GOSIM

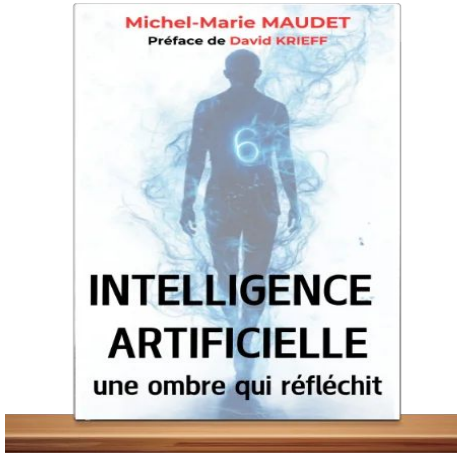
GOSIM AI Paris 2025

LUCIE : From Scratch with Love !

For a trusted and
100% Open Source Generative AI

Who I am
Mission to create a 3rd way with LINAGORA

GOSIM



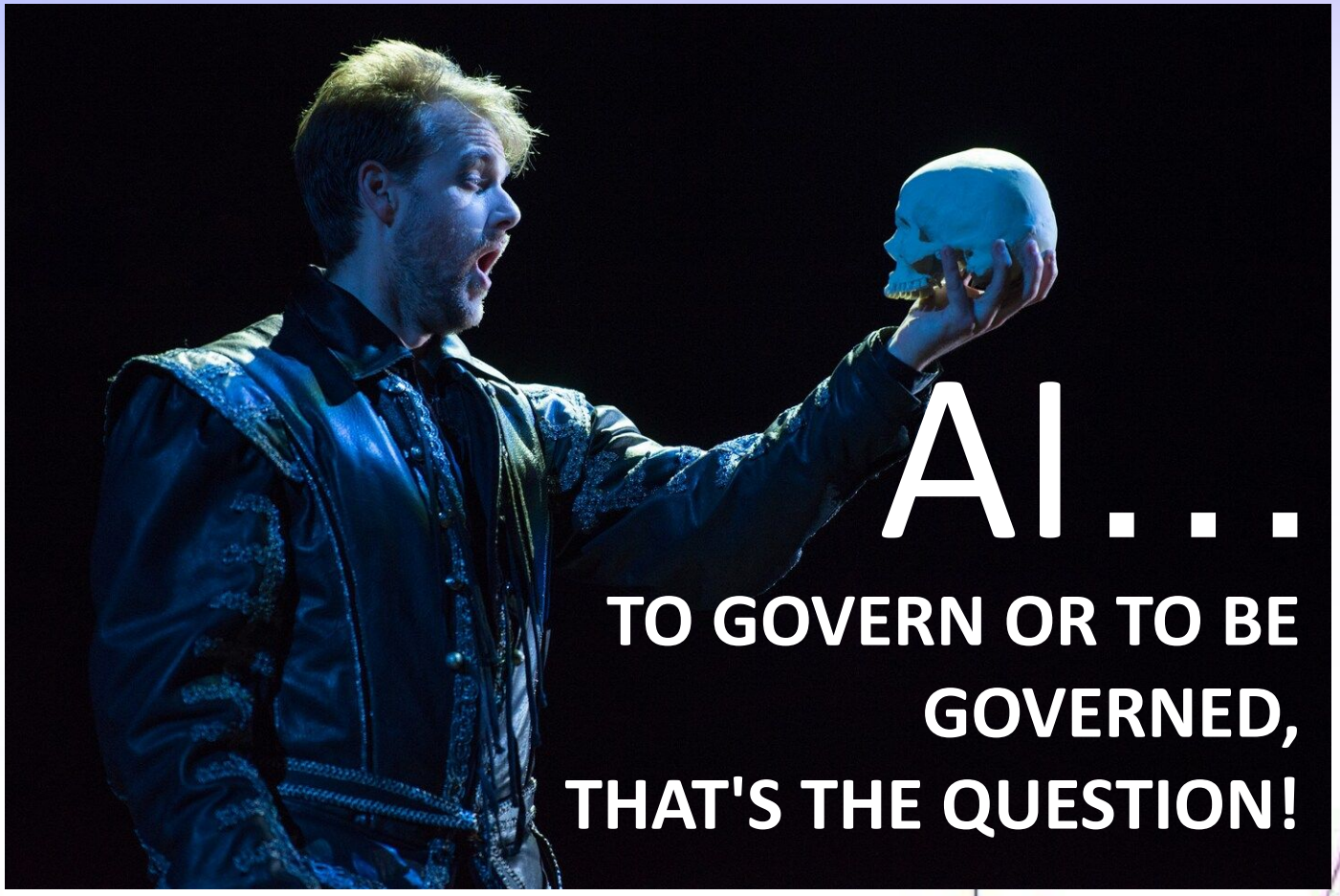
#GoodTechForGood
fantastic-enterprise.com

3rd Digital Way

“Good Tech For Good”

Ethical, responsible, sustainable and Open Source technologies to make the world a better place, with maximum positive impact on people, society and the planet.

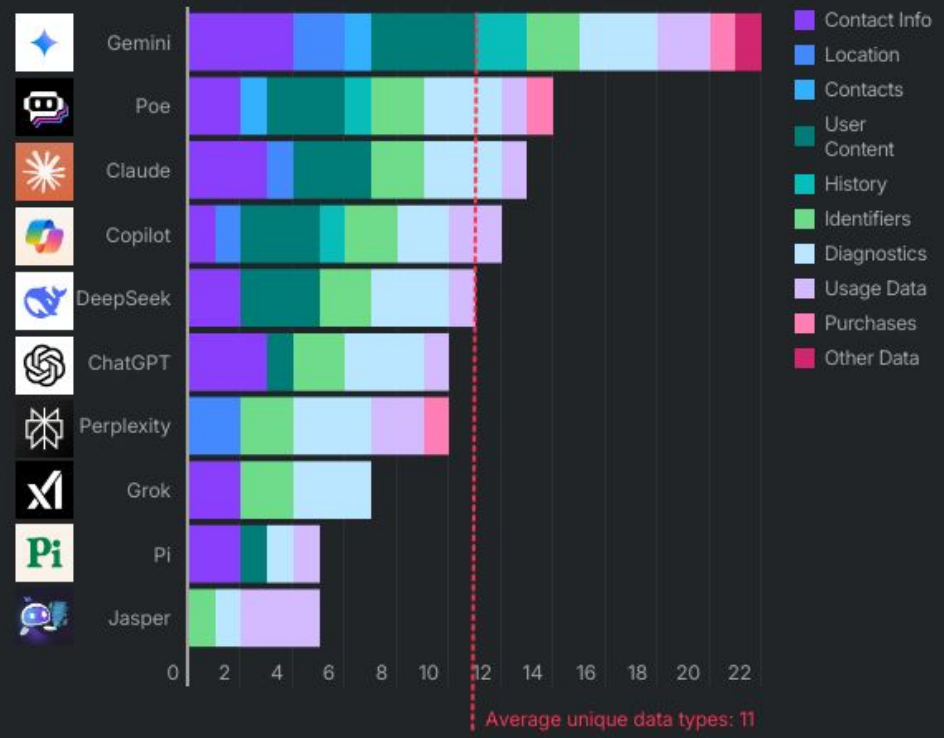
GOSIM AI Paris 2025


AI...
TO GOVERN OR TO BE GOVERNED,
THAT'S THE QUESTION!

DATA COLLECTED: February 12, 2025

40% of popular AI chatbots collect user location

Google Gemini collects the most user data among AI chatbots, gathering 22 out of 35 types of data



Surfshark®

AI...

When it's free

YOU'RE THE PRODUCT !

Overview of available language models

GOSIM

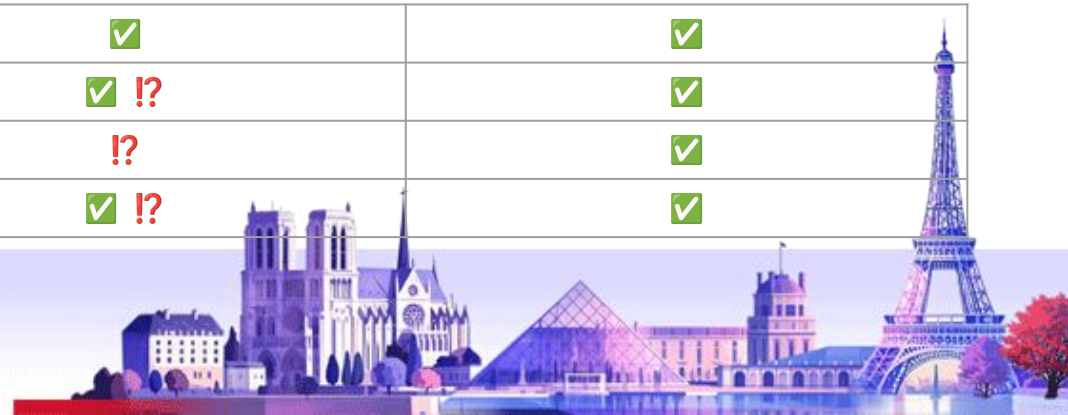
Modèles commerciaux et fermés	Models Open Weights	Models 100% Open Source
ChatGPT	LLaMA by Meta	LUCIE
Claude 3.5 Sonnet	MISTRAL AI_	Hugging Face Smol Language Model
Grok	deepseek	BLOOM

Openness and autonomy

	Modèles commerciaux et fermés	Models Open Weights	Models 100% Open Source
License to use	Commercial ❌❌❌	Gratuite mais restrictions possibles ✅ ⁉	Open Source ✅
Source code availability / “Secret Sauce”	NO ❌❌❌	Partial ✅ ⁉	YES ✅
Access to training data	NO ❌❌❌	NO ❌❌❌	YES ✅

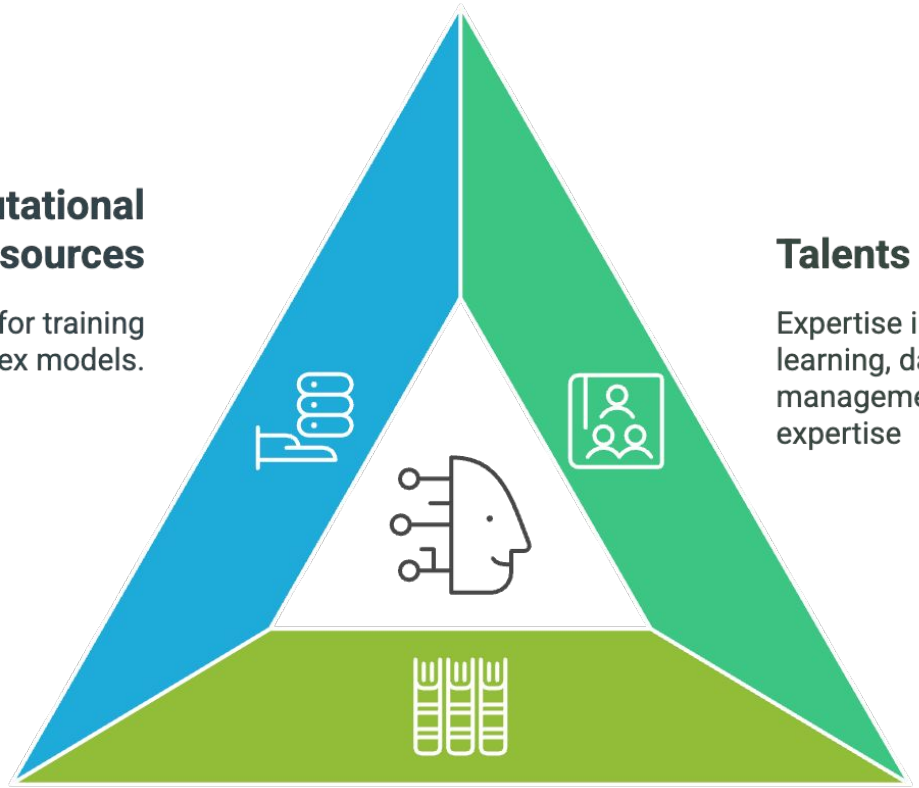
Features

Model reproduction/adaptation capability	❌❌❌	✅	✅
Choice of inference platform	❌❌❌	✅ ⁉	✅
FR and EU values taken into account	⁉	⁉	✅
Environmental impact	❌❌❌	✅ ⁉	✅



3 key components for training an LLM

GOSIM



Computational Resources

GPU hours for training complex models.

Talents

Expertise in machine learning, data management, legal expertise

Data Quality

High-quality open data with accuracy and reliability.

OpenLLM France, selected in the France 2030 call for projects "Digital commons for GenAI" call for projects

Build GenAI digital commons with trust, antonomy and transparency

GOSIM

- ➔ Development of multimodal text/voice foundation models
- ➔ Focus on French, then on European languages
- ➔ Development of an appropriate evaluation methodology
- ➔ Consideration of ethical, legal and environmental aspects
- ➔ Specialized, sober and compact models (< 24B) to limit the use of GPUs
- ➔ On the way of the revolution in autonomous AI agents (Agentic AI systems)

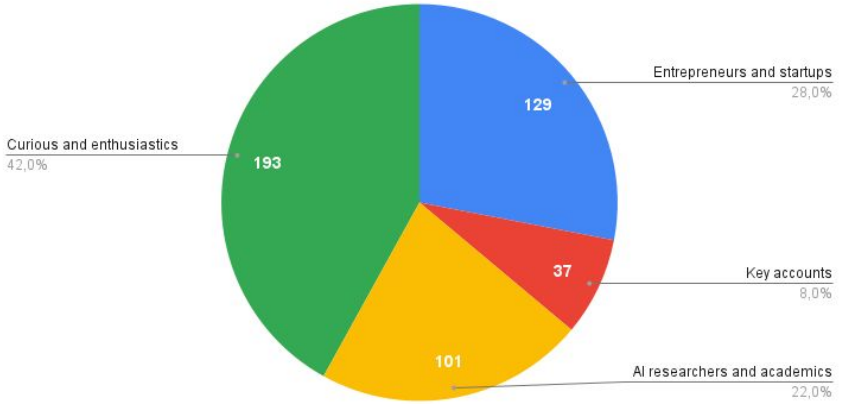
opsci TALKR.ai Class'Code cea

cnrs IDRIS Loria LX UNIVERSITÉ PARIS 1 PANTHÉON SORBONNE

OpenLLM Europe - the largest French-speaking community on Generative AI

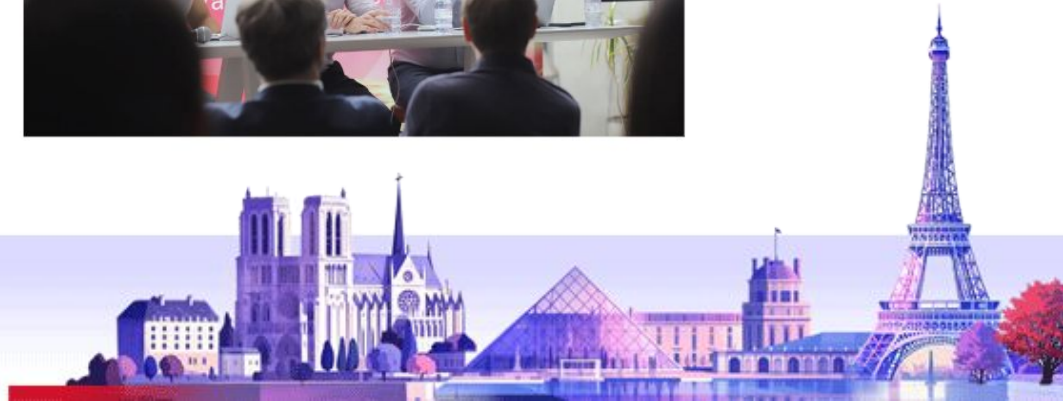
GOSIM

+1 100 researchers/engineers! (March 2025)



GENCI · IDRIS · WIKIMEDIA FRANCE · Loria · list CEA tech · ALMAnaCH Équipe-projet Inria · LX · Class'Code · UNIVERSITÉ PARIS I PANTHÉON SORBONNE · LE VOICE LAB · HUB FRANCE IA

LINAGORA.AI · opsci · THALES Together • Safer • Everywhere · FARADAY · TALKR.ai · OUTSCALE · Crédit Mutuel ARKEA · SNCF CONNECT & TECH · Giskard · Ordalie · EDF



GOSIM AI Paris 2025

LUCIE development stage

October 2023 ---------- **August 2024** ---------- **December 2024**

6 months	2 months	2,5 months	4 months	1 months
Data collection and preparation of training dataset	French-optimized tokernizer training	Additional collection on the Internet database	Model training on Jean ZAY	Alignment for instruction according to intended use
600B tokens collected, equivalent to 24 million hours of non-stop conversations, or around 8,300 years.		2.4T additional tokens added (RedPajama v2 filtered, FineWebEdu, TheStack)		Model LUCIE-7B with a version 0 instruction Ready for fine-tuning

LINAGORA's definition of truly Open Source AI

GOSIM

Three essential conditions:

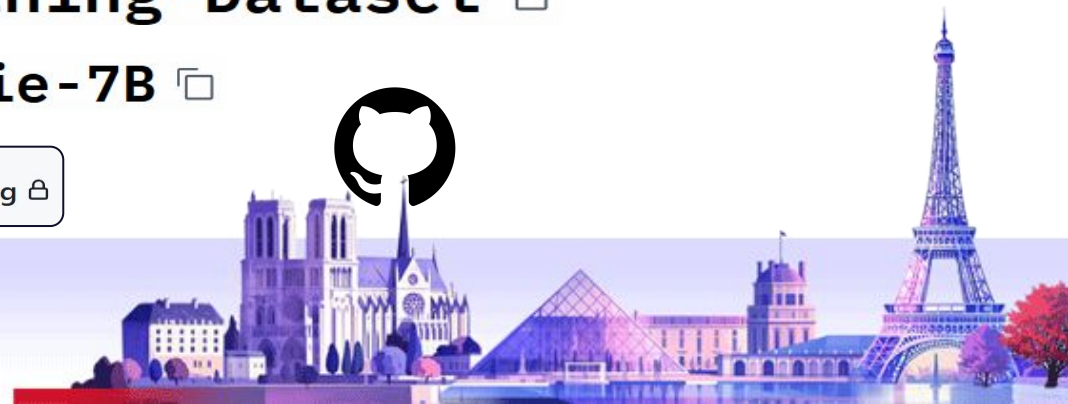
1. License to use the model without restrictions
2. Total transparency on training methods ("secret sauce" published openly)
3. **Free license for training data**

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

OpenLLM-France / **Lucie-Training-Dataset**

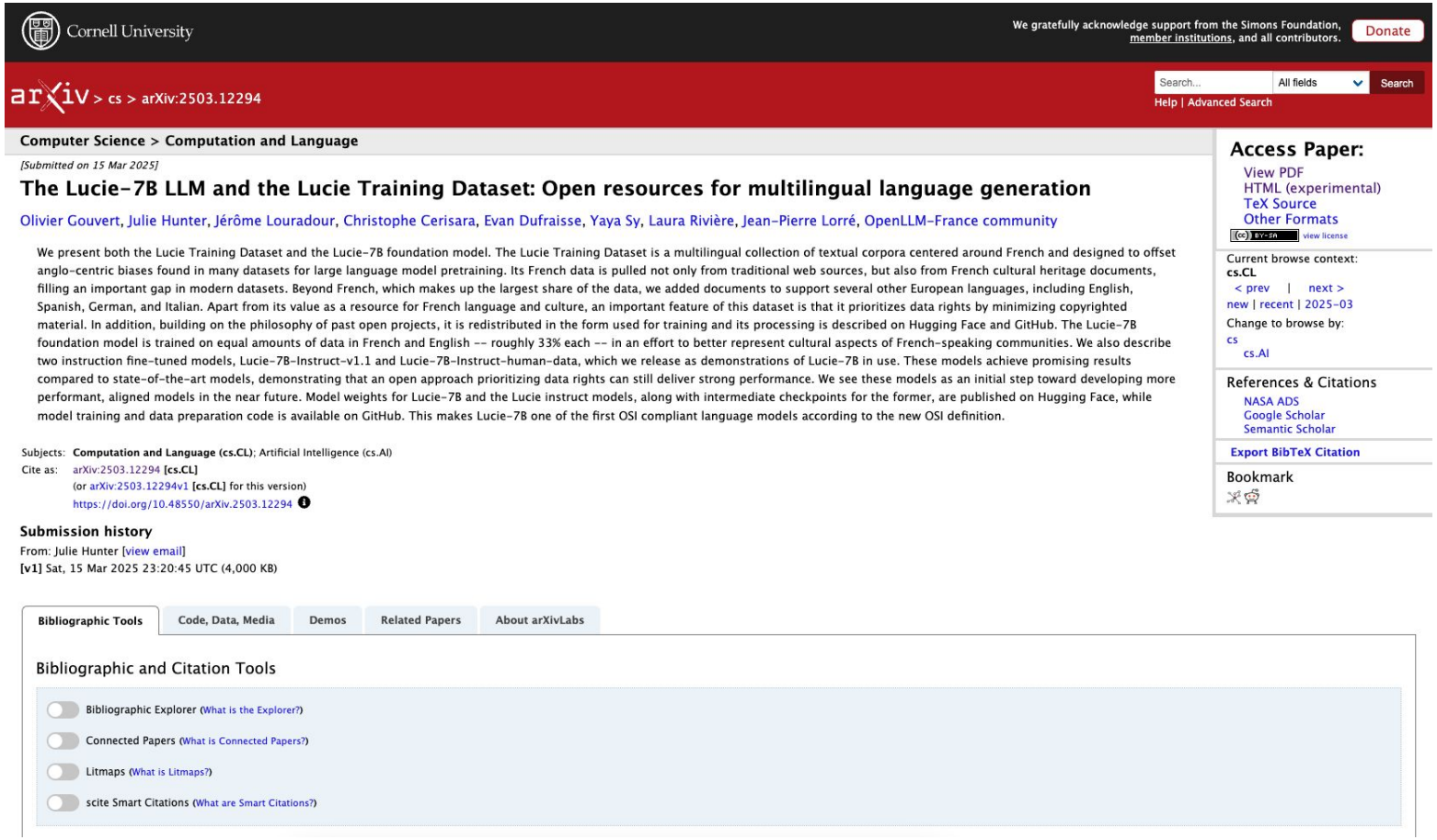
OpenLLM-France / **Lucie-7B**

OpenLLM-France / **Lucie-Training**



Total transparency on "secret sauce

GOSIM



Cornell University

We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate

arXiv > cs > arXiv:2503.12294

Search... All fields Search

Help | Advanced Search

Computer Science > Computation and Language

[Submitted on 15 Mar 2025]

The Lucie-7B LLM and the Lucie Training Dataset: Open resources for multilingual language generation

Olivier Gouvert, Julie Hunter, Jérôme Louradour, Christophe Cerisara, Evan Dufraisse, Yaya Sy, Laura Rivière, Jean-Pierre Lorré, OpenLLM-France community

We present both the Lucie Training Dataset and the Lucie-7B foundation model. The Lucie Training Dataset is a multilingual collection of textual corpora centered around French and designed to offset anglo-centric biases found in many datasets for large language model pretraining. Its French data is pulled not only from traditional web sources, but also from French cultural heritage documents, filling an important gap in modern datasets. Beyond French, which makes up the largest share of the data, we added documents to support several other European languages, including English, Spanish, German, and Italian. Apart from its value as a resource for French language and culture, an important feature of this dataset is that it prioritizes data rights by minimizing copyrighted material. In addition, building on the philosophy of past open projects, it is redistributed in the form used for training and its processing is described on Hugging Face and GitHub. The Lucie-7B foundation model is trained on equal amounts of data in French and English -- roughly 33% each -- in an effort to better represent cultural aspects of French-speaking communities. We also describe two instruction fine-tuned models, Lucie-7B-Instruct-v1.1 and Lucie-7B-Instruct-human-data, which we release as demonstrations of Lucie-7B in use. These models achieve promising results compared to state-of-the-art models, demonstrating that an open approach prioritizing data rights can still deliver strong performance. We see these models as an initial step toward developing more performant, aligned models in the near future. Model weights for Lucie-7B and the Lucie instruct models, along with intermediate checkpoints for the former, are published on Hugging Face, while model training and data preparation code is available on GitHub. This makes Lucie-7B one of the first OSI compliant language models according to the new OSI definition.

Subjects: **Computation and Language (cs.CL)**; Artificial Intelligence (cs.AI)
Cite as: arXiv:2503.12294 [cs.CL]
(or arXiv:2503.12294v1 [cs.CL] for this version)
https://doi.org/10.48550/arXiv.2503.12294

Submission history
From: Julie Hunter [view email]
[v1] Sat, 15 Mar 2025 23:20:45 UTC (4,000 KB)

Bibliographic Tools | Code, Data, Media | Demos | Related Papers | About arXivLabs

Bibliographic and Citation Tools

Bibliographic Explorer (What is the Explorer?)
Connected Papers (What is Connected Papers?)
Litmaps (What is Litmaps?)
scite Smart Citations (What are Smart Citations?)

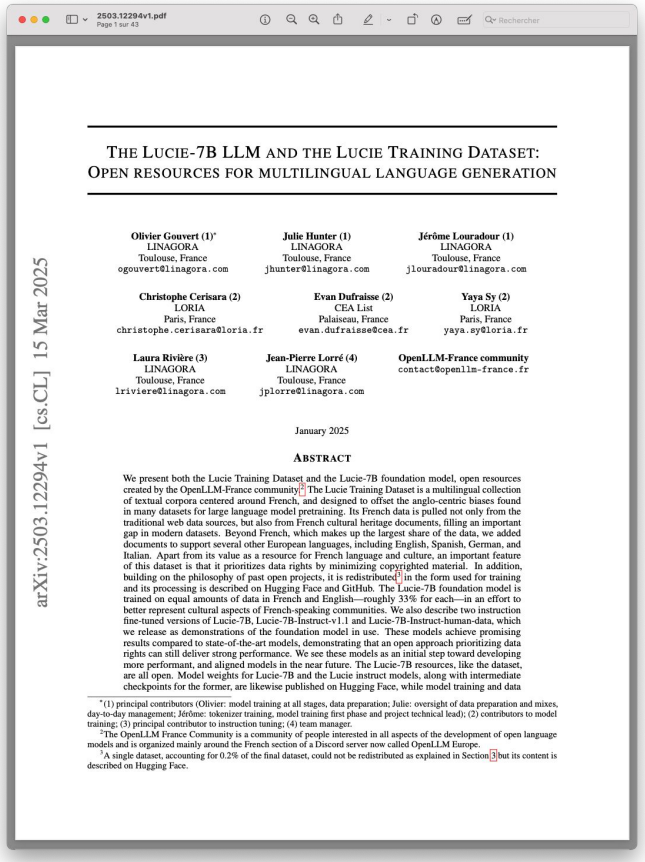
Access Paper:
View PDF
HTML (experimental)
TeX Source
Other Formats
view license

Current browse context: cs.CL
< prev | next >
new | recent | 2025-03
Change to browse by: cs, cs.AI

References & Citations
NASA ADS
Google Scholar
Semantic Scholar

Export BibTeX Citation

Bookmark



2503.12294v1.pdf

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

January 2025

ABSTRACT

We present both the Lucie Training Dataset and the Lucie-7B foundation model, open resources created by the OpenLLM-France community.[2] The Lucie Training Dataset is a multilingual collection of textual corpora centered around French, and designed to offset the anglo-centric biases found in many datasets for large language model pretraining. Its French data is pulled not only from the traditional web data sources, but also from French cultural heritage documents, filling an important gap in modern datasets. Beyond French, which makes up the largest share of the data, we added documents to support several other European languages, including English, Spanish, German, and Italian. Apart from its value as a resource for French language and culture, an important feature of this dataset is that it prioritizes data rights by minimizing copyrighted material. In addition, building on the philosophy of past open projects, it is redistributed[3] in the form used for training and its processing is described on Hugging Face and GitHub. The Lucie-7B foundation model is trained on equal amounts of data in French and English—roughly 33% for each—in an effort to better represent cultural aspects of French-speaking communities. We also describe two instruction fine-tuned versions of Lucie-7B, Lucie-7B-Instruct-v1.1 and Lucie-7B-Instruct-human-data, which we release as demonstrations of the foundation model in use. These models achieve promising results compared to state-of-the-art models, demonstrating that an open approach prioritizing data rights can still deliver strong performance. We see these models as an initial step toward developing more performant, and aligned models in the near future. The Lucie-7B resources, like the dataset, are all open. Model weights for Lucie-7B and the Lucie instruct models, along with intermediate checkpoints for the former, are likewise published on Hugging Face, while model training and data

*(1) principal contributors (Olivier: model training at all stages, data preparation; Julie: oversight of data preparation and mixes, day-to-day management; Jérôme: tokenizer training, model training first phase and project technical lead); (2) contributors to model training; (3) principal contributor to instruction tuning; (4) team manager.
[2]The OpenLLM France Community is a community of people interested in all aspects of the development of open language models and is organized mainly around the French section of a Discord server now called OpenLLM Europe.
[3]A single dataset, accounting for 0.2% of the final dataset, could not be redistributed as explained in Section 3 but its content is described on Hugging Face.

arXiv:2503.12294v1 [cs.CL] 15 Mar 2025

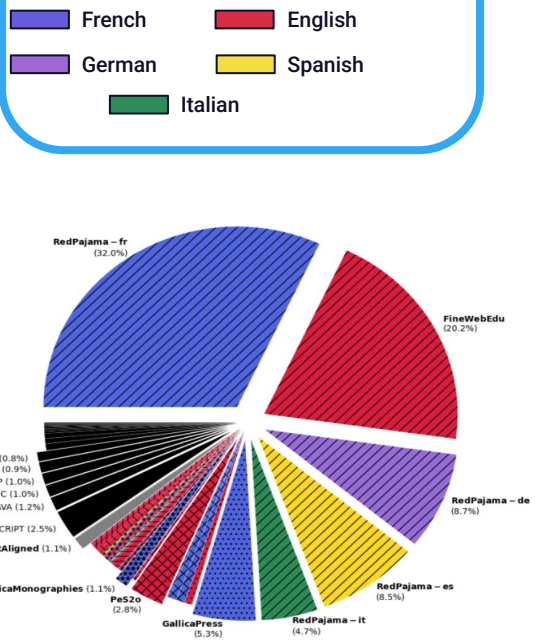
Academic paper: https://arxiv.org/abs/2503.12294

GOSIM AI Paris 2025

Training datasets

Balancing French and English

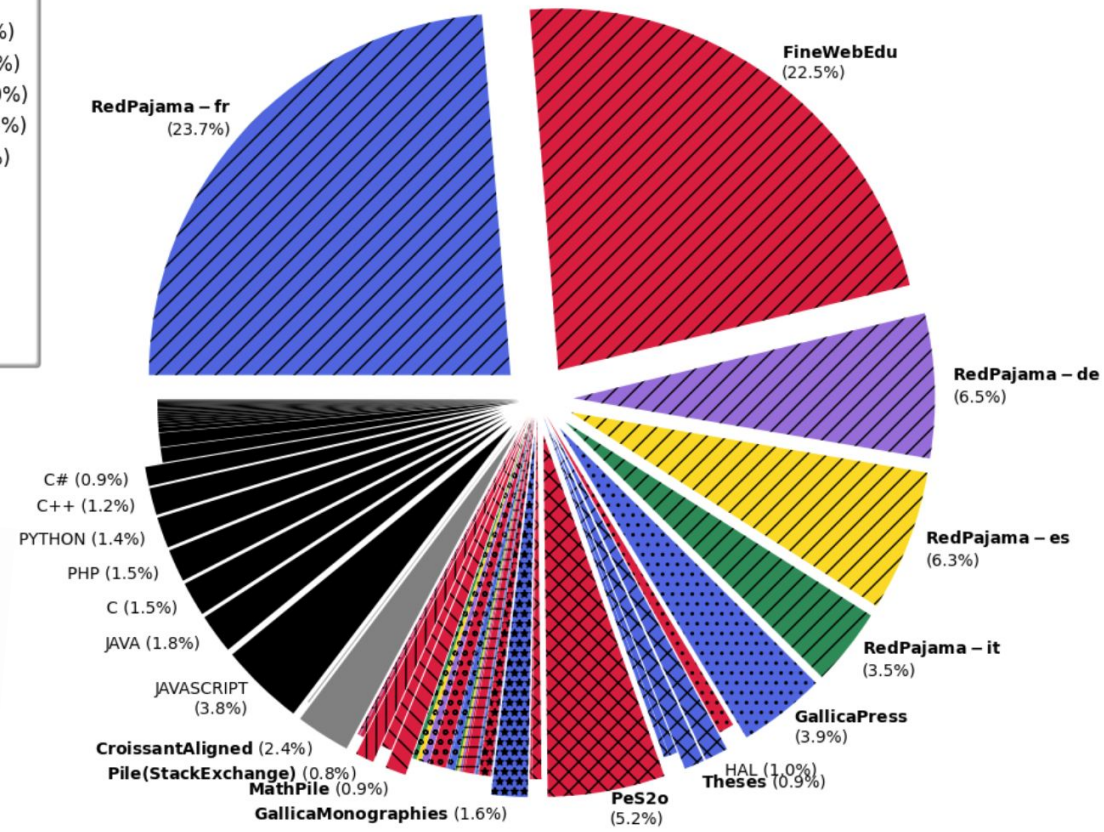
Composition of **original** training dataset (2320.616 B tokens)

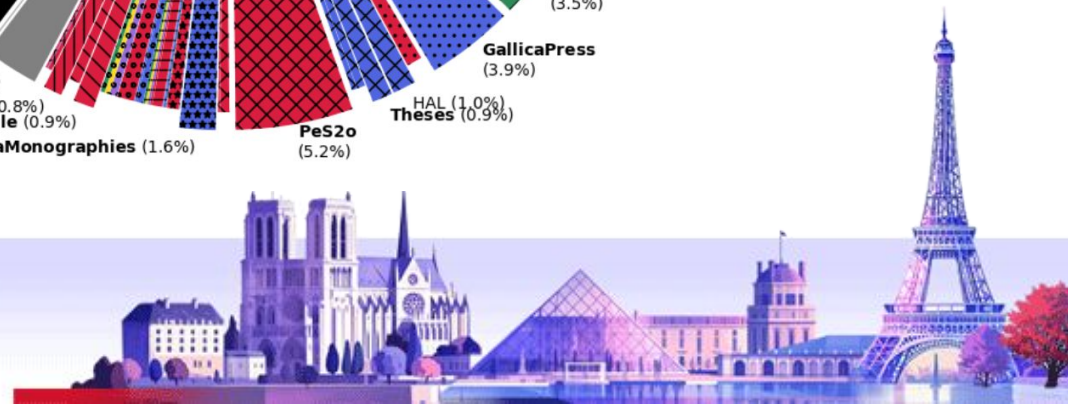


Categories		Languages	
Web	(73.9%)	French	(40.3%)
Newspaper	(5.86%)	English	(26.4%)
Technical	(4.79%)	German	(8.90%)
Book	(1.36%)	Spanish	(8.65%)
Legislative	(1.00%)	Italian	(4.83%)
Wiki	(0.832%)		
Math	(0.628%)		
Forum	(0.536%)		
Dialogue	(0.0779%)		
Multilingual	(1.12%)		
Programming	(9.87%)		

weighting by language and category

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





Tokernizer traininig

GOSIM

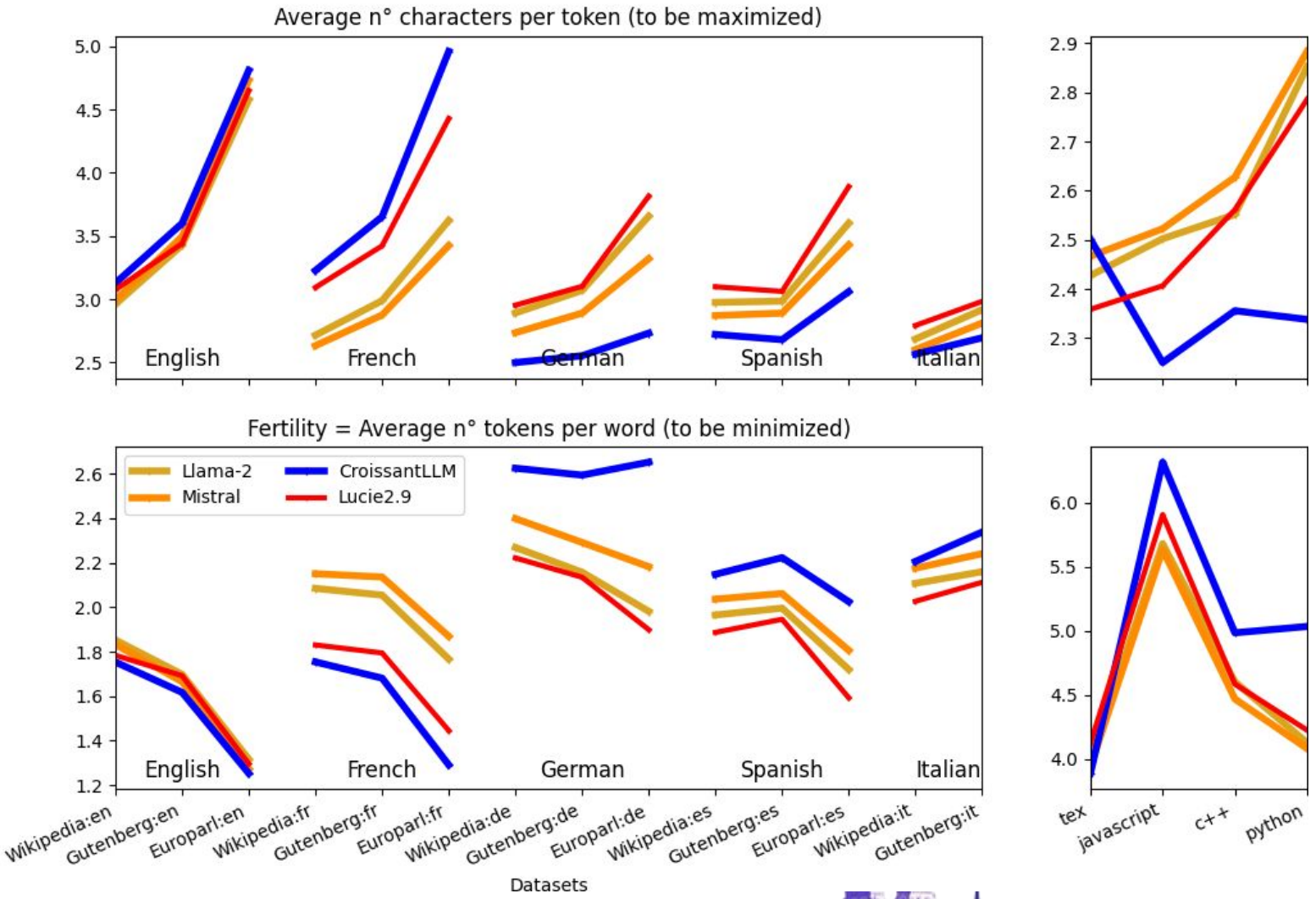
“Le chat mange la …”

tokenizer

```
'<s>', '_Le', '_chat', '_mange', '_la'
```

LLM

```
'_souris'
```



Pre-training on the Jean ZAY supercomputer

GOSIM

INSTITUT DU DÉVELOPPEMENT ET DES RESSOURCES EN INFORMATIQUE SCIENTIFIQUE

GENCI
Le calcul intensif au service de la connaissance

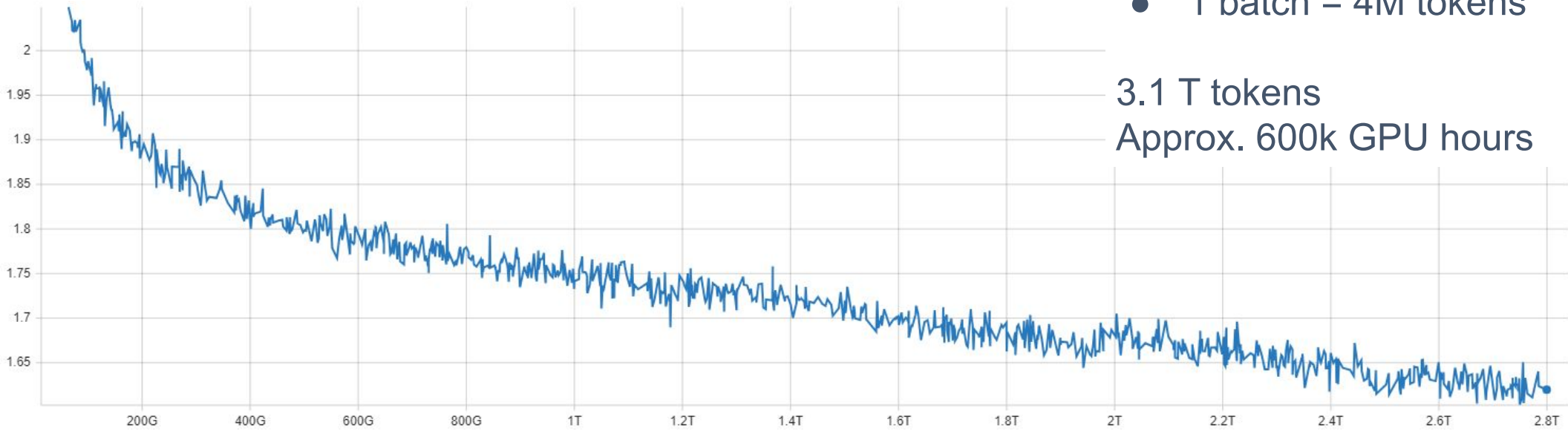
Jean-Zay cluster (CNRS)

512 GPUs H100 with 3D parallelism:

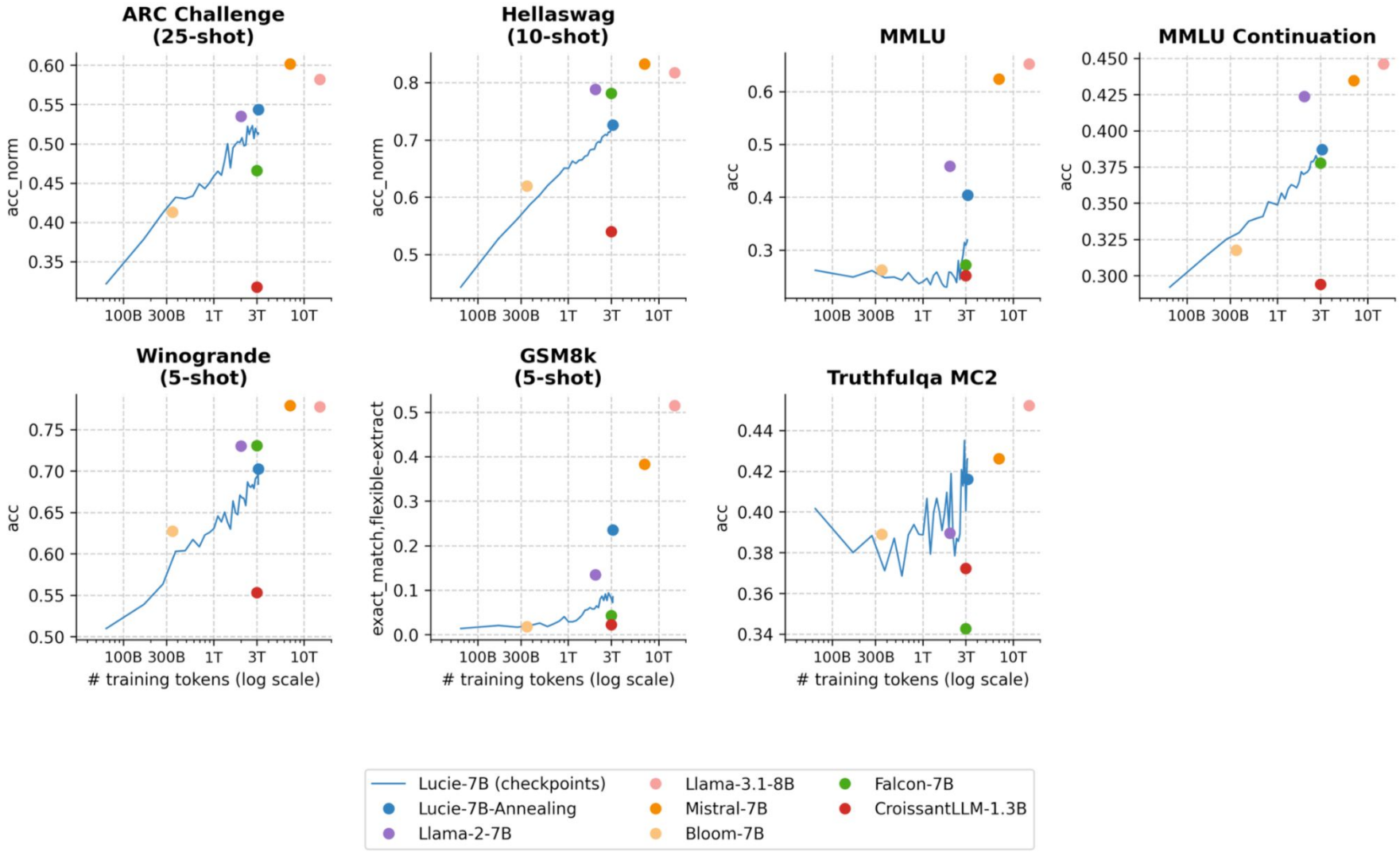
- PP = 4
- TP = 4
- DP = 32
- 1 batch = 4M tokens

3.1 T tokens
Approx. 600k GPU hours

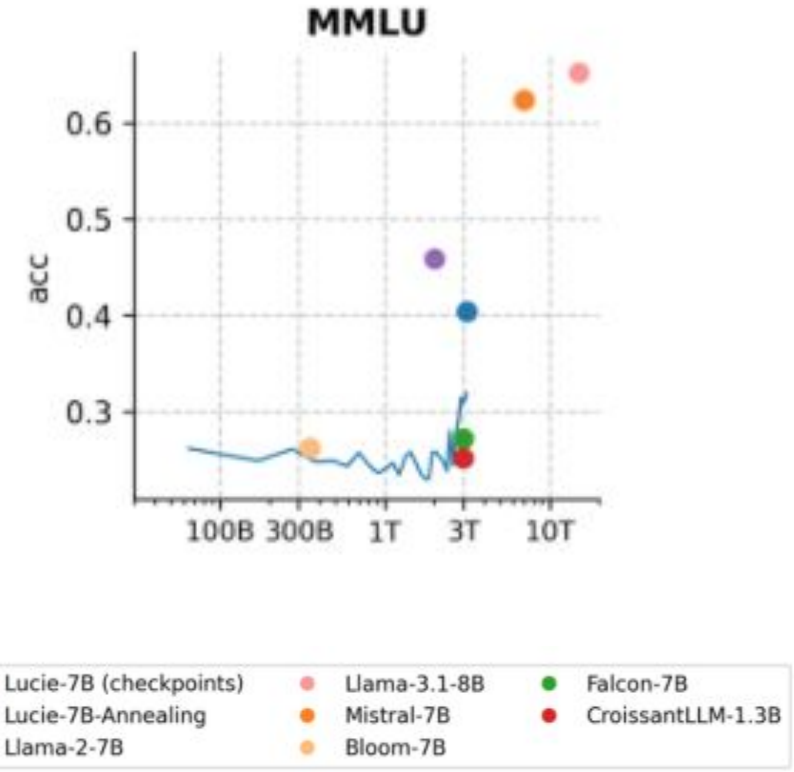
Training loss:



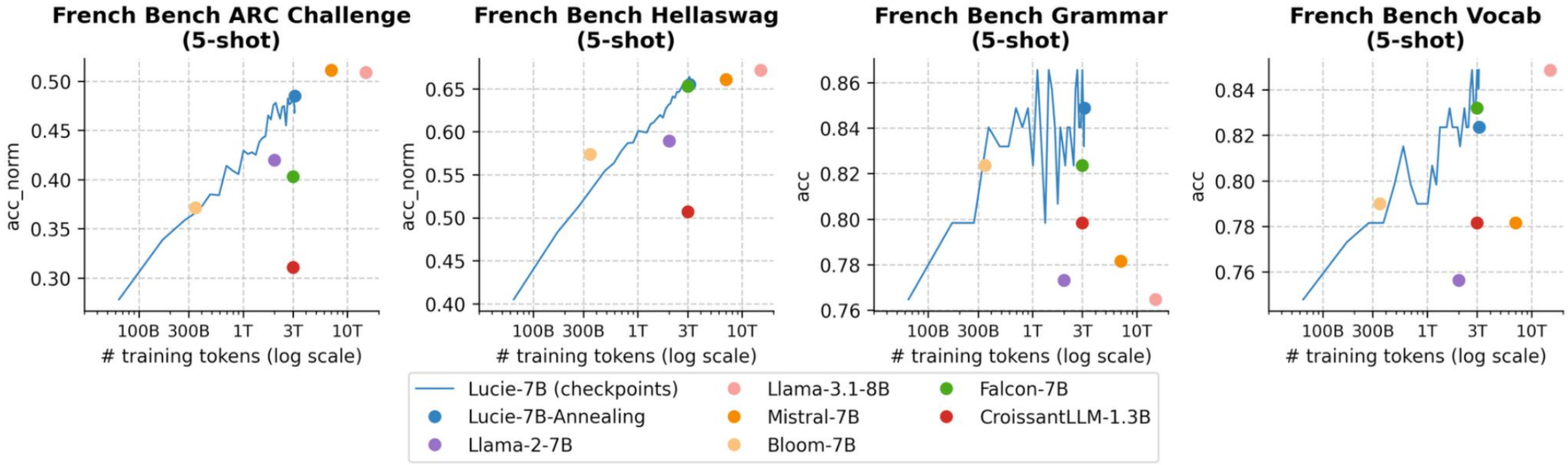
Evaluation / Benchmark



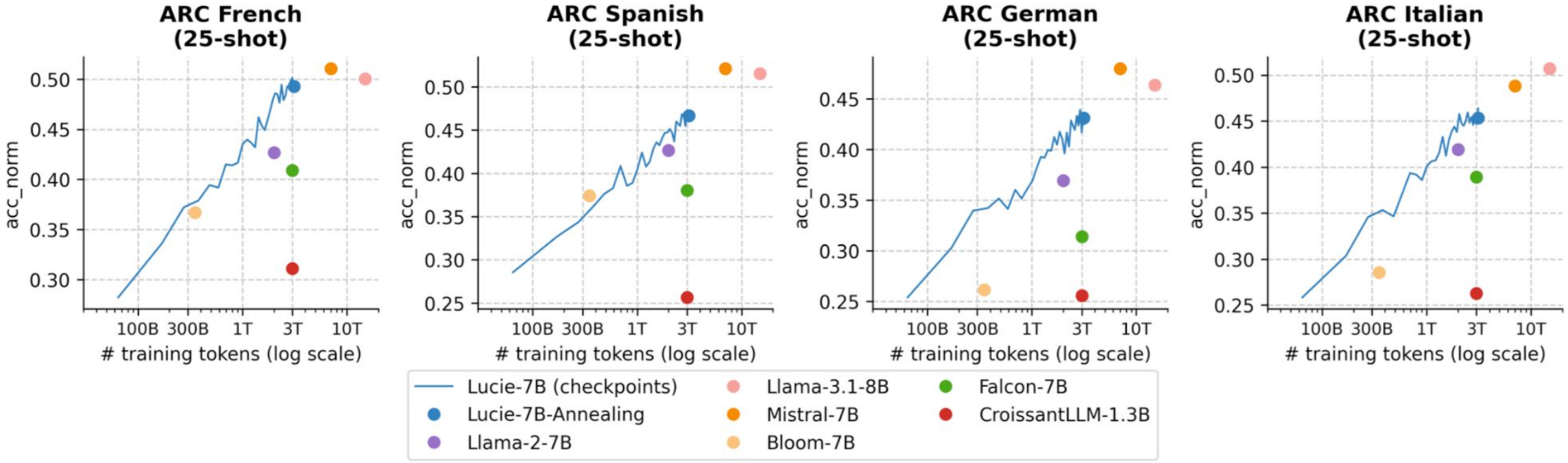
Evaluation / Benchmark



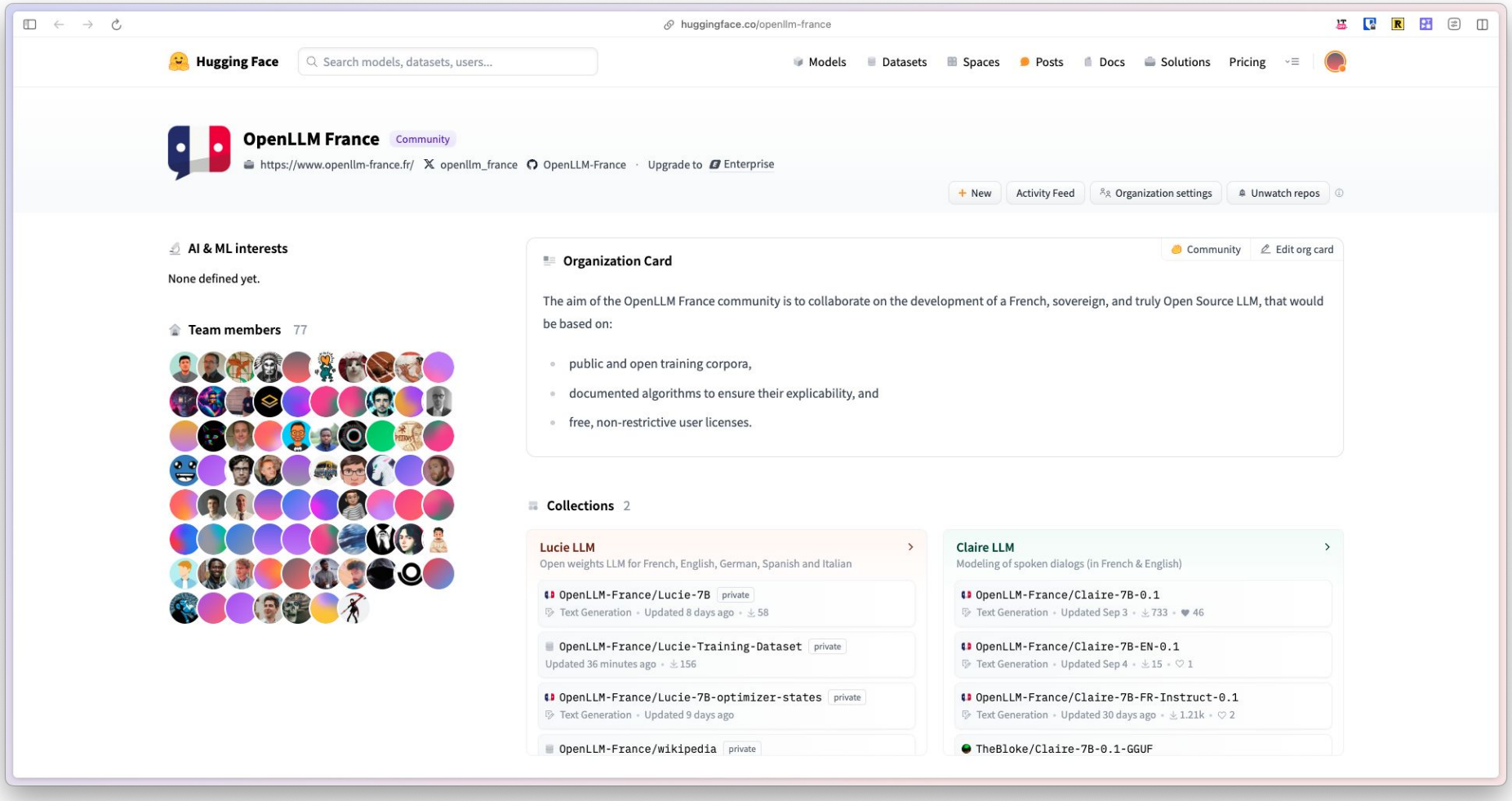
Evaluation / Benchmark



Evaluation / Benchmark

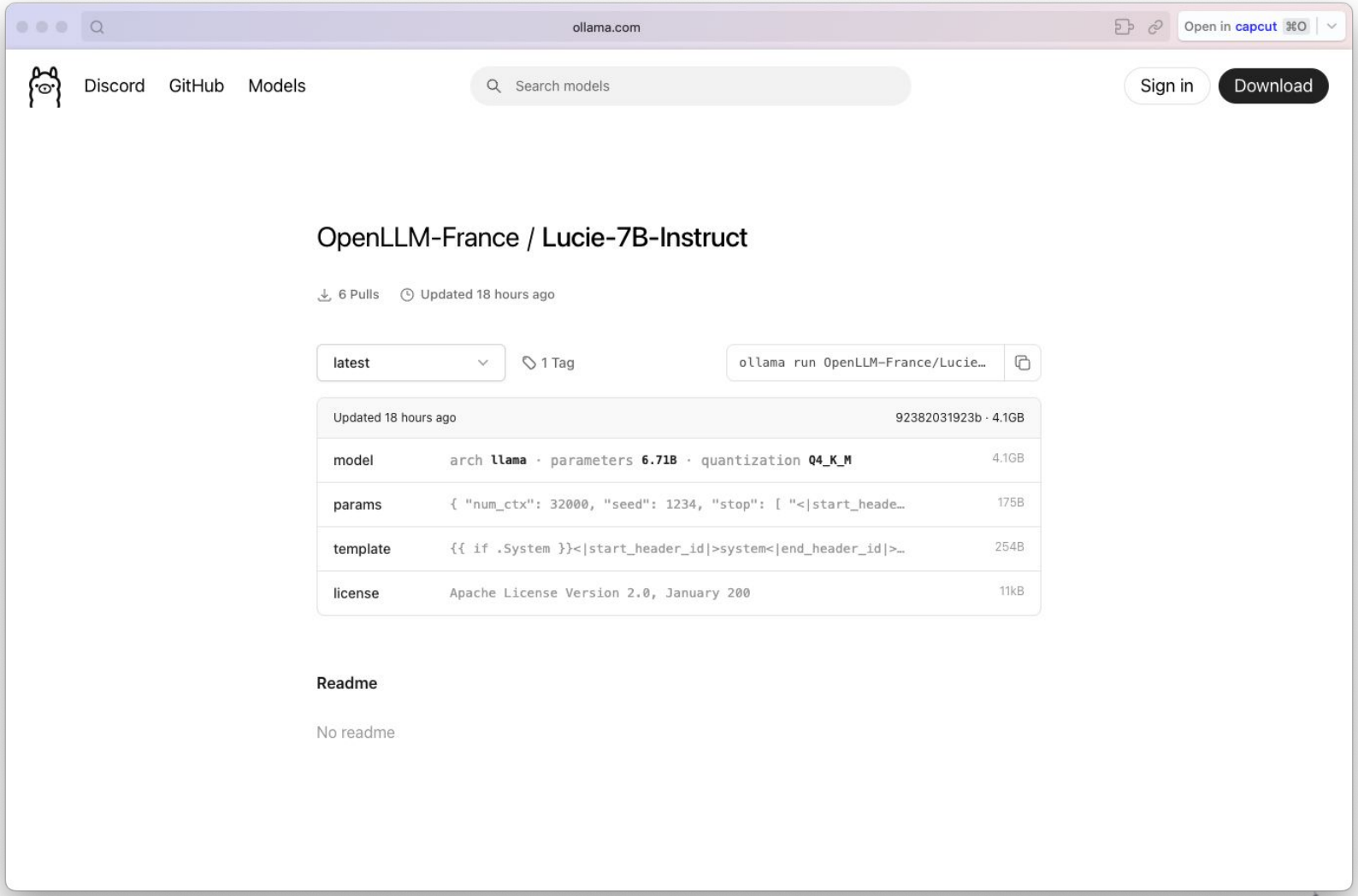


OpenLLM France on Hugging Face



https://huggingface.co/OpenLLM-France

OpenLLM France available with Ollama and LM Studio

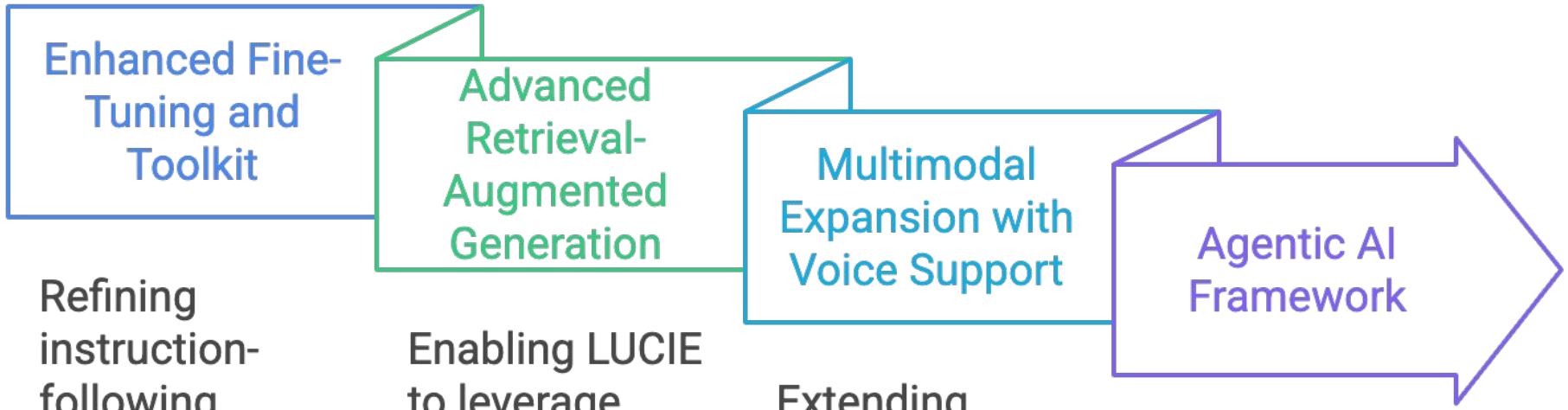


https://ollama.com/OpenLLM-France/Lucie-7B-Instruct

LUCIE Technical Roadmap 2025

GOSIM

LUCIE's 2025 Development Roadmap

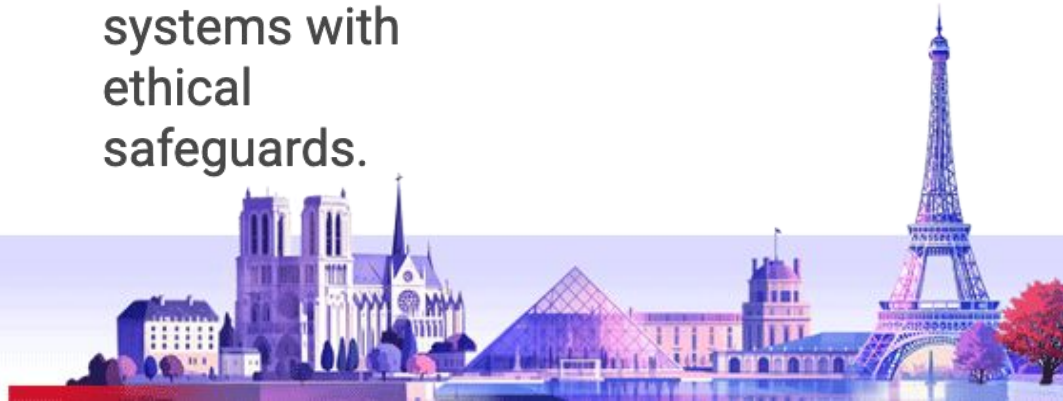


Refining instruction-following capabilities and introducing function calling for better integration.

Enabling LUCIE to leverage external knowledge bases for accurate responses.

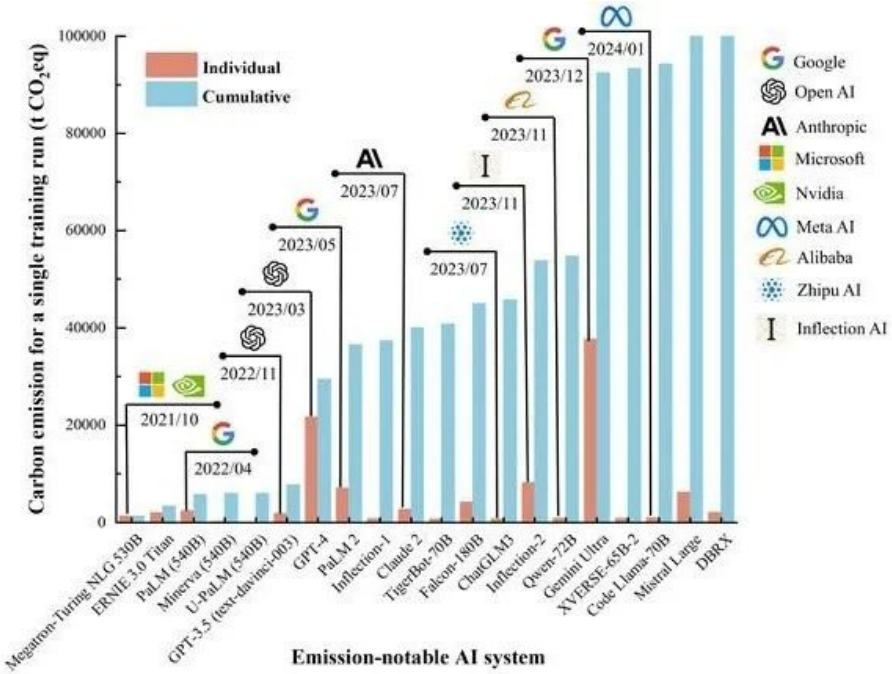
Extending capabilities into multimodal AI with voice processing in French.

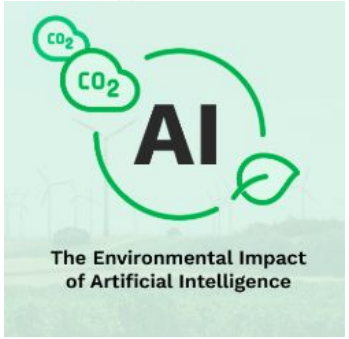
Evolving into a framework for autonomous systems with ethical safeguards.



Pre-training environmental footprint

Training Phase. As the Jean Zay supercomputer is located in France, it runs on low-carbon nuclear electricity. Our primary training took place between September and December 2024, during a period of low water stress, when the French energy mix was not reliant on Germany, where electricity production is not carbon-free. CNRS/IDRIS has implemented a waste heat recovery system that provides heating for all buildings on the Saclay Plateau. The Power Usage Effectiveness (PUE) for the H100 partition is 1.21. According to GENCI, the carbon footprint coefficient for this H100 partition is 25.7 g CO_2-eq per GPU-hour. Given that our training consumed 500,000 GPU hours, this results in a total carbon footprint of 12.85 metric tons of CO_2 equivalent—compared to 31.22 metric tons CO_2-eq for LLaMA 2 7B or 390 metric tons CO_2-eq for LLaMA 3 8B.





Use case - Commercial offer

Transcription / Traduction / Compte rendu	Agents conversationnels	Recherche et synthèse
Génération de documents	Génération de code, transcodage	Génération images / vidéos
Structuration et navigation sémantique	Automatisation messageries / mails	Automatisation traitements (no code / low code)

Secteurs / Cas d'usage

Administration / Collectivités	Banque / Assurance	Industrie	RH & Education
Automatisation des réponses citoyennes, chatbot, transcription et analyse des débats publics.	Document processing, Assistance conseiller centre appel, Analyse compliance Génération email	Génération de code, Documentations techniques, Analyse d'images, recherche	Génération job description, Résumé candidature, Aide pédagogique, auto-correction, Plateformes d'apprentissage
...	...	...	...

LUCIE Open Agents

Orchestrateurs Open Source

LinTO

Expertise Whisper / Kaldi / Parakeet, reconnaissance locuteurs, transcription temps réel

Lucie LLM

100% Open Source, résumé, synthèse, coding, etc.), Fine-tuning, RAG, Renforcement Learning

Modèles Vision & OCR

Analyse sémantique, détection de structures, tableaux, OCR

LLM Open Weight ou commerciaux

MistralAI, Meta llama, Claude, ChatGPT et autres modèles non Open Source

Want to join our community

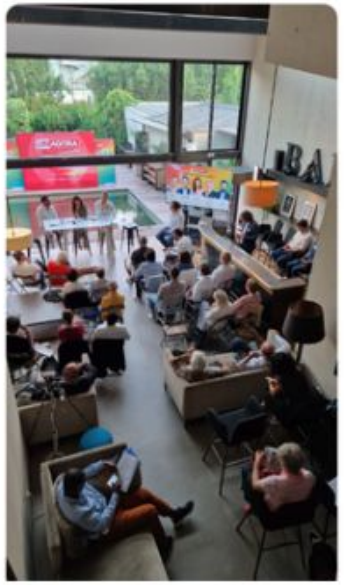
GOSIM

DISCORD

LINAGORA.AI

C'est le bon moment de rejoindre nos équipes extraordinaires...
et pas seulement pour la piscine !

Postes recherchés :
LLMOps (++), Data Scientist, Consultants IA, Chef de projet IA





GOSIM

GOSIM AI Paris 2025

THANK YOU