Vers des LLMs ouverts et transparents pour l'Europe : conception, adaptation et explicabilité

Céline Hudelot,
Professeur en Informatique, CentraleSupélec
Laboratoire MICS

Avant propos

Une rapide bio

Profession
Professeur en Informatique– MICS - CentraleSupélec, Université Paris Saclay
Directrice du laboratoire MICS

Formation
HDR en Informatique, Université Paris-Sud Paris-Sud
Doctorat en Informatique, INRIA- *Sophia Antipolis – France*

Activités de recherche

Interprétation sémantique de données non-structurées

Paradigmes d'apprentissage (frugaux, continus, hybrides, actifs)

Modèles de fondation et grands modèles de langue

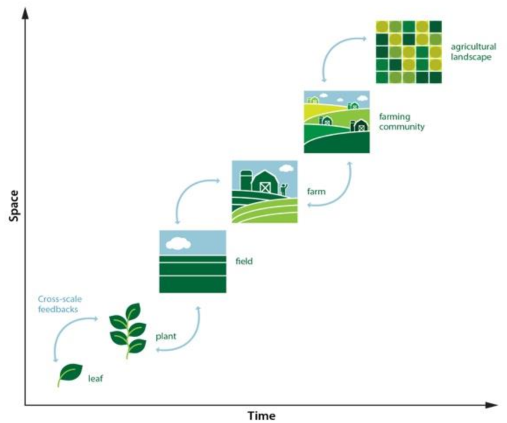
Explicabilité des systèmes d'IA

Laboratoire MICS

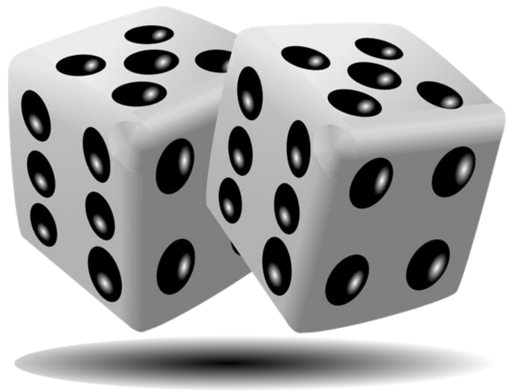
Mathématiques et Informatique pour la Complexité des Systèmes et des Données

Recherche méthodologique et fondamentale en Mathématiques et en Informatique

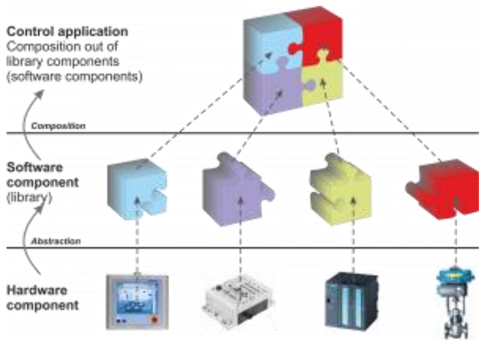
Orientée vers l'étude et la modélisation des systèmes complexes au sens large



Interaction multi-échelle
Ex : vivant



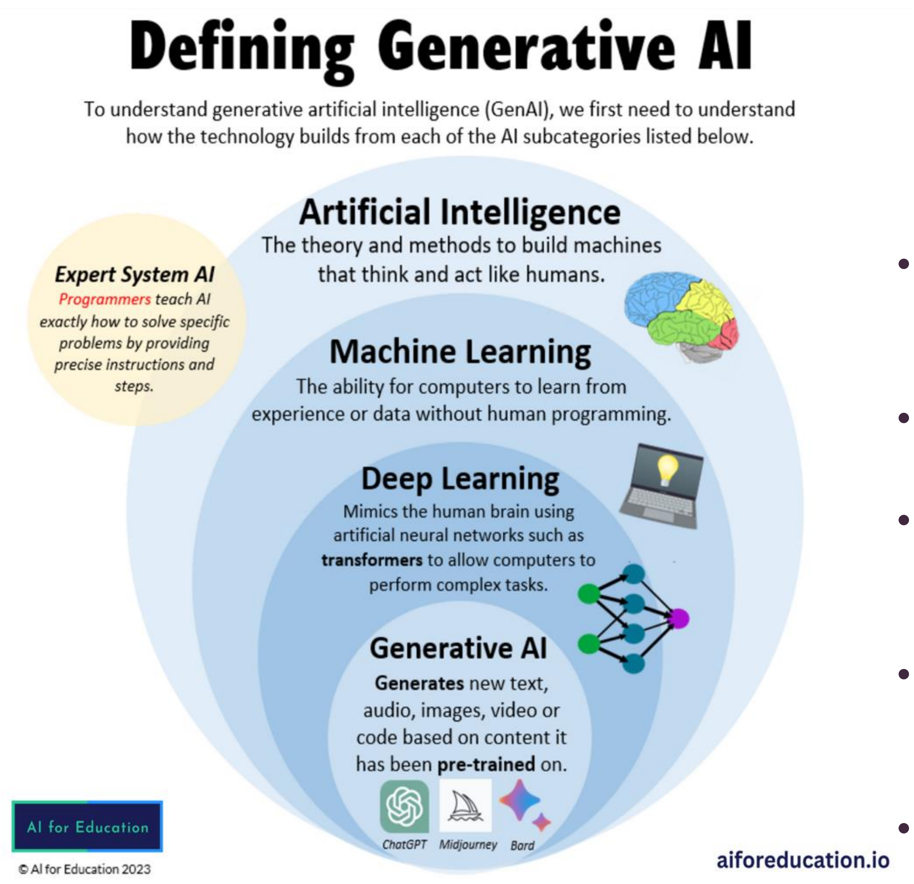
Systèmes stochastiques
Incertitude

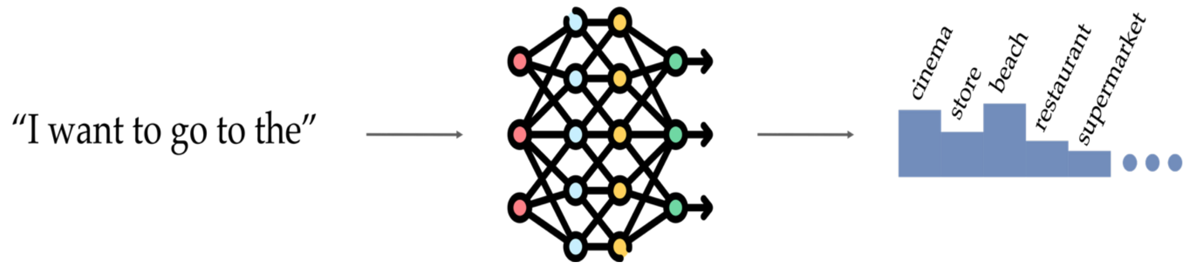


Composants hétérogènes en interaction

Données, non-structurées, hétérogènes, de grande taille

Intelligence Artificielle Générative





- Pré-entraînés à **grande échelle,** avec deux principaux objectifs: **mask language modeling (MLM)** et **causal language modeling (CLM).**
- La phase de **pré-entrainement** permet au modèle d'acquérir des **connaissances** et des **capacités** étendues.
- Lors de l'**inférence**, à partir d'une invite d'entrée (**prompt**), le LLM génère une distribution de probabilité sur un vocabulaire pour les mots ou tokens suivants possibles.
- Un même modèle peut être utilisé **avec peu ou pas d'apprentissage** pour un ensemble de tâches (e.g. traduction, classification, résumé, réponses à des questions,…).
- Vers des **modèles de fondation**.

Travaux sur les LLMs au MICS

Pré-entraînement

The EuroLLM Suite

[Faysse et al, TMLR 25]
[Alves et al, COLM 24]
[Colombo et al, NeurIPS 24]
[Boizard et al, COLM 25]

Adaptation

Safe Retrieval [Gisserot et al, TMLR 24]

Colpali : Multimodal RAG [Faysse et al, ICLR 25]

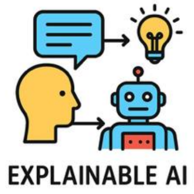
Model Distillation [Boizard et al, TMLR 25]

Évaluation

Instruction Fine-tuned Model Evaluation
[Faysse et al, EMNLP 23]

New metric, benchmarking
[Colombo et al, AAAI 22]
[Colombo et al, NeurIPS 22]

Explicabilité

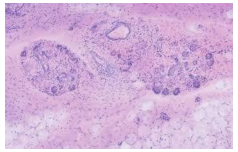


Concept-based approaches
[Claye et al, preprint 25]
[Aswal et al, preprint 25]

Et sur les modèles de fondation, e.g. santé

Modèles de fondation

Novae : graph-based FM for spatial transcriptomics data
[Blampey et al, preprint 24]



Sam-Path : Digital pathology segmentation
[Zhang et al, MICCAI Worshop 23]

rayDino : X-ray analysis
[Moutakanni et al, preprint 24]

Robustesse, Adaptation

Full Conformal Adaptation of Medical VLMs [Silva-Rodriguez et al, MICCAI 25]

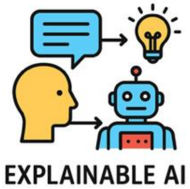
Out-of-distribution generalization [Scalbert et al, MICCAI 22]

Évaluation, Benchmarks

Fair and comprehensive comparaison of FMs in computational pathology
[Marza et al, preprint 25]

https://mics-lab.github.io/thunder/

Explicabilité



Concept-based approaches
[Claye et al, GenBio ICML 25]

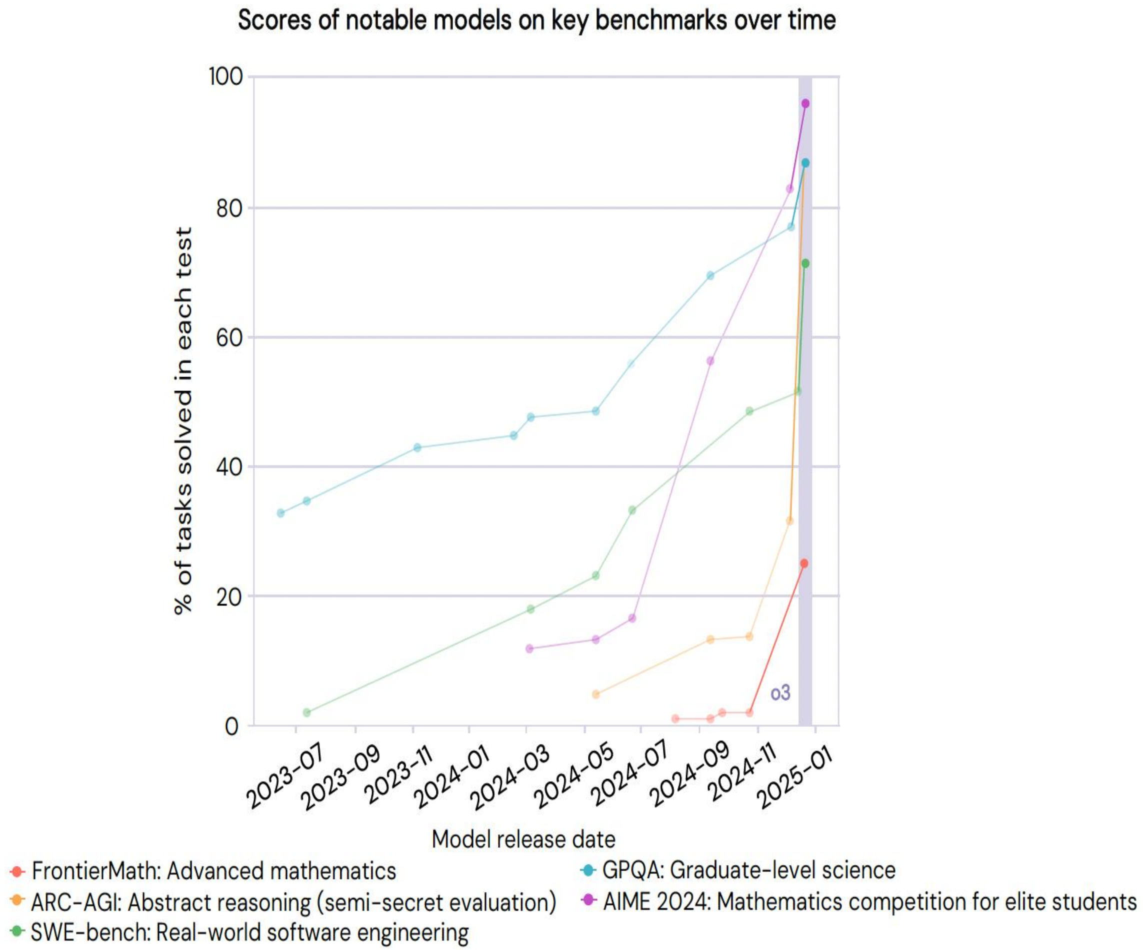
Explanation generation
[Charachon et al, FGCS 22]

Counterfactual analysis for digital histopathology
[Benkirane et al, MIDL 24]

Contexte : des modèles très performants

Des progrès majeurs à un rythme très rapide !

International AI Report [Bengio et al, 25]
« Between the end of the writing period for this report (5 December 2024) and the publication of this report in January 2025, an important development took place »



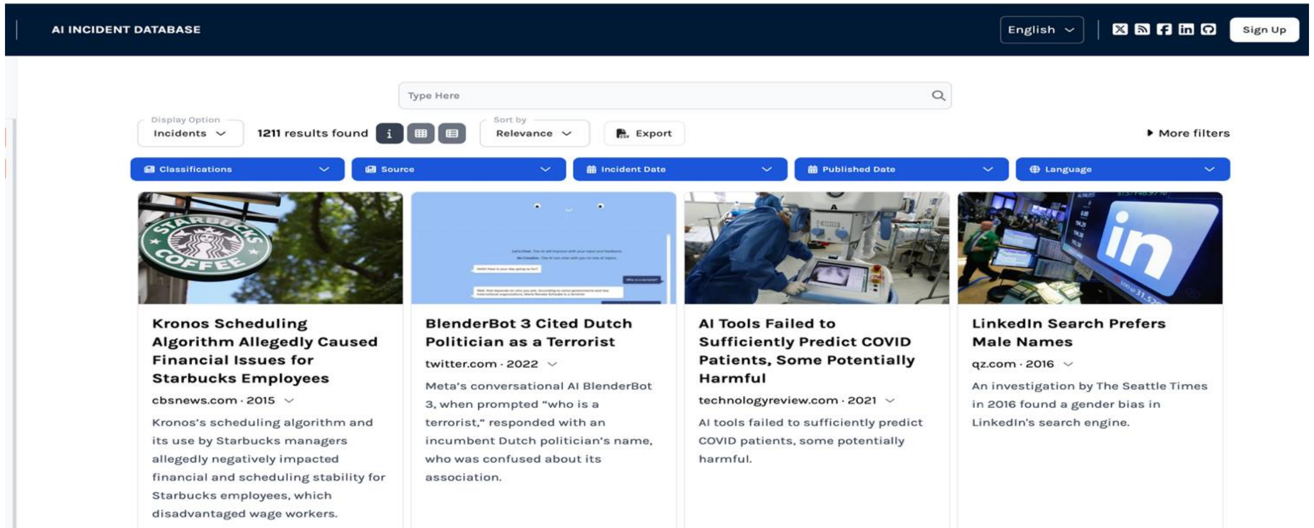
Contexte : des modèles loin d’être dignes de confiance

CNNs learn to predict pneumonia by detecting hospital which took the image

- Study on detecting pneumonia using 158,323 chest radiographs
- CNNs robustly identified hospital system and department within a hospital
- CNN has learned to detect a metal token that radiology technicians place on the patient in the corner of the image

Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. Zech JR1, Badgeley MA2, Liu M2, Costa AB3, Titano JJ4, Oermann EK3. https://www.ncbi.nlm.nih.gov/pubmed/30399157

Tesla on autopilot slam into truck



incidentdatabase.ai

NEWS

Exploring the Dangers of AI in Mental Health Care

DATE JUNE 11, 2025

TOPICS HEALTHCARE GENERATIVE AI

A new Stanford study reveals that AI therapy chatbots may not only lack effectiveness compared to human therapists but could also contribute to harmful stigma and dangerous responses.

Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.

Quelques challenges

- **Protection** des données : **confidentialité** et **intégrité**
- Contraintes liées aux **ressources** disponibles, y compris les données
- **Multimodalité** : plus que du texte et des images
- **Confiance** : **sécurité**, **robustesse**, **explicabilité**
- **Confiance** : **vérification**, **validation**
- **Certification, conformité légale**
- **Souveraineté**

1. Il est possible de concevoir des modèles européens ouverts et transparents.

Une suite de modèles européens

Tower
2B, 7B, 13B, 70B @MN5
AI & Data Intensive
200K+ downloads

CroissantLLM
1.3B @Jean Zay
100K+
downloads

EuroLLM
1.7B & 9B & 22B @MN5
Extreme-Scale
300K+ downloads

EuroBERT
<1B & 2B @Adastra
150K+ downloads

2023 — 2024 — 2025

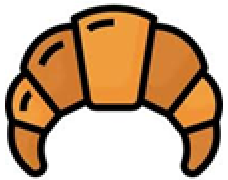
SaulLM
7B & 45B & 144 @ADASTRA
Extreme-Scale
100K+ downloads

ModernVBERT
1B <@JEANZAY
Small-Scale
10K+ downloads

La plupart de ces modèles ont été conçus avec une optique industrielle.

Une suite de modèles européens

CroissantLLM
1.3B @Jean Zay
100K+ downloads



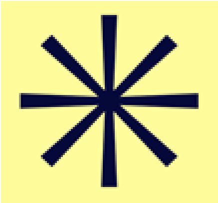
Tower
2B, 7B, 13B, 70B @MN5
AI & Data Intensive
200K+ downloads

2023

2024

SaulLM
7B & 45B & 144 @ADASTRA
Extreme-Scale
100K+ downloads

EuroLLM
1.7B & 9B & 22B @MN5
Extreme-Scale
300K+ downloads



2025

EuroBERT
<1B & 2B @Adastra
150K+ downloads

ModernVBERT
1B <@JEANZAY
Small-Scale
10K+ downloads

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'ESPACE

CROISSANT LLM : un petit modèle de langue

Un partenariat industriel et universitaire pour un SLM souverain

CroissantLLM: A Truly Bilingual French-English Language Model

Manuel Faysse[1,5] **Patrick Fernandes**[6,8,11] **Nuno M. Guerreiro**[2,5,6,8]
António Loison[1] Duarte M. Alves[6,8] Caio Corro[9] Nicolas Boizard[4,5]
João Alves[2] Ricardo Rei[2,7,8] Pedro H. Martins[2] Antoni Bigata Casademunt[10]
François Yvon[9] André F.T. Martins[2,6,8] Gautier Viaud[1] Céline Hudelot[5]
Pierre Colombo[3,5]

ILLUIN TECHNOLOGY
TÉCNICO LISBOA
Unbabel
CentraleSupélec
cnrs — Institut du Développement et des Ressources en Informatique Scientifique

Manuel Faysse
CentraleSupélec, MICS
PhD student

Pierre Colombo
CentraleSupélec, MICS
Assistant Professor

Céline Hudelot
CentraleSupélec, MICS
Professor

Nicolas Boizard
CentraleSupélec, MICS
Diabolocom
PhD Student

Croissant LLM

Comment le construire ?

Option 1: Pré-entraînement from “scratch”

- Entraîner un nouveau modèle à partir de zéro;
- Contrôle total ;
- Nécessite beaucoup de puissance de calcul !

Option 2: Pre-entraînement continu

- On part d’un modèle existant et on continue son pré-entraînement;
- On ne connait pas tout du modèle d’origine;
- Ne pas repartir de zéro.

Un projet de recherche avec des motivations industrielles

Recherche

CroissantLLM est un projet de recherche visant à étudier l'impact du **bilinguisme** sur le pré-entraînement et les performances des modèles de langue.

Industrie

Un modèle conçu pour être **suffisamment petit pour** fonctionner en local, mais suffisamment **performant** pour exécuter des tâches génératives souvent réservées à des modèles plus volumineux (**entraînement optimal pour l'inférence**). Il est entraîné uniquement sur des **données sous licence permissive.**

Open-Source

Basé sur l'open source, avec des modèles, des données, des bases de code et des critères d'évaluation **ouverts,** permettant ainsi aux **chercheurs et aux praticiens** d'en tirer profit.

Le modèle (lois d'échelle de Chinchilla)

Les modèles génératifs sont souvent des décodeurs (GPT, Mistral, LLaMa) dont les performances sont étroitement liées (1) au **nombre de paramètres du modèle** et (2) au **nombre de tokens d'entraînement.**

Entraîner le meilleur modèle étant donné un budget de calcul fixé

Pour un budget de calcul donné, il existe un rapport optimal entre le nombre de paramètres et la taille des données d'entraînement. (~20 selon les lois d'échelle de Chinchilla)

OU

Entraîner le meilleur modèle pour une taille fixée

En entraînant plus longtemps que le ratio Chinchilla, nous continuons à améliorer le modèle, mais les gains de performance sont de plus en plus coûteux.

Notre choix : surentraîner un petit modèle (1,3B) → rapport token:param de 2307 contre 20 optimal pour Chinchilla..

Plus léger

Plus rapide

Capable

Coût d'entraînement

Les données

L'entraînement des LLMs nécessite d'énormes quantités de données… En français et sous licence permissive, il est encore plus difficile d'en rassembler suffisamment !

Corpus

Culture
Données sous licence permissive

- Livres dans le domaine public
- Podcasts
- Poèmes
- Paroles de chansons
- Sous titres de films

Business
Données de l'industrie et des administrations

- Corpus de lois
- Débats parlementaires
- Décisions administratives
- Documents commerciaux publics

Connaissance
Données scientifiques et factuelles

- Encyclopédies
- Manuels scolaires
- Résumés de thèses
- Publications scientifiques

Traduction
Données parallèles Anglais-Français

- Une quantité considérable de paires de traduction provenant de différents domaines
- Filtrées à l'aide de méthodes d'estimation de la qualité à l'état de l'art

Internet
Données filtrées du web

- Données à l'échelle du Web filtrées pour obtenir des textes français et anglais de haute qualité
- Code Github sous licence libre

Les données

L'entraînement des LLMs nécessite d'énormes quantités de données… En français et sous licence permissive, il est encore plus difficile d'en rassembler suffisamment !

Corpus

Culture
Données sous licence permissive

Business
Données de l'industrie et des administrations

Connaissance
Données scientifiques et factuelles

Traduction
Données parallèles Anglais-Français

Internet
Données filtrées du web

Pré-traitement des données

Identification et scrapping → Déduplication → Filtrage → Suréchantillonnage

Transparence & Open-Source

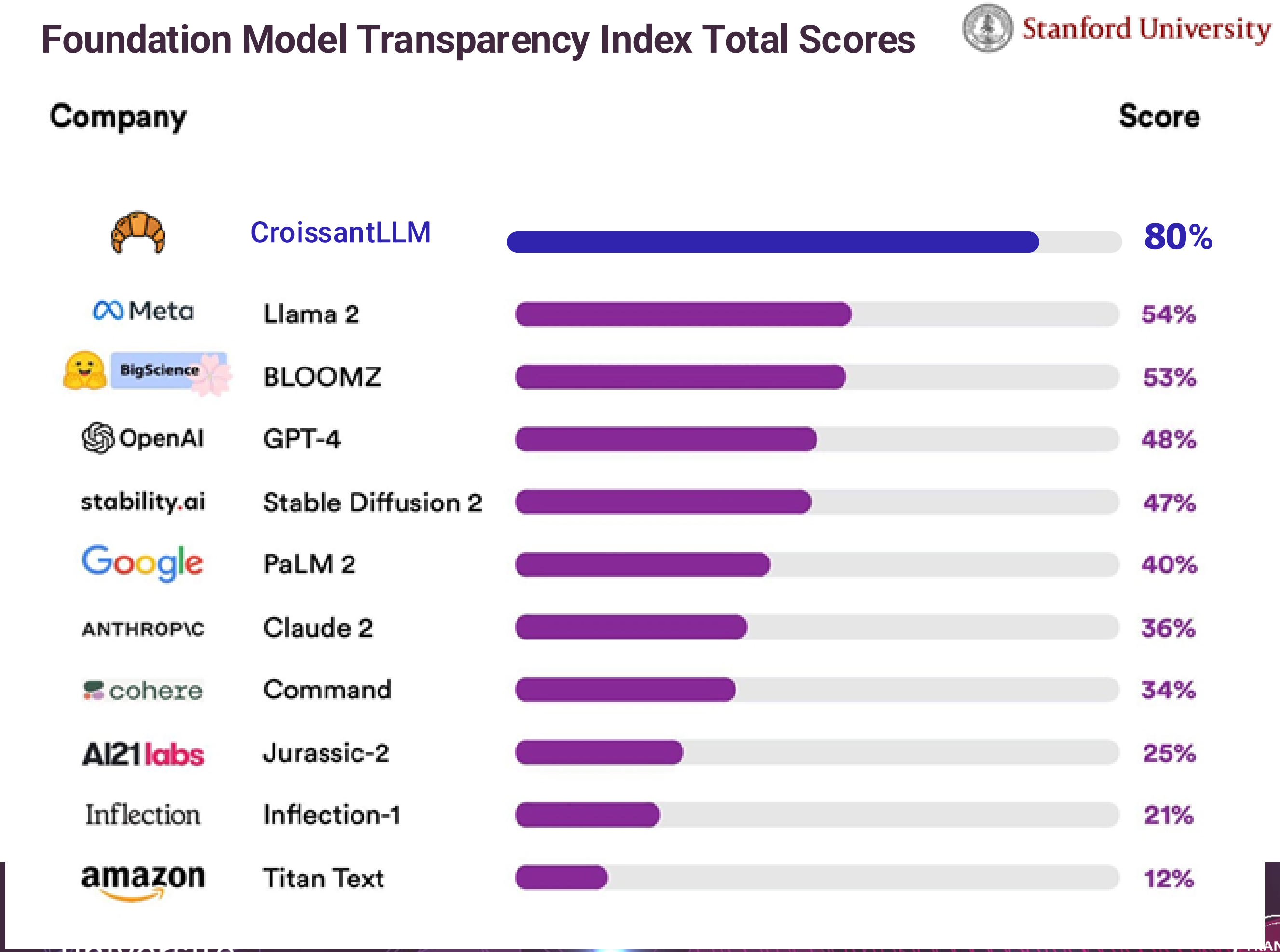
Projet fondé sur la transparence, destiné à servir de ressource utile aux professionnels de l'industrie et aux chercheurs !

Processus d'entraînement documenté du début à la fin

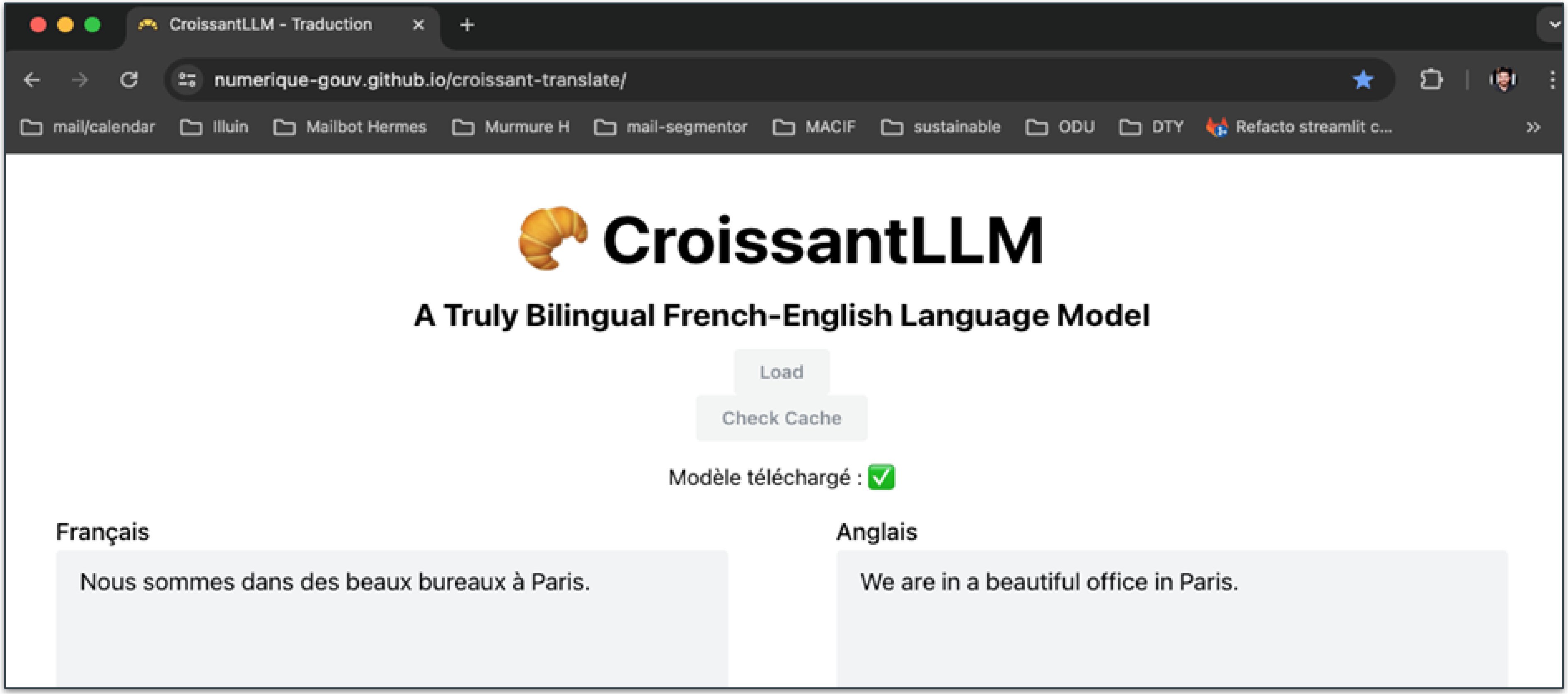
Ouvert à tous :

- **Corpus d'entraînement**
- **Points de contrôle du modèle**
- **Critères d'évaluation**
- **Bases de code**

Pas de restrictions d'usage (MIT)



Application : Traduction dans le navigateur (DINUM)



https://securitrad.centralesupelec.fr

Une suite de modèles européens

CroissantLLM
1.3B @Jean Zay
100K+ downloads

Tower
2B, 7B, 13B, 70B @MN5
AI & Data Intensive
200K+ downloads

SaulLM
7B & 45B & 144 @ADASTRA
Extreme-Scale
100K+ downloads

EuroLLM
1.7B & 9B & 22B @MN5
Extreme-Scale
300K+ downloads

EuroBERT
<1B & 2B @Adastra
150K+ downloads

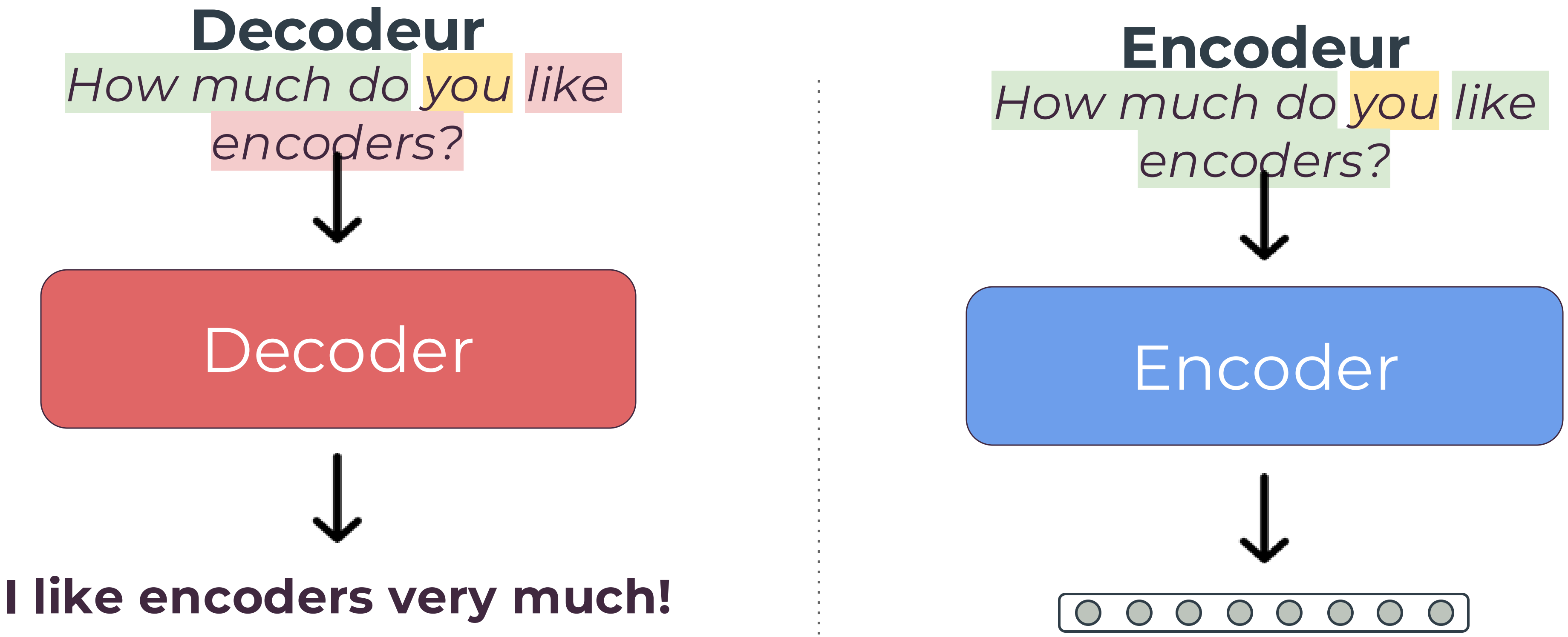
ModernVBERT
1B <@JEANZAY
Small-Scale
10K+ downloads

2023 — 2024 — 2025

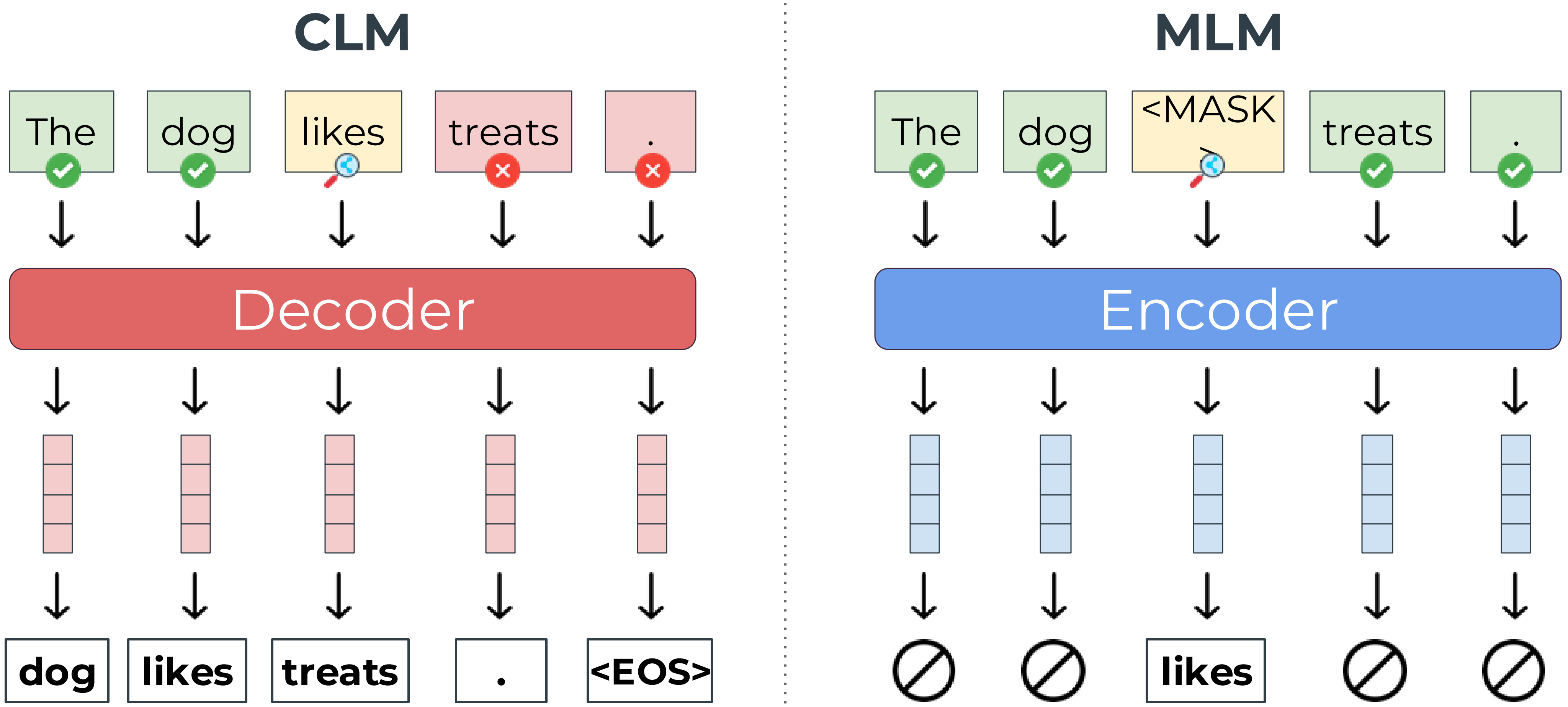
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'ESPACE

Decodeurs vs. Encodeurs

À l'ère moderne du NLP, deux types principaux de modèles dominent : les modèles génératifs à encodeur et à décodeur uniquement.



Stratégies de pré-entrainement : CLM vs. MLM



$$\mathcal{L}_{\mathrm{CLM}} = -\sum_{t=1}^{T} \log P(x_t \mid x_{<t})$$

$$\mathcal{L}_{\mathrm{MLM}} = -\sum_{i \in \mathcal{M}} \log P(x_i \mid x_{\setminus \mathcal{M}})$$

Focus sur deux travaux récents sur les encodeurs

EuroBERT: Scaling Multilingual Encoders for European Languages
COLM (2025)

Should We Still Pretrain Encoders with Masked Language Modeling?
Under review

Pourquoi des encodeurs ?

Avec l'explosion récente des modèles d'IA générative, les chercheurs ont tendance à oublier que les encodeurs restent de loin les modèles de texte d'apprentissage profond les plus utilisés dans l'industrie !

Polyvalence
Les représentations des encodeurs peuvent être utilisées pour des tâches telles que la classification, la recherche, la NER, les questions-réponses …

Efficacité
Les modèles encodeurs sont intrinsèquement plus compacts que les modèles décodeurs, ce qui les rend nettement plus rapides tant pour le réglage fin que pour l'inférence.

Opportunités
Grâce aux récentes avancées sur les décodeurs (améliorations en termes de qualité des données, d'échelle et d'architecture), il existe un fort potentiel pour réaliser des progrès similaires dans le domaine des encodeurs.

Aperçu des modèles EuroBERT

- → **3 tailles** adaptées à diverses contraintes
- → **5T tokens** vus pendant l'entraînement
- → **15 languages** supportés
- → **8k taille du contexte**
- → **Entièrement Open-Source**

EuroBERT 210M | **EuroBERT 610M** | **EuroBERT 2.1B**

Architecture

Les modèles EuroBERT comptent entre 210 millions et 2,1 milliards de paramètres, tous entraînés sur le même nombre de tokens et utilisant des configurations architecturales identiques.

Architecture inspire de Llama 3

- > Llama 3 tokenizer
- > Grouped Query Attention
- > Rotary Position Embeddings
- > RMS Normalization
- > SwiGLU layers
- > Non-causal masking

Recettes pour l'apprentissage de bout en bout

Pre-training

Data
> Large quantity of unsupervised data
> Tokens masked randomly

Learning objective
> MLM: model trained to retrieve the masked tokens

Annealing

Data
> Smaller quantity of unsupervised data
> Higher quality

Learning objective
> MLM
> Learning rate decreased to 0

Fine-tuning

Data
> Depending on use case (classification, regression…)

Learning objective
> Depending on use case (CE , MSE…)

→ **Evaluation**

Boizard et al, COLM 2025
EuroBERT: Scaling Multilingual Encoders for European Languages
https://arxiv.org/abs/2503.05500
https://huggingface.co/collections/EuroBERT/eurobert-67ceb6c01804878b1f7999c6

Recettes pour l’apprentissage de bout en bout

Pre-training

Data
> Large quantity of unsupervised data
> Tokens masked randomly

Learning objective
> MLM: model trained to retrieve the masked tokens

Annealing

Data
> Smaller quantity of unsupervised data
> Higher quality

Learning objective
> MLM
> Learning rate decreased to 0

Fine-tuning

Data
> Depending on use case (classification, regression…)

Learning objective
> Depending on use case (CE , MSE…)

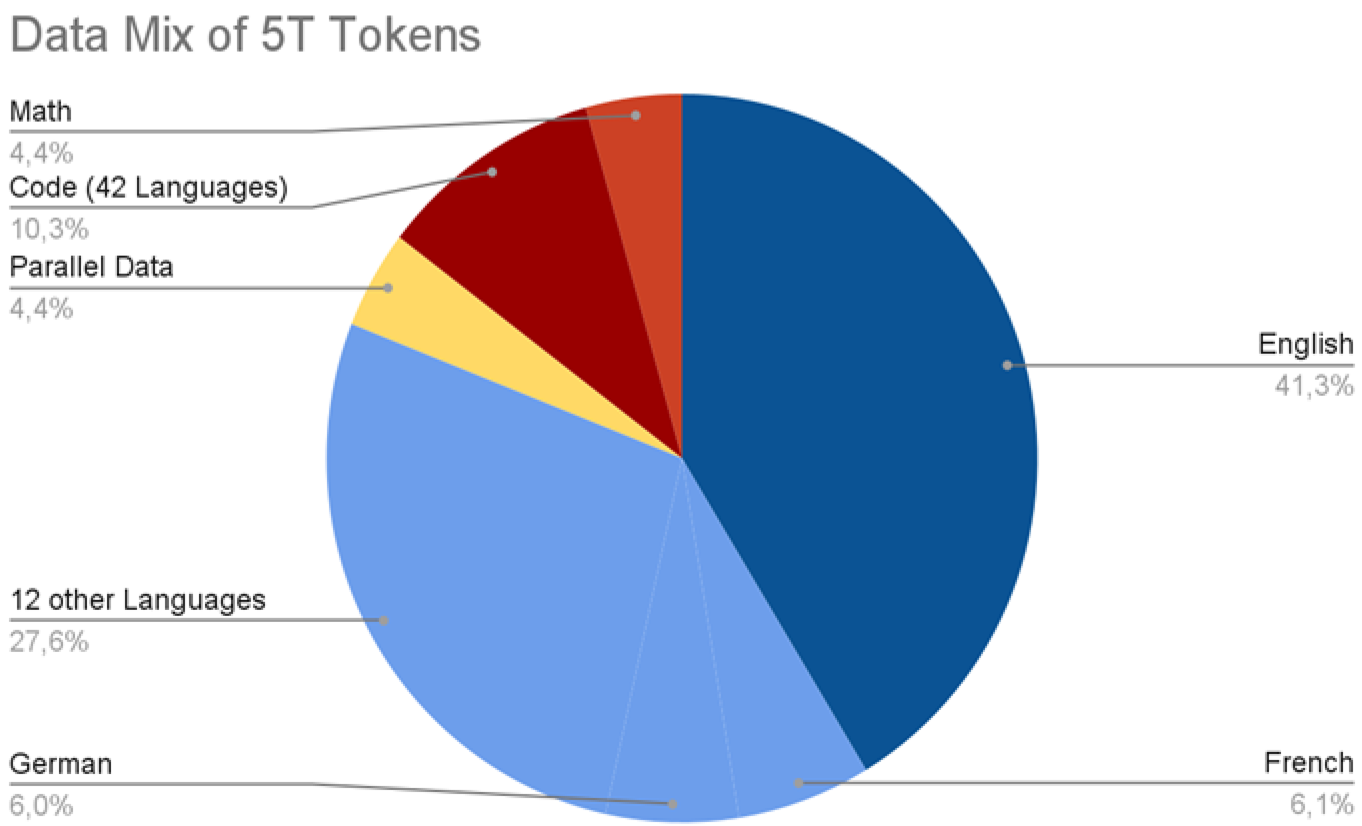
→ **Evaluation**

Boizard et al, COLM 2025
EuroBERT: Scaling Multilingual Encoders for European Languages
https://arxiv.org/abs/2503.05500
https://huggingface.co/collections/EuroBERT/eurobert-67ceb6c01804878b1f7999c6

Données pour le pré-entraînement

Un bon mix est la clé

- **Mix**: un modèle est performant sur les domaines qu'il a observés.
- **Math & code**: pour le raisonnement
- **Parallel data**: pour transférer des connaissances entre les langues

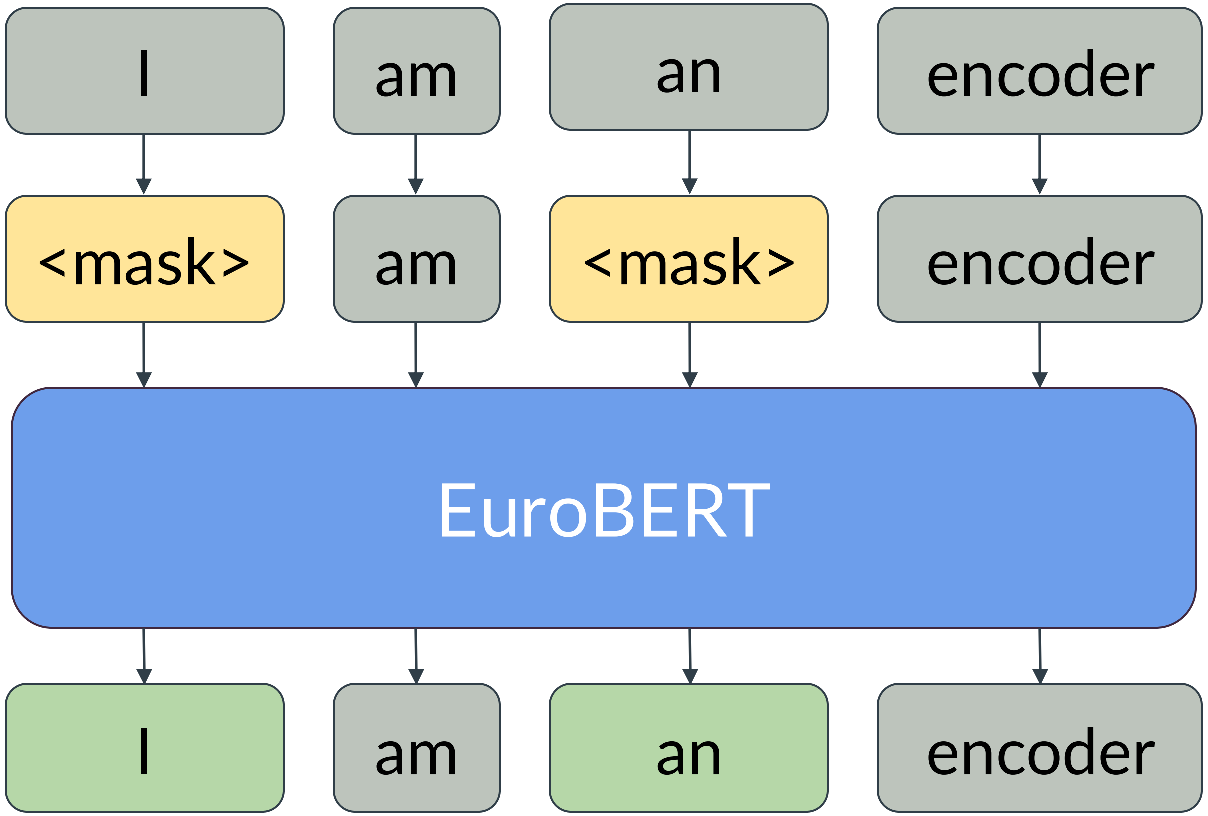


Stratégie de masquage

Stratégie de masquage

> **50%** de la phrase est masquée[1]

> **Seulement du masquage**: pas de remplacement par des jetons aléatoires (cf. entraînement de type BERT)



[1]Alexander Wettig et al., "*Should You Mask 15% in Masked Language Modeling?*"

: *Cross-entropy on vocabulary space*

Recettes pour l'apprentissage de bout en bout

Pre-training	Annealing	Fine-tuning
Data > Large quantity of unsupervised data > Tokens masked randomly	**Data** > Smaller quantity of unsupervised data > Higher quality	**Data** > Depending on use case (classification, regression…)
Learning objective > MLM: model trained to retrieve the masked tokens	**Learning objective** > MLM > Learning rate decreased to 0	**Learning objective** > Depending on use case (CE , MSE…)

→ **Evaluation**

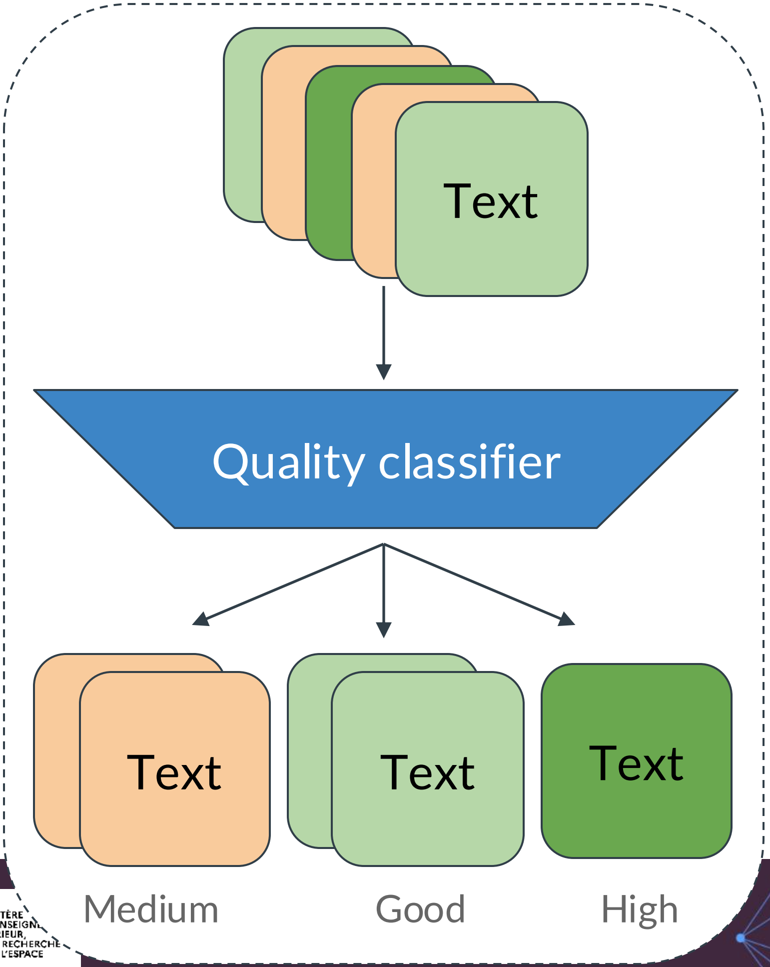
Boizard et al, COLM 2025
EuroBERT: Scaling Multilingual Encoders for European Languages
https://arxiv.org/abs/2503.05500
https://huggingface.co/collections/EuroBERT/eurobert-67ceb6c01804878b1f7999c6

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'ESPACE

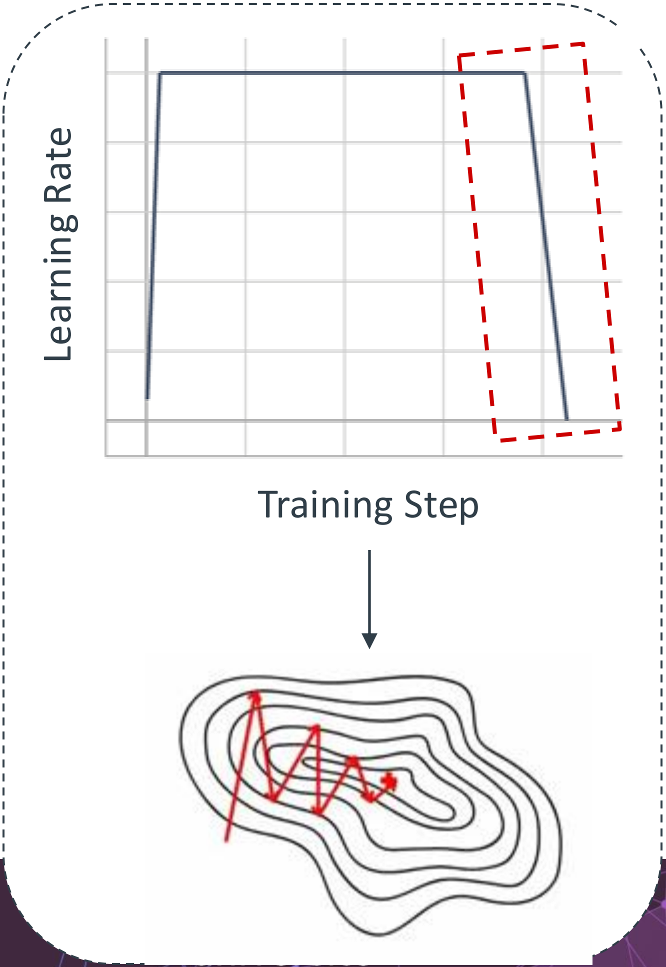
FRANCE 2030

Procédure d'Annealing

Filtrage – Qualité des données



Learning Rate Decay



Performance améliorée

- **Meilleure adéquation** avec les données du monde « réel »
- **Convergence plus efficace** vers un minimum local

Recettes pour l'apprentissage de bout en bout

Pre-training

Data
> Large quantity of unsupervised data
> Tokens masked randomly

Learning objective
> MLM: model trained to retrieve the masked tokens

Annealing

Data
> Smaller quantity of unsupervised data
> Higher quality

Learning objective
> MLM
> Learning rate decreased to 0

Fine-tuning

Data
> Depending on use case (classification, regression...)

Learning objective
> Depending on use case (CE , MSE...)

→ **Evaluation**

Boizard et al, COLM 2025
EuroBERT: Scaling Multilingual Encoders for European Languages
https://arxiv.org/abs/2503.05500
https://huggingface.co/collections/EuroBERT/eurobert-67ceb6c01804878b1f7999c6

Protocole d'évaluation

Protocole d'évaluation

> **<u>Tâches</u>**: retrieval, sequence classification & regression, token classification

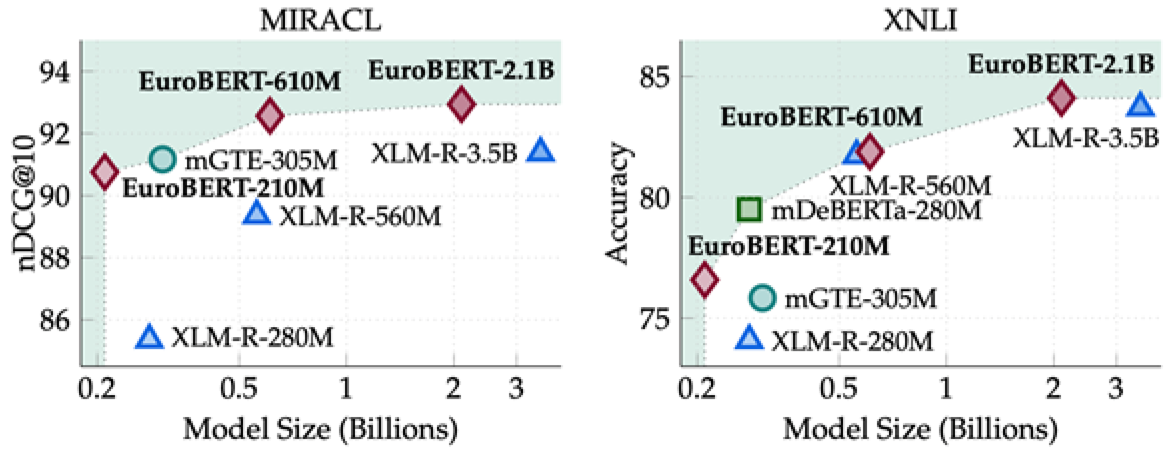
> **<u>Baselines</u>**: XLM-RoBERTa, mDeBERTa, mGTE, ModernBERT

> **<u>Fine-tuning</u>**: Tous les modèles sont soumis au même protocole afin de garantir une comparaison équitable.

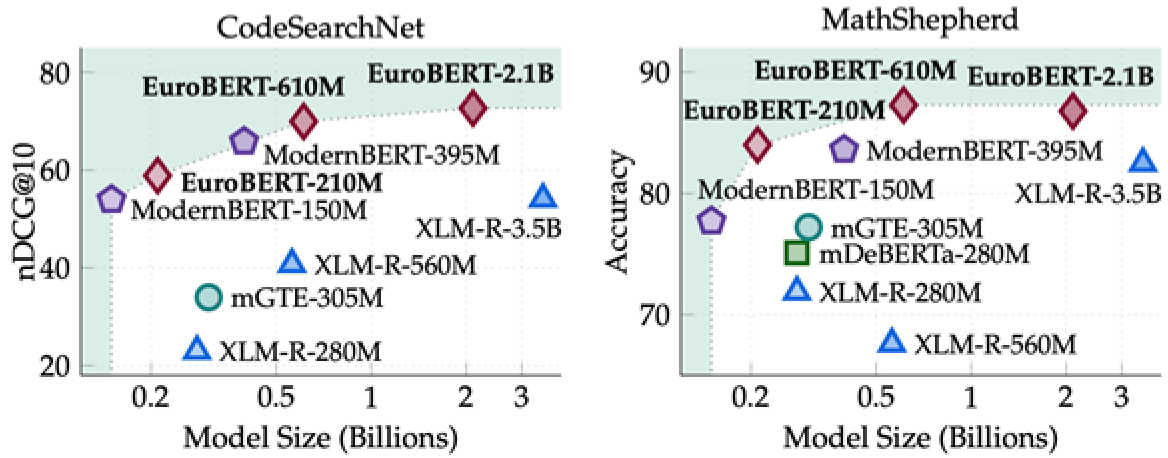
Task	European Languages								Extra-European Languages							Code	Math
	en	de	es	fr	it	nl	pl	pt	ar	hi	ja	ru	tr	vi	zh		
							Information Retrieval										
MIRACL	✓		✓	✓						✓	✓	✓	✓			✓	
MLDR	✓	✓	✓	✓	✓			✓		✓	✓	✓				✓	
Wikipedia	✓	✓			✓	✓		✓				✓					
CC-News	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
CodeSearchNet	✓															✓	
DupStackMath	✓															✓	
MathFormula	✓																✓
							Sequence Classification										
XNLI	✓	✓	✓	✓					✓	✓		✓	✓	✓	✓		
PAWS-X	✓	✓	✓	✓													
AmazonReviews	✓	✓	✓	✓							✓				✓		
MassiveIntent	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
CodeDefect	✓															✓	
CodeComplexity	✓															✓	
MathShepherd	✓																✓
							Sequence Regression										
WMT	✓	✓		✓			✓				✓	✓	✓		✓		
SeaHorse	✓	✓	✓										✓	✓			
							Token Classification										
NER	✓	✓	✓			✓											

Résultats

Tâches multilingues



Tâches codes et maths

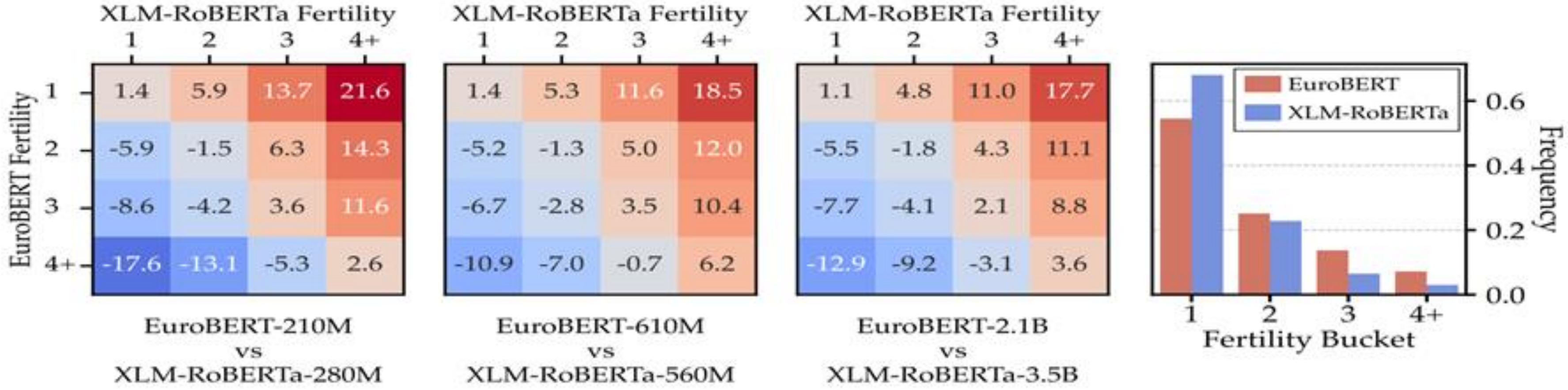


Performance d' EuroBERT

- **<u>Multilingue</u>**: EuroBERT égale ou surpasse les modèles de taille similaire sur les benchmarks multilingues.

> **<u>Math & Code</u>**: EuroBERT obtient des résultats SOTA dans les tâches mathématiques et liées au code

Token Classification



Impact de la fertilité du tokenizer sur la performance en NER

> **Observation**: Les modèles EuroBERT sont moins performants que leurs homologues de taille similaire en NER, malgré des performances globales raisonnables.

> **Explication** : Une fertilité plus élevée implique la division des entités en plusieurs sous-tokens, ce qui dilue leur cohérence sémantique et complique ainsi la supervision.

Devons-nous continuer à pré-entraîner les encodeurs avec MLM ?

EuroBERT: Scaling Multilingual Encoders for European Languages
Accepted @ COLM (10/25)

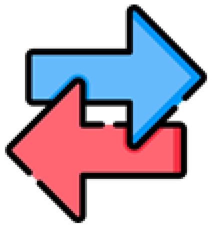
Should We Still Pretrain Encoders with Masked Language Modeling?
Preprint (07/25)

- **Les gains minimes** obtenus entre 600 millions et 2 milliards d'EuroBERT ont soulevé des questions sur la mise à l'échelle traditionnelle des encodeurs.
- **EuroBERT est axé sur les données ;** Examen du rôle de l'objectif de pré-entraînement et de la bidirectionnalité.

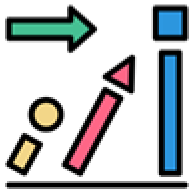
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'ESPACE

FRANCE 2030

Motivations



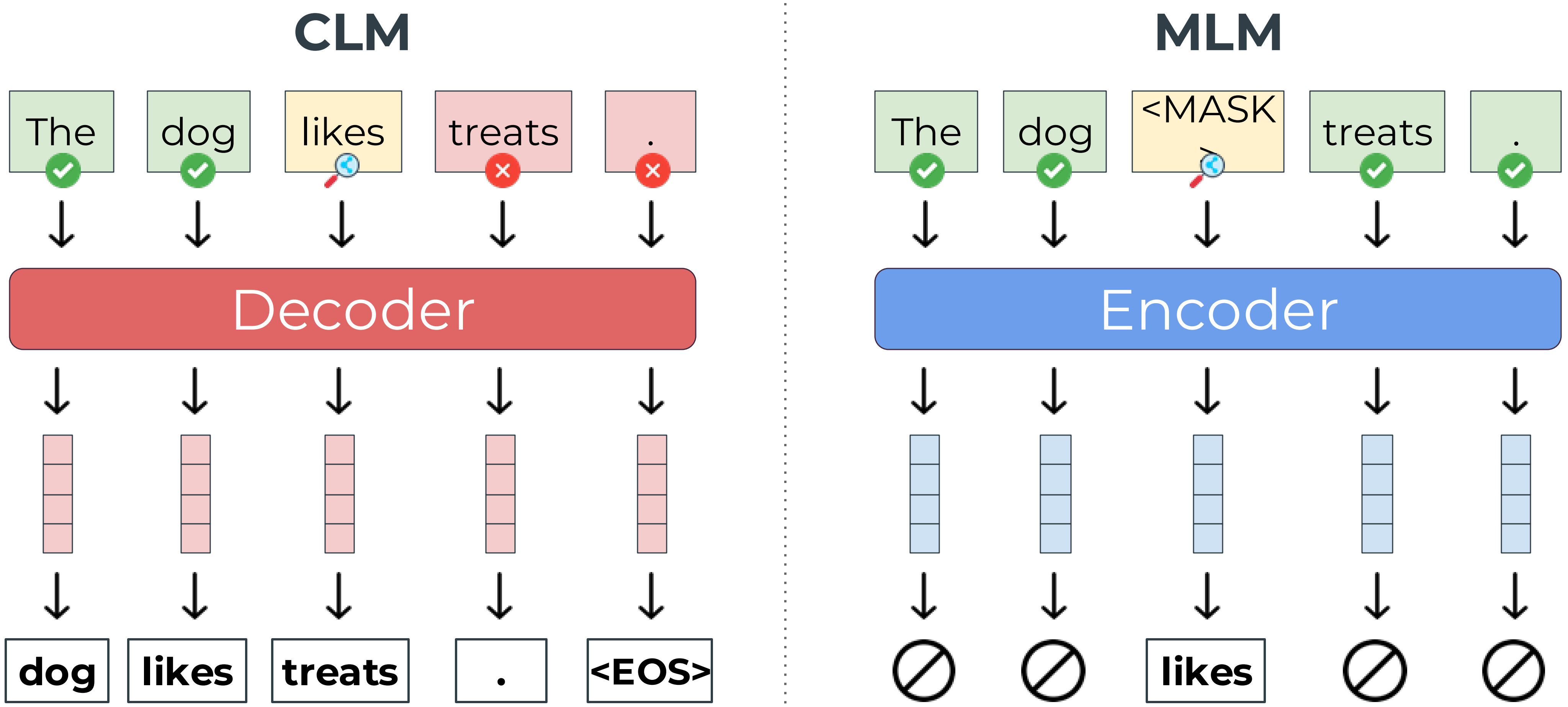
Entraînement bidirectionnel
> L'entrainement bidirectionnel est **l'approche standard** pour apprendre de bonnes representations (e.g., BERT)



Pre-entraînement causal et apprentissage de représentations
> Plus récemment, les modèles génératifs (qui fonctionnent généralement à une échelle beaucoup plus grande) ont été adaptés à l'apprentissage de représentation, obtenant des résultats **à l'état de l'art sur les benchmarks**.

Question de recherche
Dans quelle mesure le pré-entraînement causal contribue-t-il réellement à renforcer les représentations, et dans quelle mesure cet effet est-il influencé par des facteurs confondants tels que l'échelle du modèle et le régime des données ?

CLM vs. MLM



$$\mathcal{L}_{\mathrm{CLM}} = -\sum_{t=1}^{T} \log P(x_t \mid x_{<t})$$

$$\mathcal{L}_{\mathrm{MLM}} = -\sum_{i \in \mathcal{M}} \log P(x_i \mid x_{\setminus \mathcal{M}})$$

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'ESPACE

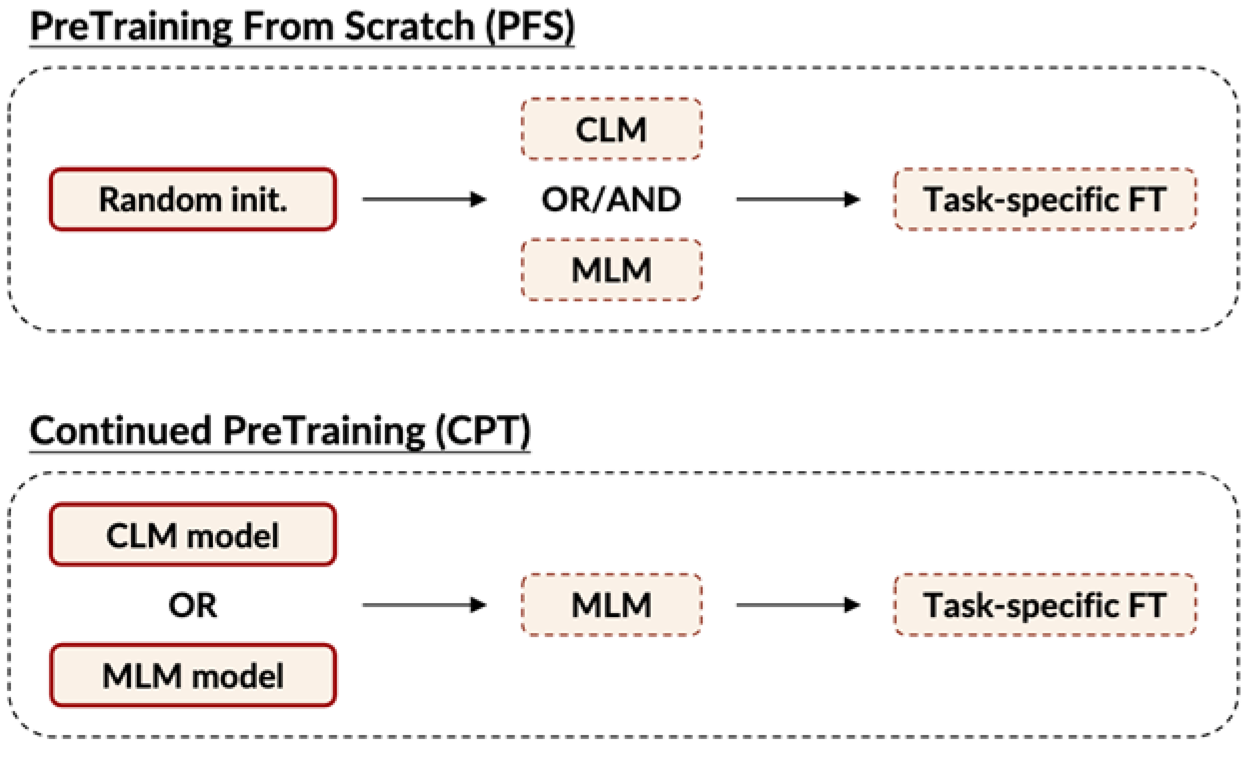
université PARIS-SACLAY

FRANCE 2030

Protocole expérimental

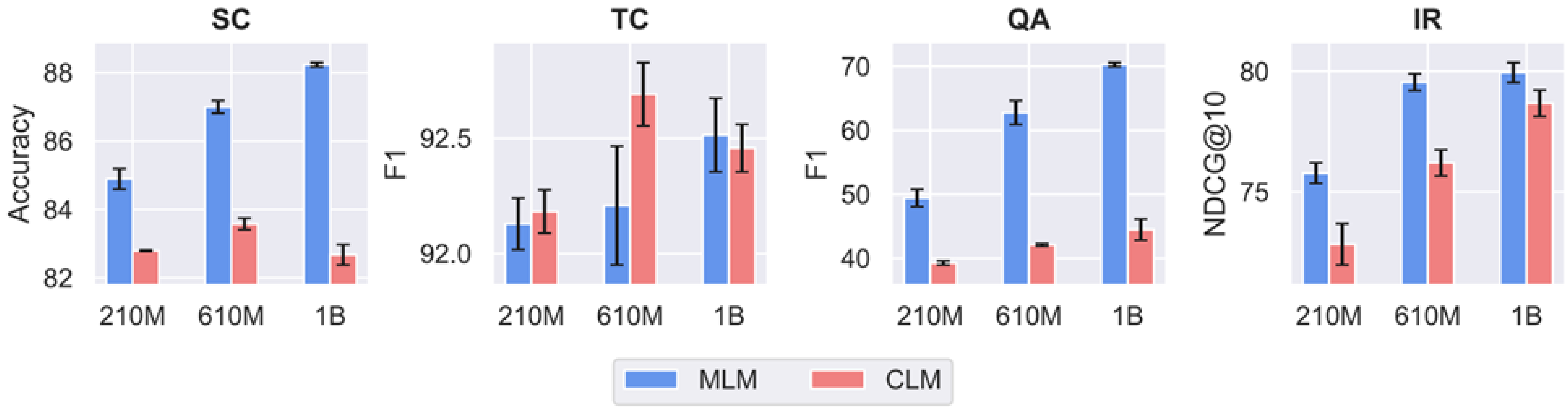
But : **Évaluer l'impact spécifique de l'objectif de pré-entraînement sur la qualité de la représentation**

> **2 cadres réalistes**: PFS, CPT
> **3 tailles de modèles**: 210M, 610M, 1B
> **Cadre unifié**: Pipelines unifiés de pré-entraînement et d'ajustement / d'évaluation



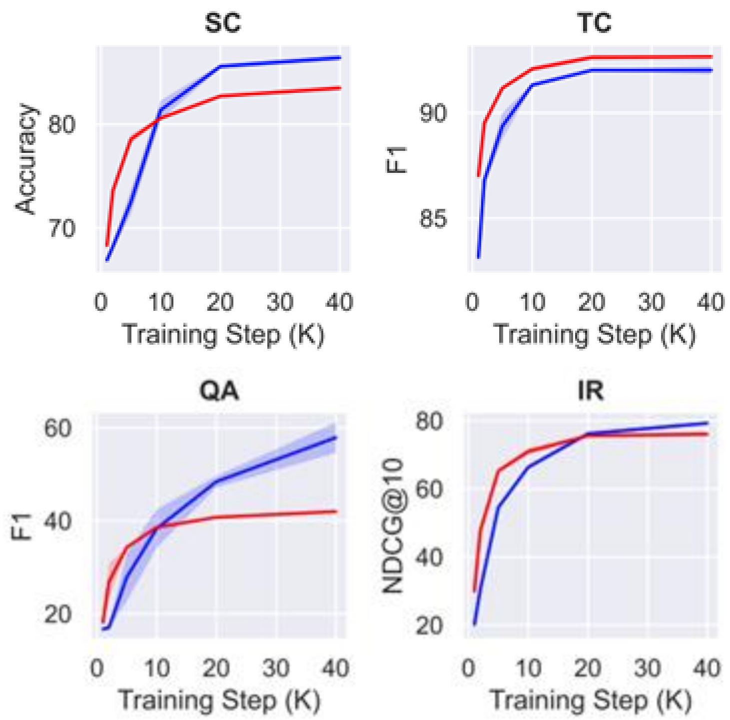
Pré-entraînement

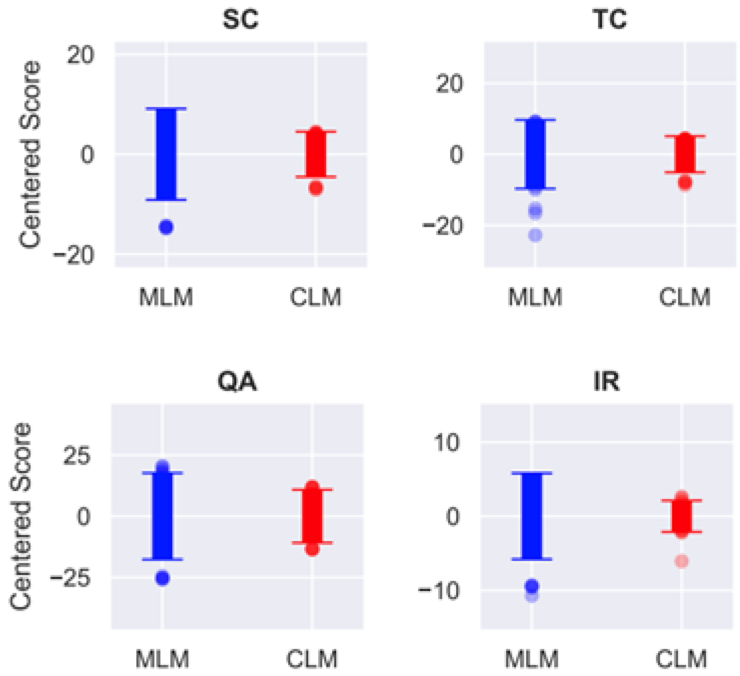
Question: Entre CLM et MLM, quel objectif est le plus efficace pour apprendre de bonnes représentations textuelles ?



- **MLM surpasse globalement CLM pour l'apprentissage des représentations textuelles.**
- **Le passage à l'échelle diffère selon la tâche** (e.g., QA vs. IR)
- **CLM égale ou surpasse MLM en matière de TC.**

Pré-entraînement avec MLM ou CML

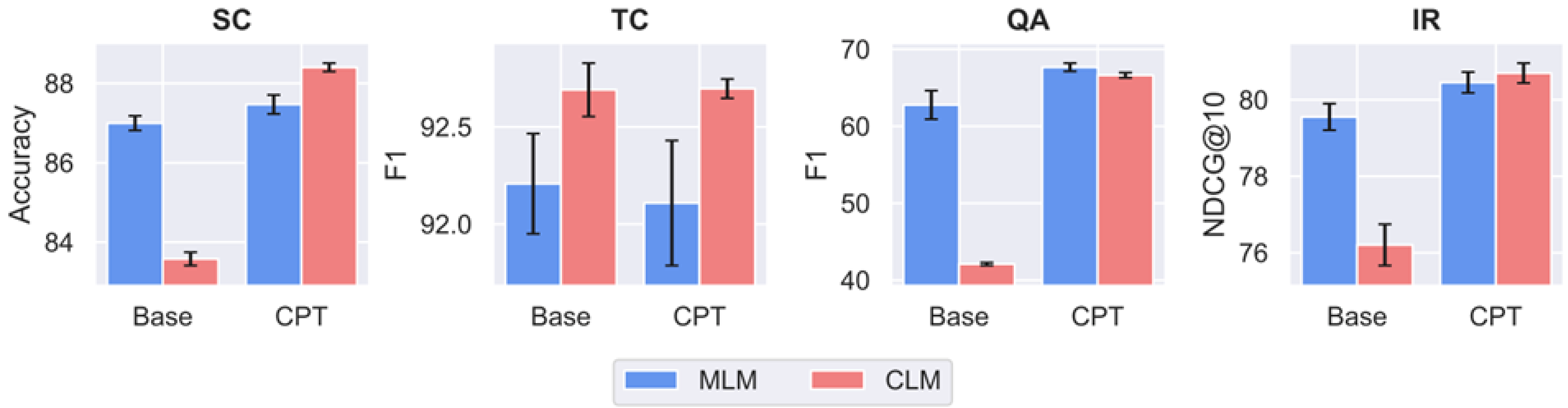




> CLM est plus **data-efficient** que MLM (gauche)
> CLM génère des modèles plus **faciles à adapter** que MLM (droite)

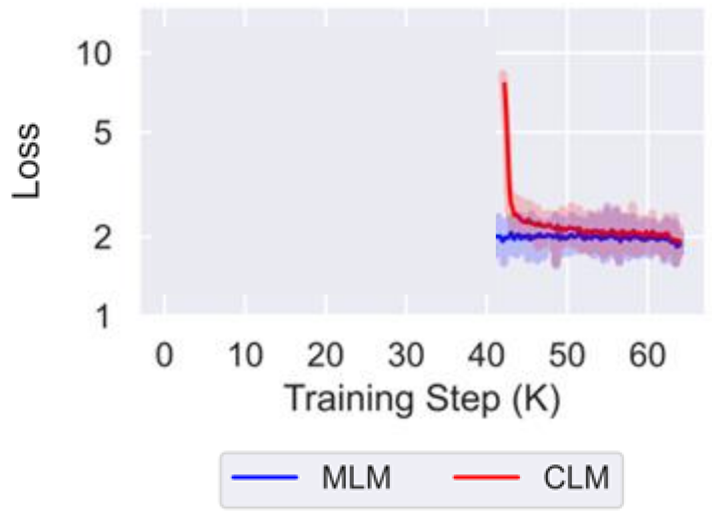
Recyclage de modèles via un pré-entraînement continu

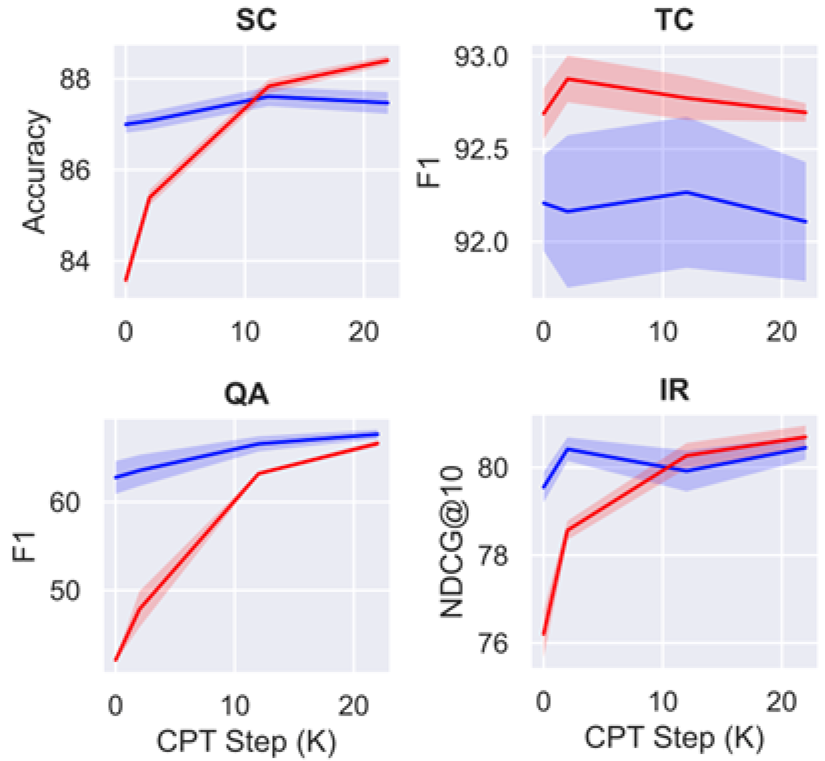
Question: Si nous voulons poursuivre le pré-entraînement d'un modèle existant, devons-nous partir d'un point de contrôle MLM ou CLM ?



> Dans l'ensemble, adapter un CLM avec MLM donne de meilleurs résultats que la poursuite du MLM sur un modèle pré-entraîné avec MLM.

Pré-entraînement continu





> **Quelques étapes de CPT** suffisent pour atteindre de bonnes performances
> Les modèles pré-entraînés avec CLM montrent une **trajectoire pus prometteuse.**

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'ESPACE

FRANCE 2030

Conclusions

- Il est possible de construire des **modèles européens ouverts, transparents et compétitifs**
- Importance de l' **ouverture des modèles, codes et données** :
  - Plus que les modèles eux-mêmes, l'ouverture dans ce domaine permet de **partager le savoir-faire et des recettes importantes**.
  - Vers des modèles entièrement ouverts incluant les **points de contrôle d'entraînement** pour étudier de nouvelles pistes d'amélioration mais aussi comprendre la dynamique de l'entraînement

2. Importance des données pour l’adaptation

Adaptation

Adapter les LLMs et les rendre accessibles au plus grand nombre impliquent de les **adapter à des données spécifiques**, de remédier aux **contraintes liées aux coûts** et de résoudre les limites en **matière d'évaluation.**

Safe Retrieval
Trustworthy Reranking: abstention mechanism

TMLR-2024

ULD - Distillation Loss
Loss using optimal transport theory to enable distillation across different architectures & tokenizers.

TMLR - 2025

ColPali – Multimodal LM
Simple Document Retrieval in the Vision Space.

ICLR - 2025

Adaptation

Adapter les LLMs et les rendre accessibles au plus grand nombre impliquent de les **adapter à des données spécifiques**, de remédier aux **contraintes liées aux coûts** et de résoudre les limites en **matière d'évaluation.**

Safe Retrieval

Trustworthy Reranking: abstention mechanism

TMLR-2024

ULD - Distillation Loss

Loss using optimal transport theory to enable distillation across different architectures & tokenizers.

TMLR - 2025

ColPali – Multimodal LM

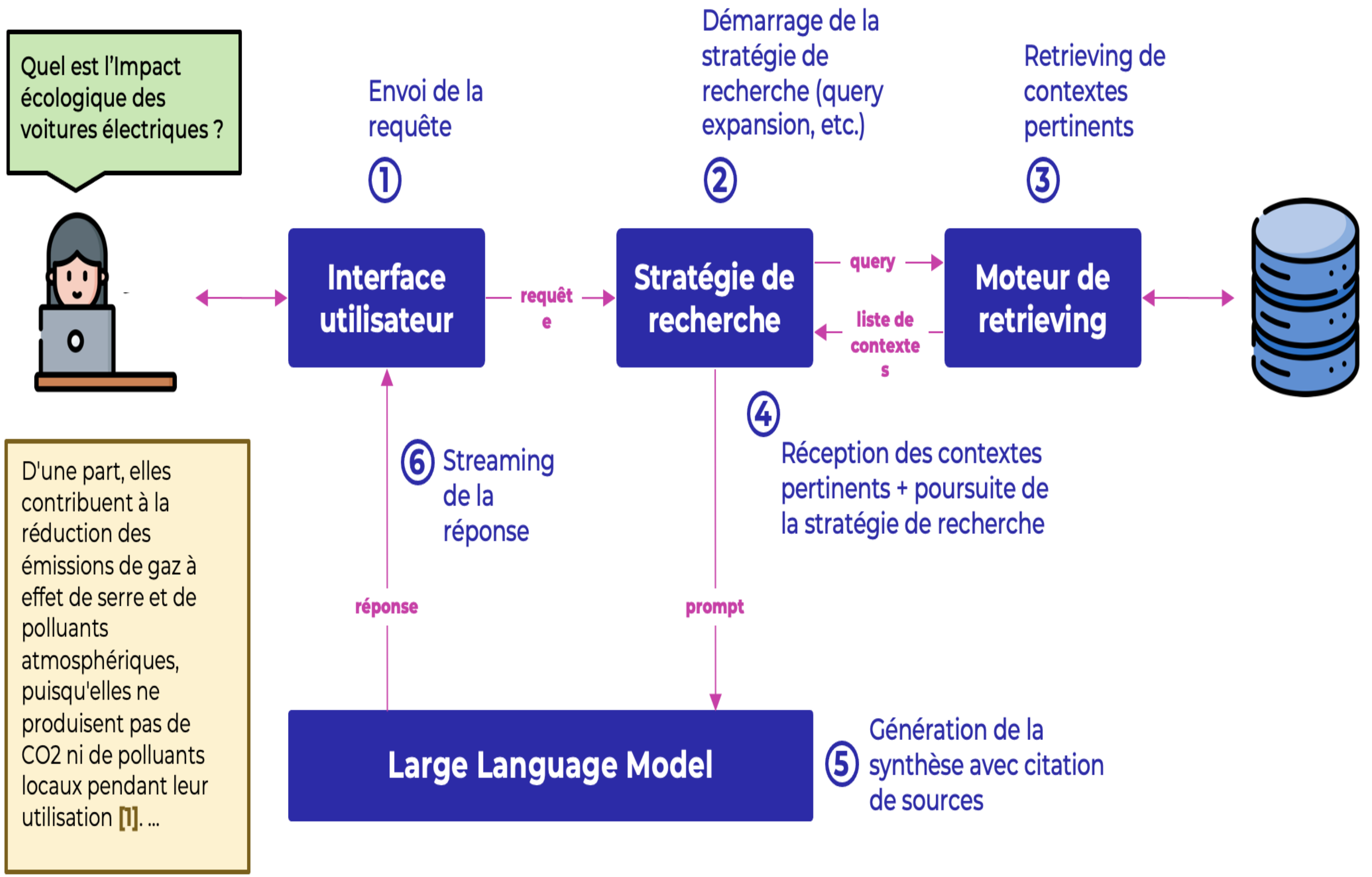
Simple Document Retrieval in the Vision Space.

ICLR - 2025

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'ESPACE

université PARIS-SACLAY

RAG: Retrieval Augmented Generation
Le cas d'usage n°1 en entreprise



Corpus privé

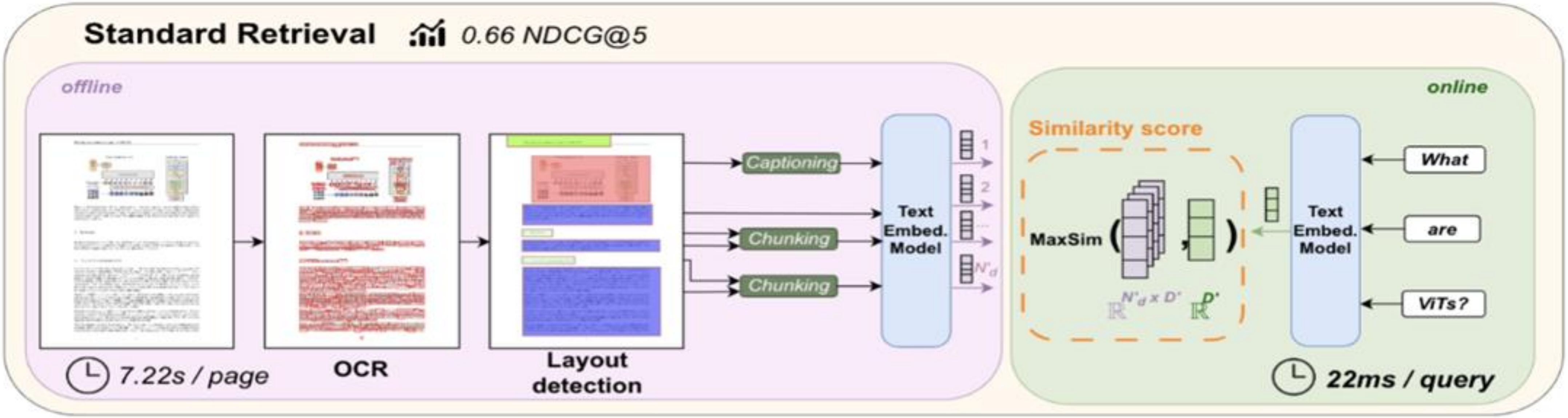
Mise à jour

Informations récentes sans ré-apprentissage

Sources

Fondements des réponses dans les documents sources

Pipeline classique pour la recherche de documents



Une chaine complexe et lente

Colpali : recherche dans le domaine visuel

Idée : supprimer complètement l'analyse des documents, en travaillant avec des « captures d'écran » des documents.

Les pipelines standard fonctionnent
avec le contenu textuel extrait.

Des pipelines complexes sont nécessaires pour extraire du texte, détecter la mise en page du document, légender les éléments visuels, intégrer le contenu textuel à l'aide de modèles d'intégration de texte...

vs.

Peut-on former des retrievers à travailler à partir d'images de documents ?

Les modèles linguistiques visuels sont utilisés pour créer directement des représentations de l'image de chaque page du document !

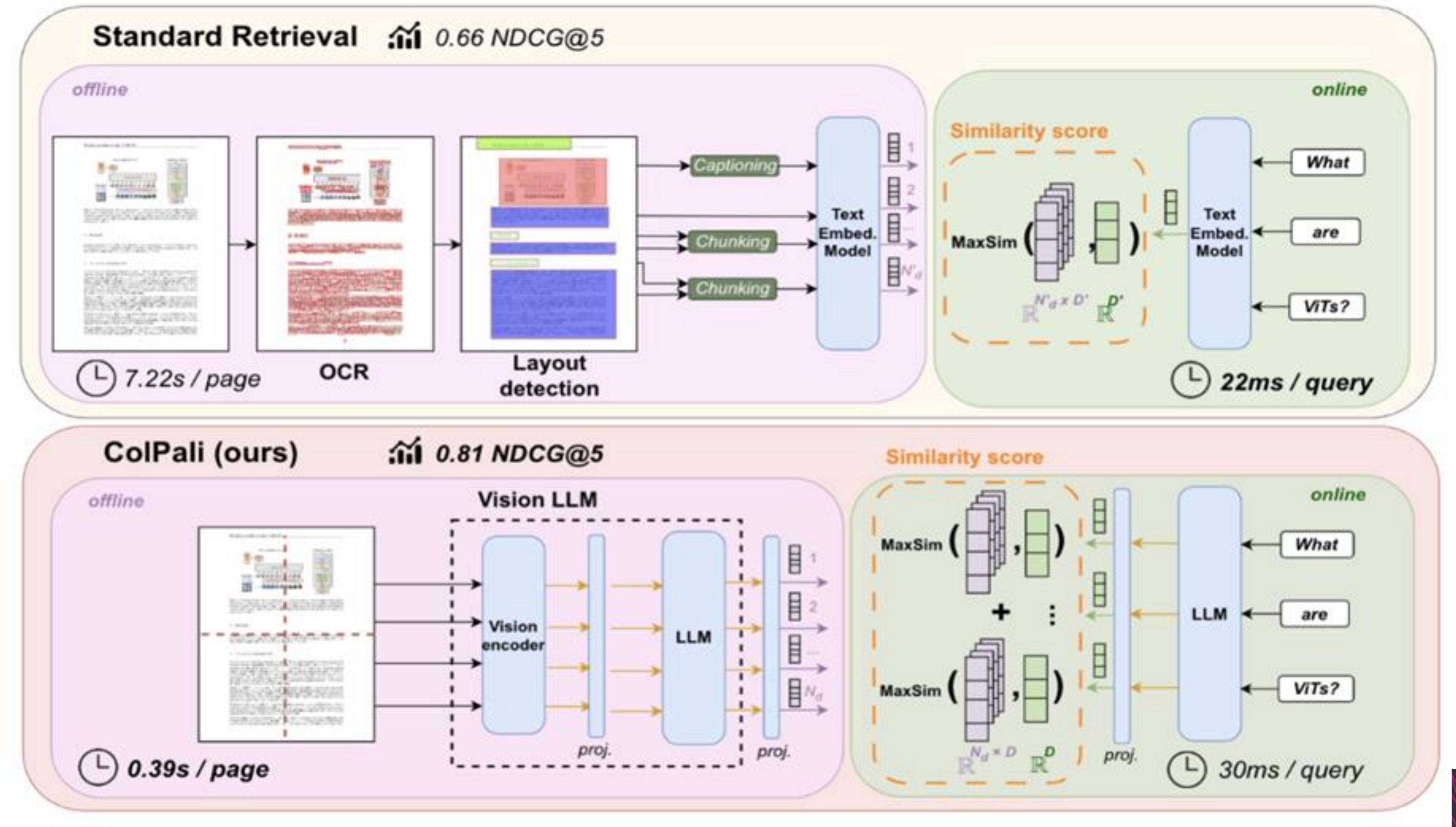
La récupération de documents à partir d'images est un concept totalement nouveau... Est-ce que cela fonctionne bien ?

Capable

Plus rapide

Entrainable

Colpali : recherche dans le domaine visuel



Colpali : le benchmark ViDoRe

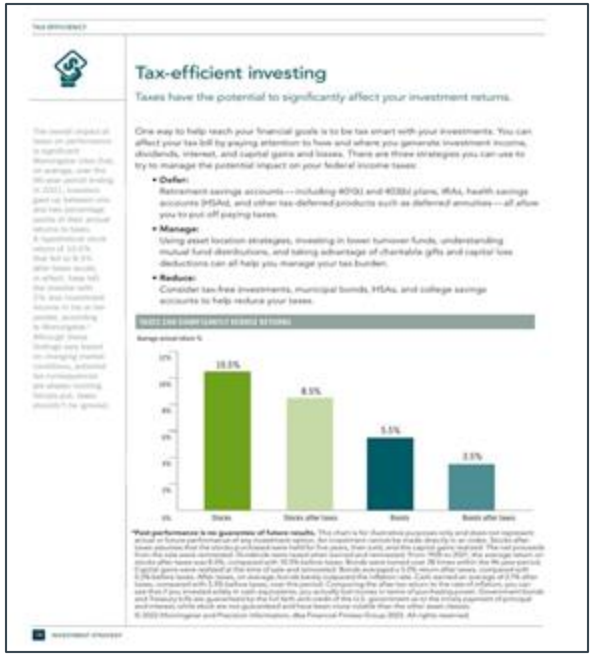
ViDoRe, Visual Document Retrieval Benchmark (référentiel de recherche documentaire visuelle), qui permet d'évaluer la capacité des moteurs de recherche à **extraire des informations riches sur le plan visuel** dans des documents, avec des tâches couvrant divers **sujets, modalités** (figures, tableaux, texte) et **langues** !

Dataset	# Queries	Domain
Academic Tasks		
DocVQA (eng)	500 (500)	Industrial
InfoVQA (eng)	500 (500)	Infographics
TAT-DQA (eng)	1600 (1600)	Varied Modalities
arXiVQA (eng)	500 (500)	Scientific Figures
TabFQuAD (fra)	210 (210)	Tables
Practical Tasks		
Energy (eng)	100 (1000)	Scientific
Government (eng)	100 (1000)	Administrative
Healthcare (eng)	100 (1000)	Medical
AI (eng)	100 (1000)	Scientific
Shift Project (fra)	100 (1000)	Environment

Table 1: *ViDoRe* comprehensively evaluates multimodal retrieval methods. The size of the document corpus is indicated in parentheses.

Colpali : le benchmark ViDoRe

Exemples de paires (document, requête)


Tax-efficient investing

Taxes have the potential to significantly affect your investment returns.

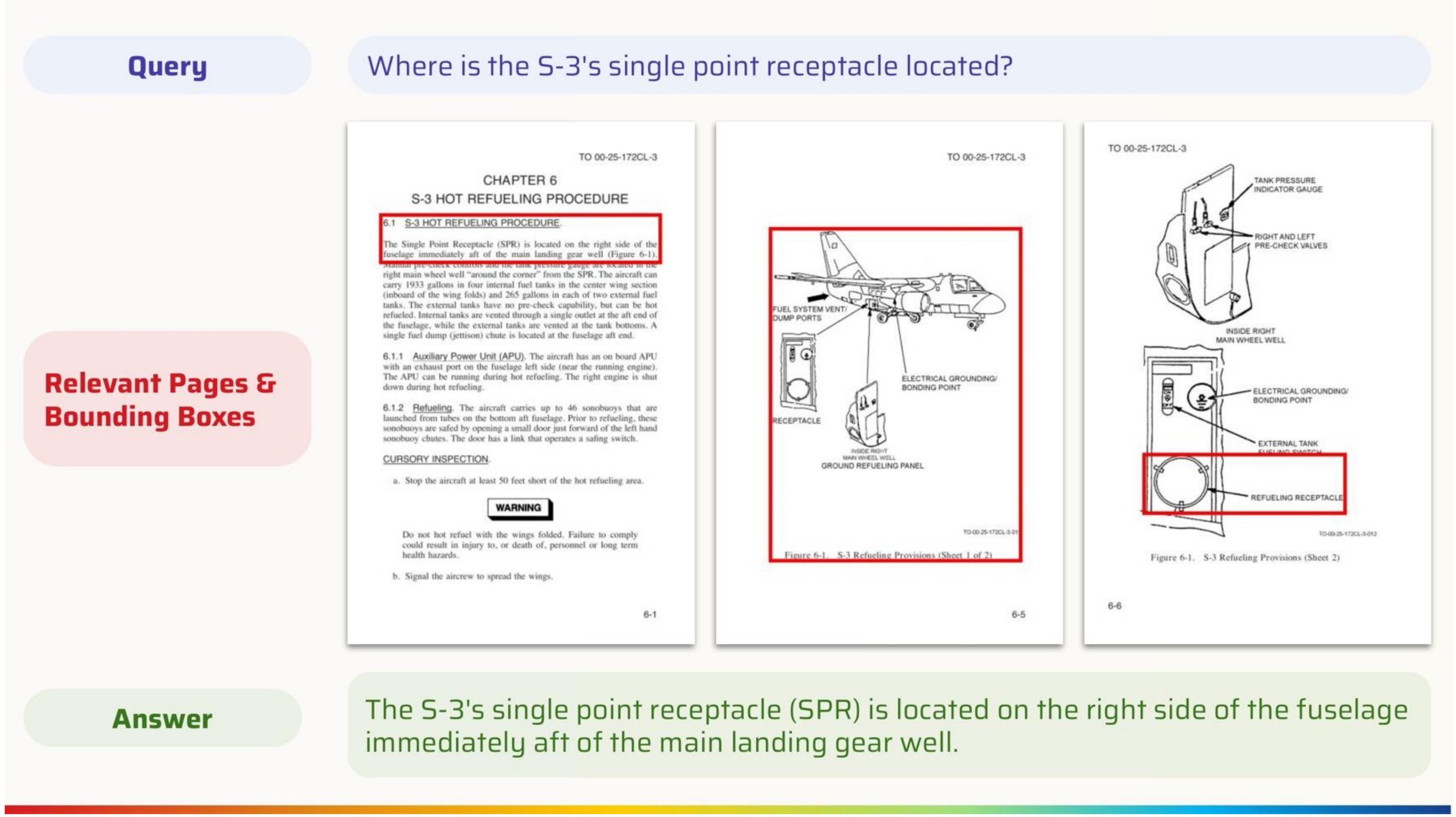
What types of accounts or products allow investors to defer paying taxes?


Technology-Aided Instruction and Intervention (TAII)

What are some common outcome areas targeted by TAII for different age groups?

ViDoRe V2 et V3

- ViDoRe V2 a étendu le benchmark à des requêtes plus ouvertes.
- ViDoRe V3 conçu pour établir une nouvelle norme d'excellence dans le secteur de l'évaluation multimodale de la recherche de documents d'entreprise (avec NVIDIA)



Quelques résultats

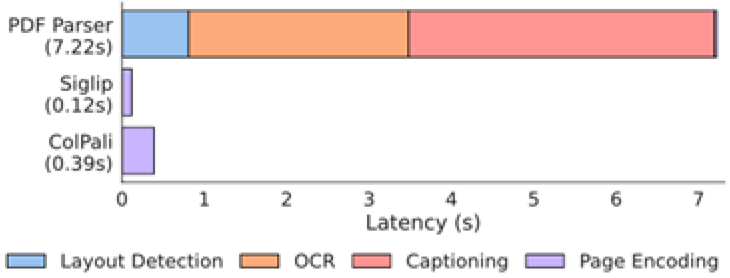


Figure 3: Offline indexing with *ColPali* is much simpler and faster compared to standard retrieval methods. Indexing speeds reported are computed on Nvidia L4 GPUs and detailed in subsection B.5.

ColPali est simple & rapide

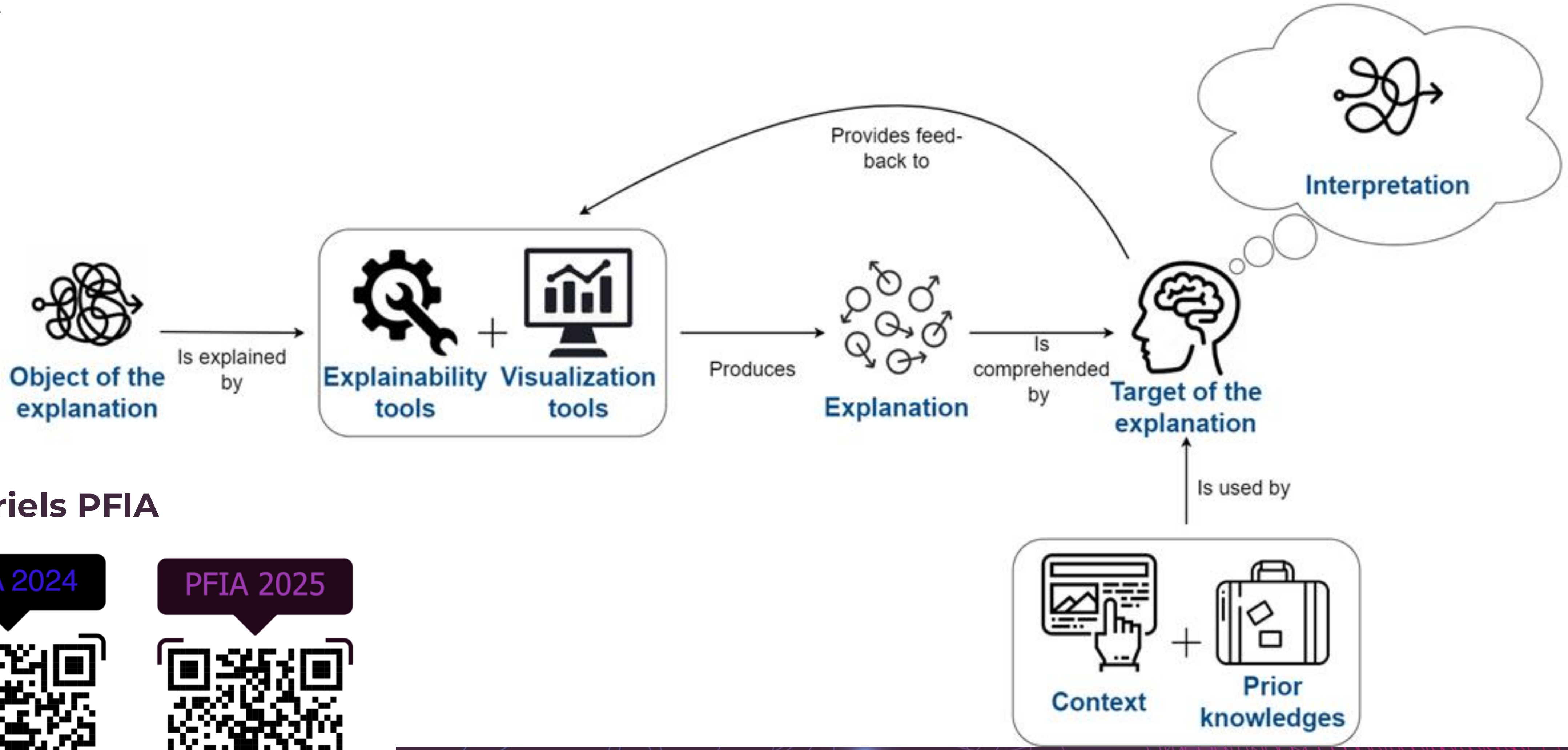
L'encodage s'effectue par simple transfert, ce qui permet d'intégrer les documents 18 fois plus rapidement qu'avec des pipelines PDF/OCR complexes.

	ArxivQ	DocQ	InfoQ	TabF	TATQ	Shift	AI	Energy	Gov.	Health.	Avg.
Unstructured Text only											
- BM25	-	34.1	-	-	44.0	59.6	90.4	78.3	78.8	82.6	-
- BGE-M3	-	28.4$_{\downarrow 5.7}$	-	-	36.1$_{\downarrow 7.9}$	68.5$_{\uparrow 8.9}$	88.4$_{\downarrow 2.0}$	76.8$_{\downarrow 1.5}$	77.7$_{\downarrow 1.1}$	84.6$_{\uparrow 2.0}$	-
Unstructured + OCR											
- BM25	31.6	36.8	62.9	46.5	62.7	64.3	92.8	85.9	83.9	87.2	65.5
- BGE-M3	31.4$_{\downarrow 0.2}$	25.7$_{\downarrow 11.1}$	60.1$_{\downarrow 2.8}$	70.8$_{\uparrow 24.3}$	50.5$_{\downarrow 12.2}$	**73.2**$_{\uparrow 8.9}$	90.2$_{\downarrow 2.6}$	83.6$_{\downarrow 2.3}$	84.9$_{\uparrow 1.0}$	91.1$_{\uparrow 3.9}$	66.1$_{\uparrow 0.6}$
Unstructured + Captioning											
- BM25	40.1	38.4	70.0	35.4	61.5	60.9	88.0	84.7	82.7	89.2	65.1
- BGE-M3	35.7$_{\downarrow 4.4}$	32.9$_{\downarrow 5.4}$	71.9$_{\uparrow 1.9}$	69.1$_{\uparrow 33.7}$	43.8$_{\downarrow 17.7}$	73.1$_{\uparrow 12.2}$	88.8$_{\uparrow 0.8}$	83.3$_{\downarrow 1.4}$	80.4$_{\downarrow 2.3}$	91.3$_{\uparrow 2.1}$	67.0$_{\uparrow 1.9}$
Contrastive VLMs											
Jina-CLIP	25.4	11.9	35.5	20.2	3.3	3.8	15.2	19.7	21.4	20.8	17.7
Nomic-vision	17.1	10.7	30.1	16.3	2.7	1.1	12.9	10.9	11.4	15.7	12.9
SigLIP (Vanilla)	43.2	30.3	64.1	58.1	26.2	18.7	62.5	65.7	66.1	79.1	51.4
Ours											
SigLIP (Vanilla)	43.2	30.3	64.1	58.1	26.2	18.7	62.5	65.7	66.1	79.1	51.4
BiSigLIP (+fine-tuning)	58.5$_{\uparrow 15.3}$	32.9$_{\uparrow 2.6}$	70.5$_{\uparrow 6.4}$	62.7$_{\uparrow 4.6}$	30.5$_{\uparrow 4.3}$	26.5$_{\uparrow 7.8}$	74.3$_{\uparrow 11.8}$	73.7$_{\uparrow 8.0}$	74.2$_{\uparrow 8.1}$	82.3$_{\uparrow 3.2}$	58.6$_{\uparrow 7.2}$
BiPali (+LLM)	56.5$_{\downarrow -2.0}$	30.0$_{\downarrow -2.9}$	67.4$_{\downarrow -3.1}$	76.9$_{\uparrow 14.2}$	33.4$_{\uparrow 2.9}$	43.7$_{\uparrow 17.2}$	71.2$_{\downarrow -3.1}$	61.9$_{\downarrow -11.7}$	73.8$_{\downarrow -0.4}$	73.6$_{\downarrow -8.8}$	58.8$_{\uparrow 0.2}$
ColPali (+Late Inter.)	**79.1**$_{\uparrow 22.6}$	**54.4**$_{\uparrow 24.5}$	**81.8**$_{\uparrow 14.4}$	**83.9**$_{\uparrow 7.0}$	**65.8**$_{\uparrow 32.4}$	**73.2**$_{\uparrow 29.5}$	**96.2**$_{\uparrow 25.0}$	**91.0**$_{\uparrow 29.1}$	**92.7**$_{\uparrow 18.9}$	**94.4**$_{\uparrow 20.8}$	**81.3**$_{\uparrow 22.5}$

Table 2: **Comprehensive evaluation of baseline models and our proposed method on *ViDoRe*.** Results are presented using NDCG@5 metrics, and illustrate the impact of different components. Text-only metrics are not computed for benchmarks with only visual elements.

3. Importance de l'explicabilité : vers des outils ouverts

Composantes clés de l'explicabilité



Tutoriels PFIA

PFIA 2024

PFIA 2025

Des outils en open-source

Interpreto

Interpretability Toolbox for LLMs

DEEL FOR ANITI

DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'ESPACE

Evolution of learned concepts during training

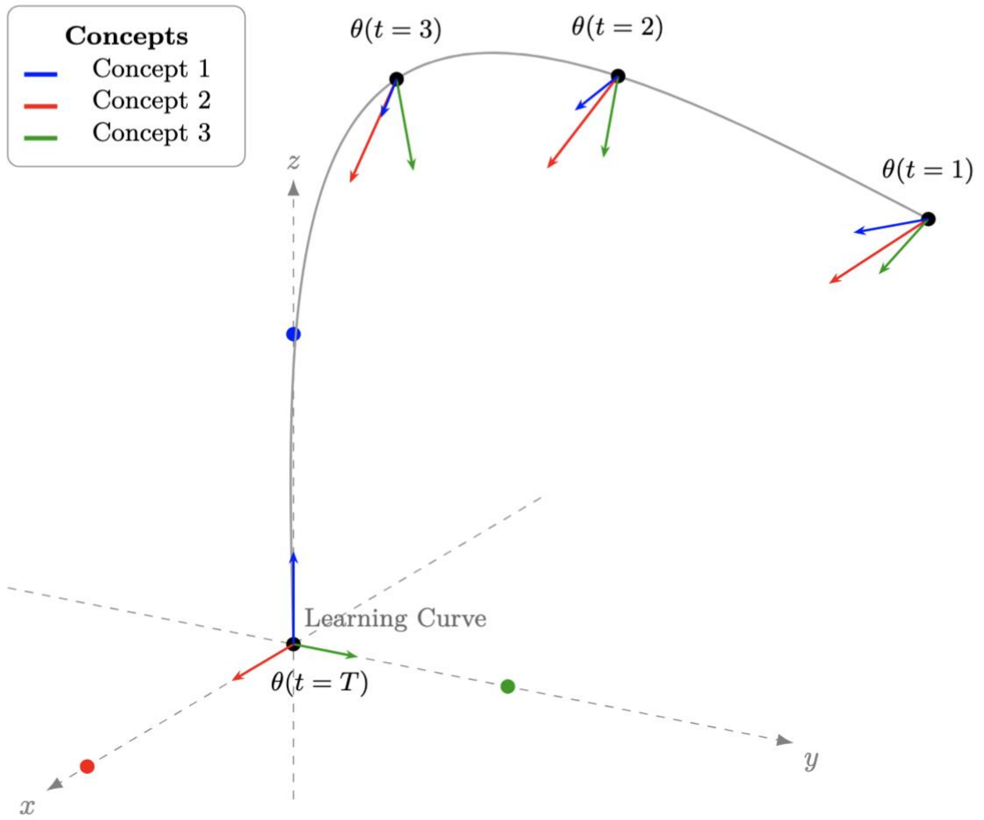


Figure 8: Evolution of θ over discrete time steps, with vectors indicating orientations toward three fixed concepts. The vector lengths represent the coefficients associated with each concept. The main intuition behind this figure is that, during training, the model may learn similar concepts but represent them differently (here, concept directions do not evolve linearly over time). In such a case, our process may fail to properly track concepts for direct comparison. However, the relative importance of concepts should remain stable-hence, we should focus on comparing their significance rather than their raw orientation (illustrated here by the arrow lengths).

Internship of Raphael Bernas with Fanny Jourdan and A. Poché, IRT Saint Exupery

Interpreto

Interpretability Toolbox for LLMs

DEEL FOR ANITI

Conclusions

- Il est possible de construire des **modèles européens ouverts, transparents et compétitifs**
- Importance de **l'ouverture des données, modèles, code de bases, points de contrôle d'entraînement.**
- Problème du **contexte-shift :** les jeux d'évaluation sont souvent peu représentatifs de l'ensemble des cas rééls, e.g. ViDOrE.
- Importance de supporter la mise à disposition d'outils pour la conception de systèmes de confiance.

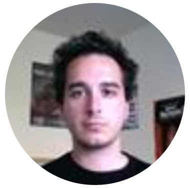
The team










Many more





MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'ESPACE
CentraleSupélec
TÉCNICO LISBOA
ARTEFACT RESEARCH CENTER
université PARIS-SACLAY
ILLUIN TECHNOLOGY
diabolocom
Unbabel
FRANCE 2030

Q&A