Langage, IA générative et robotique

ANITI Days 2024

25 novembre 2024

ANITI

Towards more open LLM pretraining

Julie Hunter, LINAGORA

November 25, 2024

ANITI

① Some questions about LLM training

② Concerns about openness

③ Model pretraining at LINAGORA/OpenLLM

④ Next steps in ANITI

Some concerns (academic and industrial) about LLMs

They’re (very) large.

They’re black holes for (personal) data.

They’re often opaque and hard to train.

Can task-specific LLMs minimize these concerns?

Code, document understanding, classroom assistance, cobots, …

? How do different kinds of data impact performance on different tasks?

(Even easier to wonder once you start looking at the data 😱)

Can externalizing (certain) knowledge help?

- RAG
- External modules (ALMs)

What knowledge can be externalized?

? To what extent can we separate language modeling capacities from the capacity to retain and exploit world knowledge?

ANITI

Hurdles to scientific study

Pretraining is crucial. We need ablation studies. But:

- hardware
- scaling questions
- quantity of pretraining data
- preprocessing and annotation
- conception of evaluation
- …

Open collaboration

Opening up the process is crucial for getting smaller actors involved.

Data!

- collection and preparation practices
- pretraining, fine-tuning, evaluation
- the most important factor, and the most difficult to get

Code: we can’t keep reinventing the (engineering) wheel in our own corners

ANITI

Open collaboration

BigScience

EleutherAI

CroissantLLM

AI2 Allen Institute for AI

TII Technology Innovation Institute

⚠ **Data !** ⚠

Meta

G

IBM®

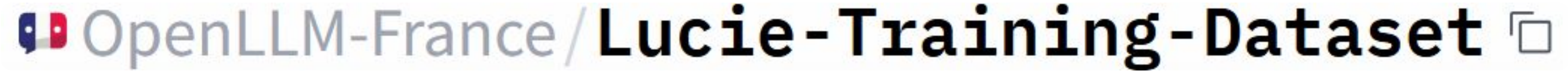
Qwen

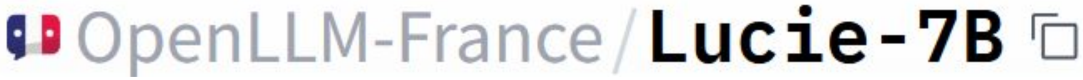
ANITI

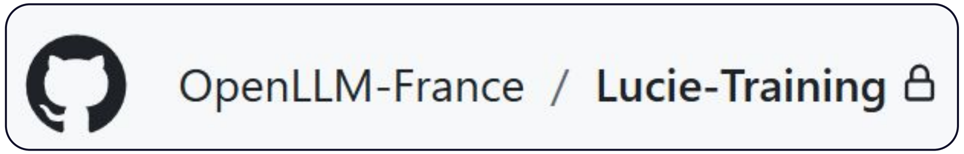
Lucie

- Causal decoder-only model ⬦ 7B parameters
- 512 H100 80GB GPUs ➜ ~550,000 GPU hrs on Jean Zay
- ETA for Lucie Instruct: mid-December

❖ **Open data** ❖ **Open weights (+ checkpoints)** ❖ **Open code**


OpenLLM-France/Lucie-Training-Dataset


OpenLLM-France/Lucie-7B

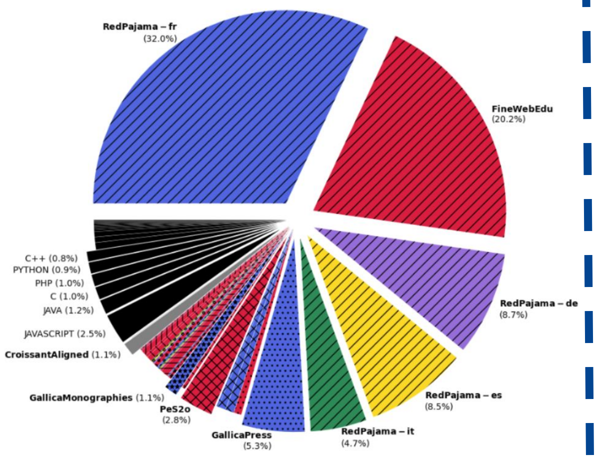

OpenLLM-France / Lucie-Training

ANITI

Balancing French and English

Composition of **original** training dataset (2320.616 B tokens)

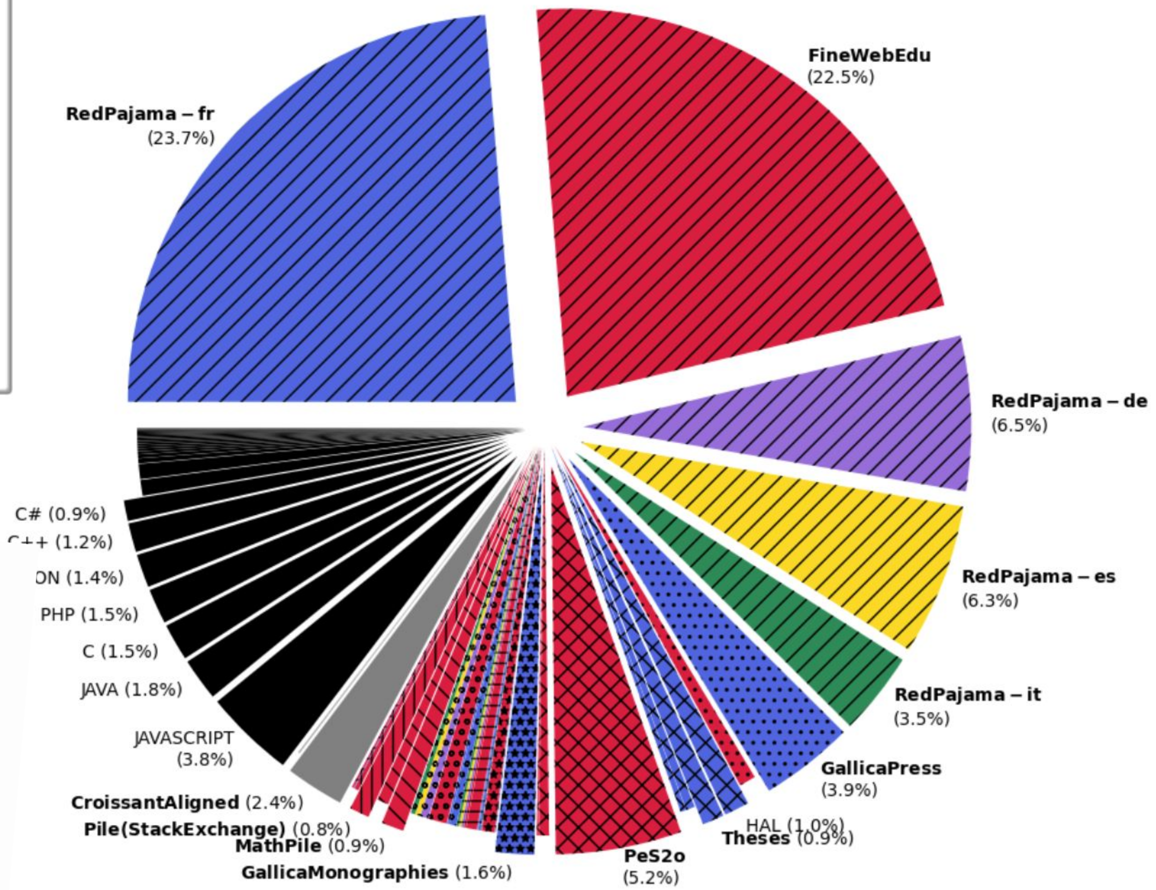
🟦 French 🟥 English
🟪 German 🟨 Spanish
🟩 Italian



Categories	Languages
▨ Web (73.9%)	🟦 French (40.3%)
⬚ Newspaper (5.86%)	🟥 English (26.4%)
⊠ Technical (4.79%)	🟪 German (8.90%)
★ Book (1.36%)	🟨 Spanish (8.65%)
▤ Legislative (1.00%)	🟩 Italian (4.83%)
● Wiki (0.832%)	
▧ Math (0.628%)	
▥ Forum (0.536%)	
▦ Dialogue (0.0779%)	
■ Multilingual (1.12%)	
■ Programming (9.87%)	

weighting by language and category

Composition of **final** training dataset (3121.743 B tokens)



ANITI

OpenLLM France

An open approach leads to a heavy bias towards certain types of data

LLMs for assisting students of French

- targeted instruction tuning
- task-specific evaluation
  - factuality
  - French
  - cultural norms
- exploiting RAG approaches

ANITI

C3PO 🤖

Impact of data proportions and training approaches on model performance

- multilinguality (but not from a low-resource perspective)
- factuality (education) - overlap with Airbus

Starting points

- Lucie data
- Small English models (OLMO, Pythia)
- Lucie checkpoints - change data mixtures at the end

Multimodality: strengthening the language component

ANITI

Thank you!