Construire des LLM de taille moyenne véritablement ouverts, centrés sur le français

Pourquoi et comment

Julie Hunter

Open Science Days @ UGA 2025

À propos de LINAGORA

Leader dans le logiciel libre depuis 25 ans

Avec ses logiciels et services, LINAGORA permet aux grandes organisations publiques et privées de developer leur independence technologique.

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

Aujourd'hui

1. Pourquoi faire des LLMs open source ?
2. Introduction à l'initiative OpenLLM France
3. Aperçu des étapes d'entrainement d'un LLM
4. Les modèles de OpenLLM
5. Au-delà de l'entraînement

Pourquoi faire des LLMs open source ?

Des LLMs partout !

Assistance chatbot : planification des voyages, réponses aux urgences, emergency response, soutien scolaire, interrogation des documents, ...

Production de documents : résumés, exercices, recommandations, ...

Plusieurs modèles open weights !

A-t-on besoin de plus ?

Que faire si un LLM générique n'est pas adapté à votre cas d'utilisation ou ne maîtrise pas votre langue cible ?

... si vous avez besoin de modèles plus petits en raison d'un manque de matériel (GPUs) ou d'accès à des API payantes ?

... si vous souhaitez simplement comprendre pourquoi les LLMs fonctionnent (ou pas) ?

... si vous avez besoin de savoir quelles données d'entraînement ont été utilisées ?

Quelques obstacles

De nombreux LLM sont dans les mains de grandes entreprises qui communiquent peu d'informations sur leurs données ou leurs méthodes.

- Coût informatique et accès aux données

Les modèles open-weights permettant l'affinage.

Mais :

- l'affinage ne peut pas tout faire
- les modèles génériques très performants restent plutôt gros
- on hérite les biais/problèmes du modèle de base
- la recherche est limitée aux étapes finales de l'entraînement

OpenLLM France

OpenLLM France

Projet financé par la BPI (09.2024 – 08.2026) issu d'un communauté

Vise à construire des technologies d'IA générative véritablement open source, éthiques, compactes et souveraines, centrés sur le français.

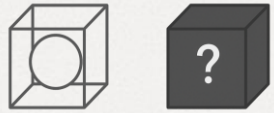
LINAGORA

opsci TALKR.ai Class'Code cea

cnrs IDRIS Loria Laboratoire lorrain de recherche en informatique et ses applications LX UNIVERSITÉ PARIS 1 PANTHÉON SORBONNE

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

IA véritablement open source



Trois conditions essentielles :

1. **Licence** d'utilisation sans restriction pour le modèle
2. **Transparence** totale sur les méthodes d'entraînement
3. **Disponibilité des données** d'entraînement sous une licence ouverte

Peu d'initiatives ouvertes

Quelques exemples notables :

Ressources pour l'anglais

Multilingues avec du français

Sur-representation de l'anglais

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%

La situation s'est améliorée un peu, mais l'anglais semble rester dominant (sachant que nous n'avons peu d'infos sur la plupart des LLMs)

Une question de langue ET de culture

- Histoire
- Politique
- Art
- Cuisine
- Normes étiques…

Comment entraîner un LLM ?

Les grands modèles de langage

Modèles génératifs

J'aime manger des céréales au ___

- **I like to eat cereal for breakfast.**
- Les céréales sont généralement mangées à quel repas ? **Le petit déjeuner**
- Parfois je mange des viennoiseries ou des crêpes au petit déjeuner. De temps en temps, je mange de la saucisse, mais je n'aime vraiment pas prendre de viande le matin. D'habitude, je mange juste des céréales → **Julie aime manger des céréales pour le petit déjeuner.**

Etapes d'entraînement

Entraînement générique

1. Tokenisation
2. Pré-entraînement
3. Ajout des tâches
4. Ajout de préférences

Tokenisation

Trouver des unites minimales pour constuire des sequences

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

Caractères : `J | ‘ | a | d | o | r | e | l | a | c | h | o | c | o | l | a | t | e | r | i | e`

Sous-mots : `J ' | ad | ore | la | ch | ocol | ater | ie`

Une bonne tokenisation permet d'économiser de l'argent et améliore les performances !

Entraînée sur les mêmes données que le LLM à entraîner

Pré-entraînement

Prédiction du prochain token dépend des séquences vues lors de l'entraînement

Corpus d'entraînement : J'aime manger des céréales au dîner. J'aime manger des céréales au dîner. J'aime manger des céréales au dîner. J'aime manger des céréales au dîner.

Prédiction: J'aime manger des céréales au **dîner**

→ Besoin de diversité et de bonne couverture

→ Et beaucoup de données (des "trillions" de tokens)

Apprendre à faire des tâches

Quelle est la capitale de la France?

Nous vous proposons d'apprendre ou d'améliorer votre français avec un professeur particulier ou en groupe dans une ambiance conviviale. Le professeur particulier vous accompagnera sur le chemin de la réussite pour vous aider à atteindre vos objectifs.

Paris.

Pretrained model

Instruction fine-tuned model

Instruction fine-tuning

Apprendre au LLM à effectuer des tâches

- Peux-tu résumer le texte suivant {texte} ?
- Peux-tu écrire un poème sur les cacahuètes ?
- Peux-tu expliquer la relativité en deux paragraphes ?

Training templates:

```
### Instruction:
{instruction} \n###
Response:
```

Données

- Moins que le pré-entraînement mais tout de même des (centaines de) millions d'exemples
- Plus concentré sur la qualité que la diversité, mais il faut quand même pouvoir gérer un large éventail de requêtes ouvertes des utilisateurs

Sensible aux données de pré-entraînement

Préférences

Votre LLM a appris à imiter beaucoup de choses

Mais comme un enfant qui aurait appris à imiter ses parents en disant des gros mots, il est également important d'enseigner au modèle quel est le comportement souhaitable…

Classer des réponses de LLMs comme meilleures ou moins bonnes

Apprend au modèle d'imiter ces préférences → besoin d'annotation humaine ou par un autre modèle

Egalement sensible aux étapes précédentes

Les modèles d’OpenLLM

Lucie 7B (pre-trained model)

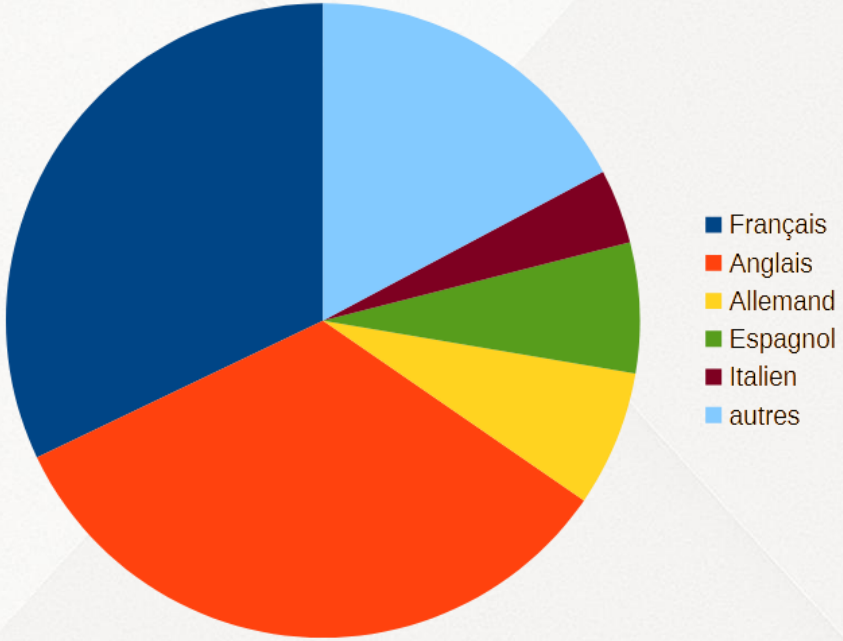
« Lucie » vient de « lux »
Lucie apporte de la lumière à la boîte noire de l'IA.

LLM Open Source
données, code, poids (finaux et intermédiaires)

Focus sur le français

Compliant RGPD et AI-Act

Représentation du français



3 000 milliards de tokens (2 300 uniques)

Entraînement d'un **nouvel tokeniser** pour équilibrer la représentation de toutes les langues dans les données d'entraînement

Les nouveaux modèles

Toujours 100% open, toujours centrés sur le français (30%), toujours compliants RGPD et AI-Act

- 1B, 8B, 23B
- 5 000 de tokens
- + po, nl, ar
- Entrainement en phases pour améliorer le raisonnement

Défi 1 : les données web (a)

Très difficile à éviter – la majorité des données de pré-entraînement

Qualité : filtrage par heuristiques, classifieurs (via des LLMs), déduplication (Hugging Face, AI2, …)

Propriété intellectuelle : filtrage par

- robots.txt : imparfait, à appliquer rétroactivement (cf. Common Crawl) ? Quel impact pour l'open source ?
- licence (Common Pile), domaines connus (Common Corpus) : restriction sévère sur la quantité de tokens

Défi 1 : les données web (b)

Toxicité et biais :

- Blacklists, no-go words
- Usage de classifieurs entrainés sur des annotations des LLMs (Hugging Face, AI2, Apertus, …)
- Argument que les LLMs doivent voir des données nuisibles pour pouvoir apprendre qu'elles sont mauvaises (phase de préférences)

Défi 2 : les données françaises (a)

Concentration sur les données anglaises (Common Pile, AI2, Hugging Face jusqu'à très récemment ...)

Données françaises :

- Difficulté des licences ouvertes (cf l'énorme travail du Common Pile)
- Données du domaine publique (gros effort de Common Corpus) : vieux documents (des biais !), documents OCRisés (!!!), données du gouvernement (bonnes, mais peu)

Défi 2 : les données françaises (b)

Quantité donc très limitée

En choisissant les meilleures données web en français à notre disposition (toujours moins bonnes que l'anglais), nous arrivons à :

- ~500 milliards de tokens web
- ~225 milliards de tokens non-web (presque entièrement OCR).

Difficile d'imaginer un entrainement « long horizon » avec une important proportion de données françaises

Instructions et préférences

De nouveaux défis !

Encore moins d'ouverture, encore moins de français

Pas plus facile à obtenir : besoin de moins de quantité mais plus de qualité et de format spécifique et l'annotation humaine est couteuse.

Génération de données avec des LLMs

- Biais introduits
- Questions d'ouverture et souveraineté : pas facile à respecter nos valeurs

La construction de données souveraines pour le post-training va prendre du temps !

Au-delà de l’entraînement

Publication des données

Hugging Face | Search models, datasets, users... | Models | Datasets | Spaces | Community | Docs | Enterprise | Pricing

Datasets: OpenLLM-France / Lucie-Training-Dataset | like 30 | Following OpenLLM France 297

Tasks: Text Generation | Modalities: Text | Formats: parquet | Sub-tasks: language-modeling | Languages: English | French | German +3 | Size: 10B - 100B

ArXiv: arxiv:2308.12477 | arxiv:2311.16840 | arxiv:2402.00786 +11 | Tags: text-generation | conditional-text-generation | Libraries: Datasets | Dask | Croissant +1 | License: cc-by-nc-sa-4.0

Lucie : données NC mais modèle Apache 2.0

Nouveaux modèles : on évite les données NC pour éviter cette restriction

Subset (108): Gutenberg-fr · 3.45k rows | Split (1): train · 3.45k rows

Search this dataset

source string · classes	id string · lengths	language string · classes	date string · lengths	author string · lengths	url string · classes	title string · lengths
1 value	3 5	1 value	0 33	68 140	1 value	3 42
Gutenberg	10053	fr	November 1, 2003 [eBook #10053]	{"author": "Féval, Paul",…		La vampire
Gutenberg	10061	fr	November 1, 2003 [eBook #10061]	{"author": "Verhaeren, Emile"…		Les Heures Claire
Gutenberg	10160	fr	November 1, 2003 [eBook #10160]	{"author": "France, Anatole",…		Pierre Nozière
Gutenberg	10289	fr	November 1, 2003 [eBook #10289]	{"author": "Hervilly, Ernest…		Le Chat du Neptun
Gutenberg	10346	fr	December 1, 2003 [eBook #10346]	{"author": "Buysse, Cyriel",…		C'était ainsi...
Gutenberg	10384	fr	December 1, 2003	{"author":		Le Pays de l'or

Trouver la place de l'open source

Etape 1 : savoir bien communiquer sur votre projet !

Si votre focus est sur la création d'un modèle fondation 100% à forte coloration française, assurez vous de prévenir le monde que vous ne faites pas un ChatGPT français ;-)

Prochaines étapes : contribuer aux projets ouverts – les modèles ouverts ont du mal à concurrencer les modèles open-weights (a-t-on besoin de les concurrencer ?)

Conclusions

- L’IA open source peut faciliter la création de modèles plus petits et focalisés sur des cas d’usage spécifiques
- Elle permet aussi la recherche sur les LLMs
- OpenLLM se focalise sur le développement de modèles francophones 100% open et transparents, comme Lucie 7B
- Contribution aussi aux connaissances en Europe ; focus ici sur les données, mais l’entraînement est aussi important – besoin de rapports technique et bases de code

Nos publications

- https://arxiv.org/abs/2503.12294
- https://github.com/orgs/OpenLLM-France/repositories
- https://huggingface.co/OpenLLM-France
  (data, final checkpoints, intermediate checkpoints for base model)


Q&A

Merci !