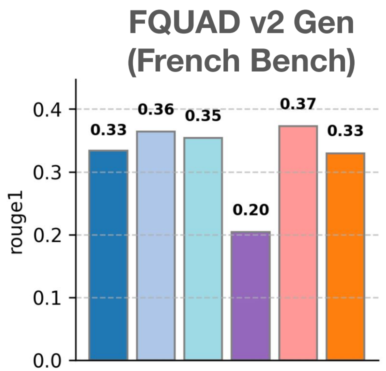
This branch is 38 commits ahead of, 6 commits behind deep-spin/Megatron-DeepSpeed:main.

📈 Learning Curves & Benchmark Evaluations

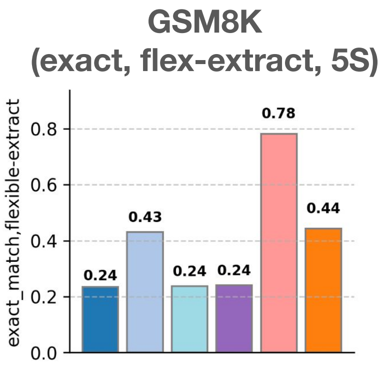


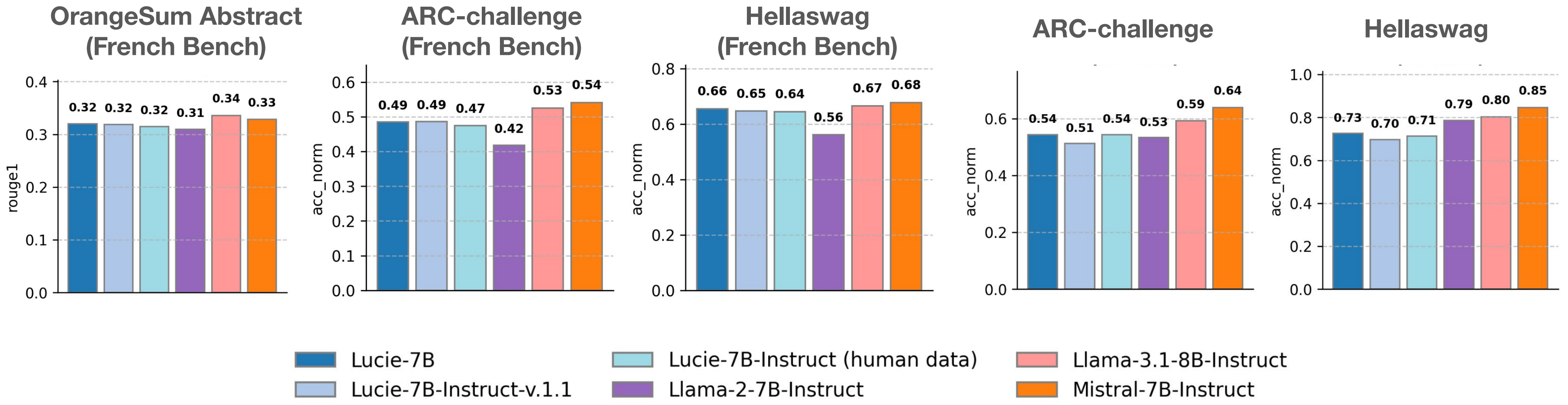
- Lucie state-of-the-art on French
- Right direction for German, Spanish and Italian
- OK in English, but not at the level of Llama-3 (8B) and Mistral-7B for Multitask Language Understanding (MMLU, Winogrande, GSM8k)

📈 Instruction Tuning (the start...)



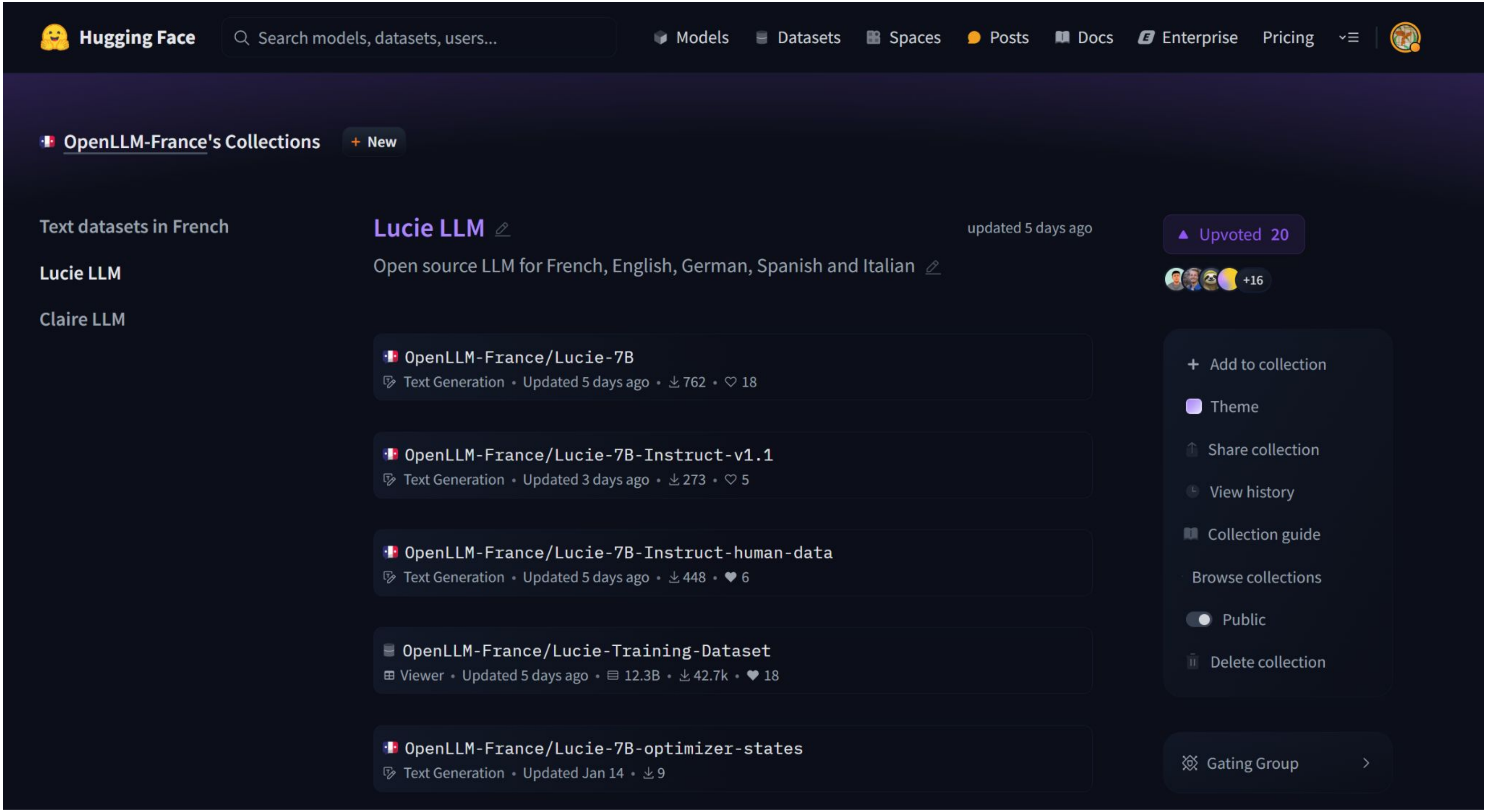
- Math
  - Beats Llama 2, competitive with Mistral Instruct
  - Falls short of Llama 3.1
- Language-dependence
  - Competitive on French benchmarks
  - Less so when the benchmarks are in English





🏛 Shared & Open Resources

https://github.com/OpenLLM-France/Lucie-Training



OpenLLM-France's Collections + New

Text datasets in French

Lucie LLM

Claire LLM

Lucie LLM

updated 5 days ago

Open source LLM for French, English, German, Spanish and Italian

▲ Upvoted 20

+16

- OpenLLM-France/Lucie-7B
  Text Generation • Updated 5 days ago • 762 • ♡ 18
- OpenLLM-France/Lucie-7B-Instruct-v1.1
  Text Generation • Updated 3 days ago • 273 • ♡ 5
- OpenLLM-France/Lucie-7B-Instruct-human-data
  Text Generation • Updated 5 days ago • 448 • ♥ 6
- OpenLLM-France/Lucie-Training-Dataset
  Viewer • Updated 5 days ago • 12.3B • 42.7k • ♥ 18
- OpenLLM-France/Lucie-7B-optimizer-states
  Text Generation • Updated Jan 14 • 9

Future work

Future Work

- Model Alignement
  - Model for Education (OpenLLM project just started – kick-off was 21/01/2025)
  - Propose open test platform to get community feedbacks
  - Start improvement loop with Reinforcement Learning (GRPO …)
- Reasoning
  - Function Calling (for math, physics, …) to calculators and API
  - **R**etrieval-**A**ugmented **G**eneration
- Multi-modality (Text prompt + Audio [+ Image/Video])
- Scale to more languages & alphabets (Greek, …), handle code switching and multilingual inputs
- Smaller model (1B) – Distillation and/or Training from scratch
- Data mix improvement (quantity, quality, nature)
- New Architectures
  - MAMBA (more linear, more efficient)
  - Hybrid Transformers(Attention) / RNN(LSTM) – TITAN

LINAGORA

Open LLM Europe

MERCI

Discord
OpenLLM-Europe

Jean-Pierre Lorré
jplorre@linagora.com
+33 6 88 34 63 85