université PARIS-SACLAY

IA GÉNÉRATIVES: DES OUTILS POUR LES ÉTUDIANTS, LES ENSEIGNANTS ET LES CHERCHEURS

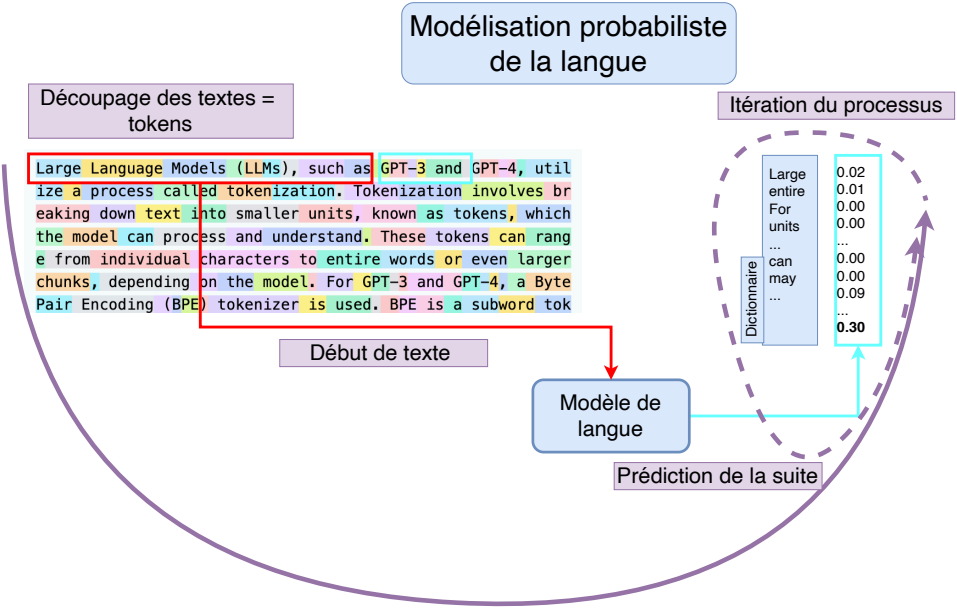
Lundi 24 mars 2025
IA générative à l'Université Paris-Saclay

Vincent Guigue
https://vguigue.github.io

MIA PARIS-SACLAY
EKINOCS

AGROPARISTECH
Institut des Sciences et Industries du Vivant et de l'Environnement

Les modèles de langue en 5 tableaux



Les modèles de langue en 5 tableaux

Pré-entrainement
Compléter des phrases

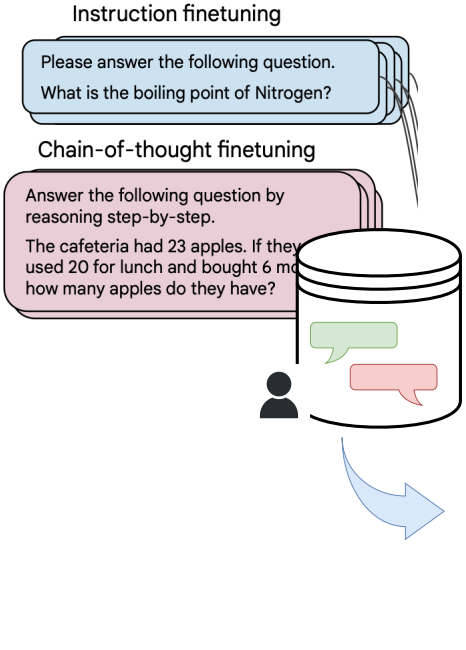
1963, he was was assassinated in Dallas ...

Modèle de langue

JFK died in

Acquisition de:
grammaire
entités
connaissances
raisonnement (?)

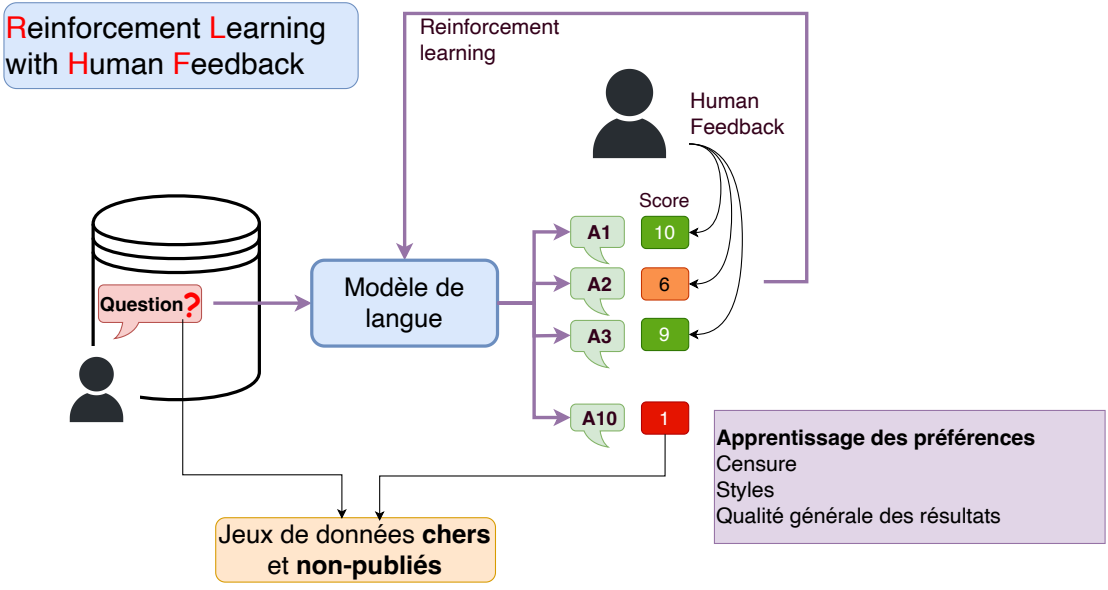
Les modèles de langue en 5 tableaux



Spécialisation sur des tâches

Acquisition de:
Capacité de répondre à une question
Suivre un dialogue
Connaissances physiques
Bases de raisonnement

Les modèles de langue en 5 tableaux



Chaine des compétences et de la souvenraineté

Construction du modèle de base

Maitrise des données
- Collecte/équilibrage
- Nettoyage

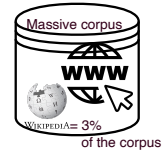
Entrainement
- Puissance machine (milliers de GPU)
- Architecture/recherche ML

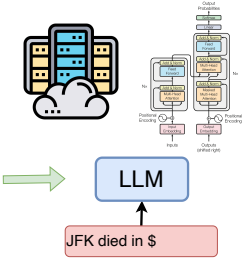
Raffinement du modèle

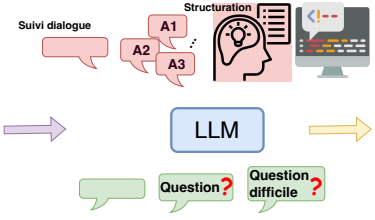
Maitrise & construction des données
- interactions humaines +++
- prix des données
- spécialisation à la demande

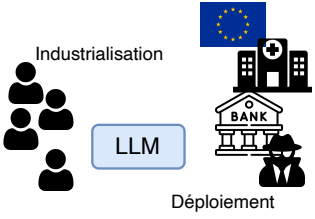
Exploitation du modèle

Optimisation / Limitation du coût
- compétence MLOps
- déploiement local

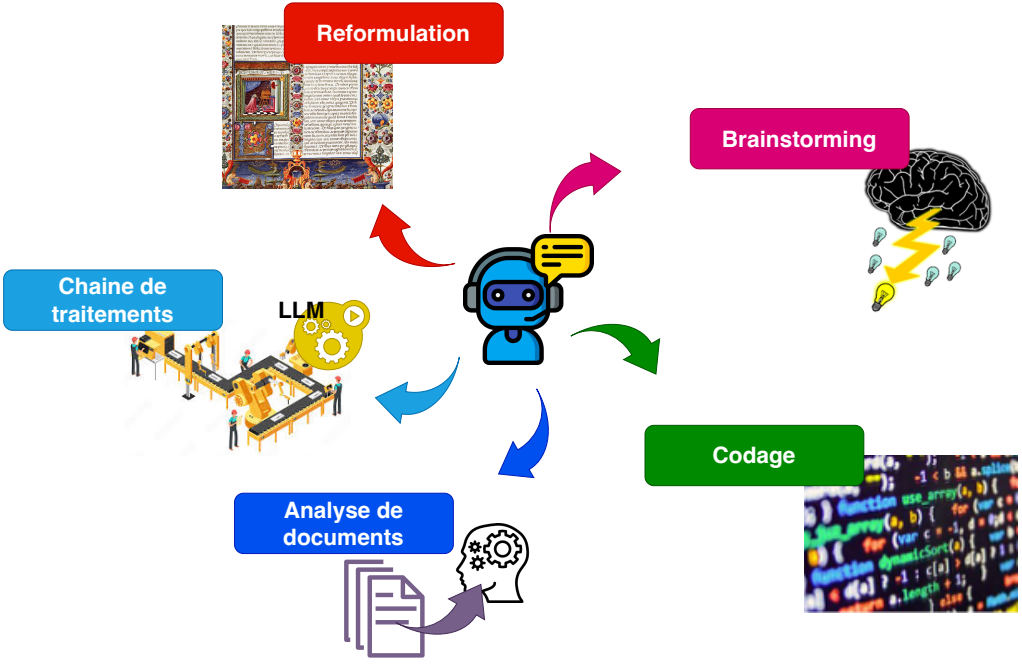








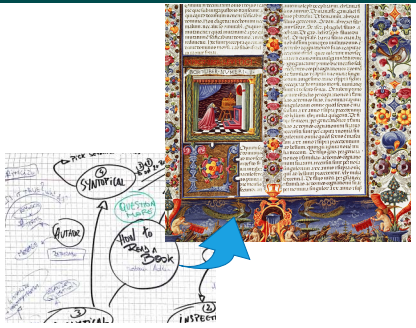
Les principaux usages en 5 tableaux



(1) Reformulation, mise en page, traduction

- Lettre de recommandation, motivation etc...
- Ecriture d'article
- Traduction

$\Rightarrow$ Gain en rapidité, gain en qualité (?)



- Faut-il indiquer l'usage des LLM dans les articles?
  - Quid de la détectabilité?
- Comment traiter les lettres de motivation dans les recrutements?
- Quels comportements sont éthiques ou pas? [Charte?]

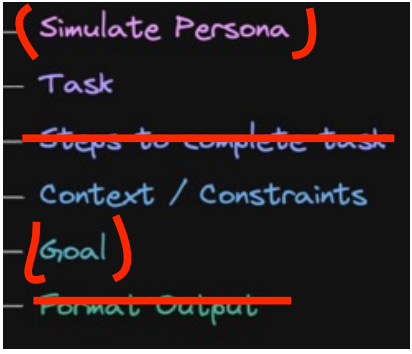
- Mauvaises tournures (et/ou phrases caractéristiques) [cf G. Cabanac]

In the era of... Overlearning...

(2) Brainstorming / plan de cours / révision de stats

- Chercher l'inspiration [synd. page blanche]
- Organiser ses idées rapidement
- Eviter les oublis
- Se documenter de manière ciblée, adaptée à ses besoins

⇒ Des réponses bluffantes parfois incomplètes ou partiellement fausses... Mais souvent utiles



3 articles de références sur l'usage des transformers en recommandation
A quoi sert la loi poisson log normale?
Proposer 10 parties pour un cours sur les Transformers en IA

- Dans quel périmètre les LLM sont-ils fiables?
- Quels risques pour les sources d'information primaires?
- Quels risques sociétaux sur l'information?

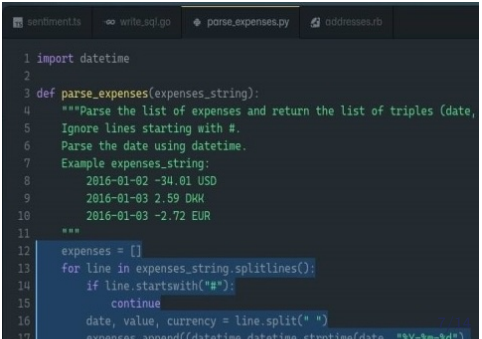
(3) Codage: différents outils, différents niveaux

- Donner la solution d'un exercice
- Apprendre à coder ou s'y remettre
  - Nouveaux langages, nouvelles approches (ML?)
  - Bénéfier des explications...

    Mais comment gérer les erreurs?
- Aide sur une lib [*getting started*]
- Codage + rapide

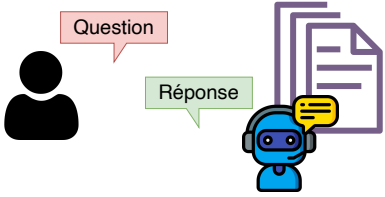
- Quid des copyrights?
  - Quel impact sur le traitement futur des codes?
- Comment adapter la pédagogies ?
- Combien d'appels pour la complétion?

  Quel bilan carbone?
- Quel risque de propagation d'erreurs?



(4) Analyse de documents

- Synthèse de documents / articles
- Dialogue avec une base documentaire
- Assistance à l'écriture de reviews
- FAQ, service d'assitance interne dans les entreprises
- Veille
- Génération de quizz à partir d'un poly de cours

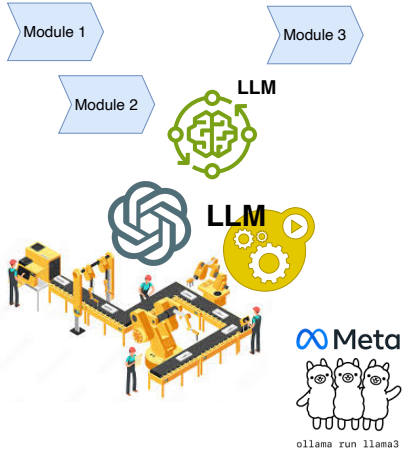


- Les articles seront-ils encore lus demain?
  - Faut-il rendre nos articles NotebookLM-proof?
- Comment gagner du temps mais rester honnête & éthique?

(5) LLM dans une chaine de production / Agentic AI

- Extraire des connaissances
- Trier des documents / produire résumés
- Générer des exemples pour entrainer un modèle
  [Teacher/student - distillation]
- Variantes d'exemples ↗↗ taille données
  [Data augmentation]

⇒ Intégrer le LLM dans une chaine de traitement
= peu/moins de supervision = **Agentic AI**

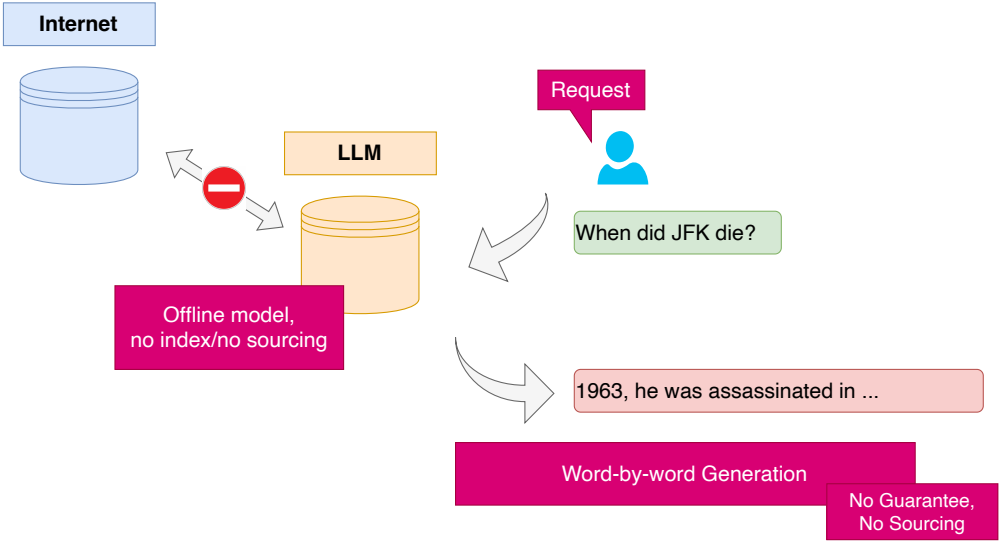


⚠ format de sortie ⇒ regex / contrainte de formattage JSON

- Est ce que je peux entrainer des modèles sur des données générées?
- Combien ça coute? ($ + CO_2) Besoin de GPU?
- Que valent les modèles open-weight?

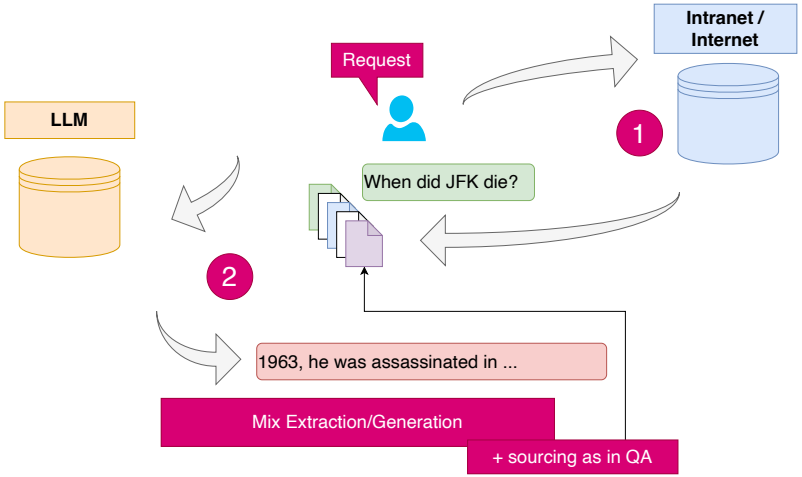
Mémoire paramétrique vs extraction d'information

Besoins spécifiques: métriques, enjeux...



Mémoire paramétrique vs extraction d'information

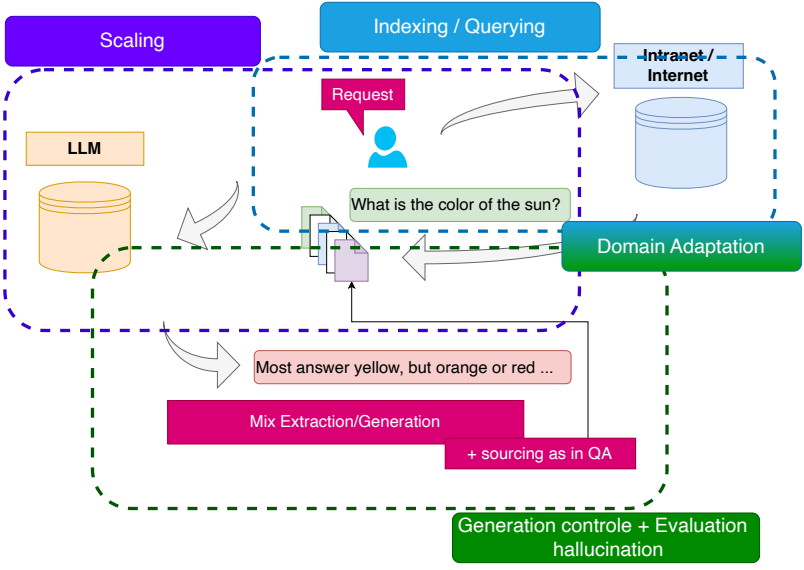
Besoins spécifiques: métriques, enjeux... [$\neq$ hallucinations !!]



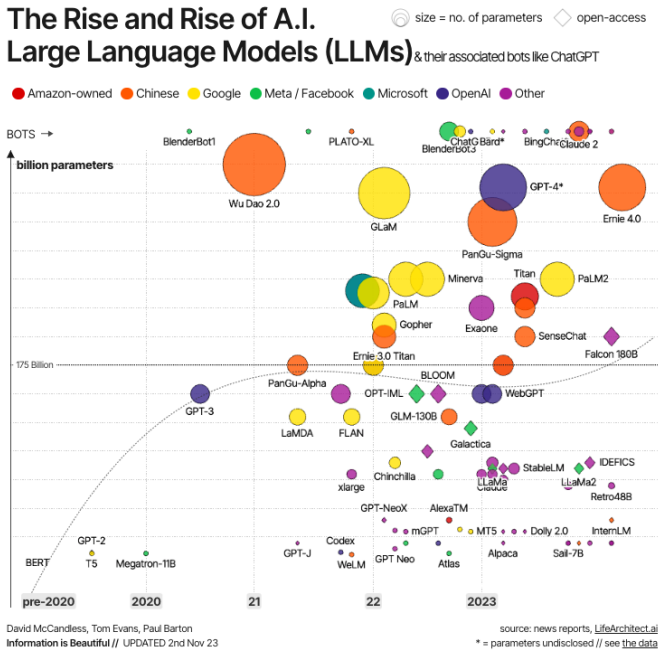
NotebookLM : penser qu'un ensemble de documents (longs) peuvent constituer une source pour le dialogue

Mémoire paramétrique vs extraction d'information

Besoins spécifiques: métriques, enjeux... [≠ hallucinations !!]



Les LLMs & la frugalité



Paramètres

1998 LeNet-5 = 0.06M

2011 Senna = 7.3M

2012 AlexNet = 60M

2017 Transformer = 65M / 210M

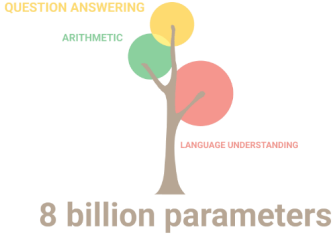
2018 ELMo = 94M

2018 BERT = 110M / 340M

2019 GPT2 = 1,500M

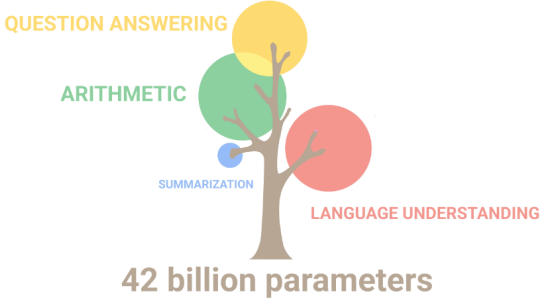
2020 GPT3 = 175,000M

Capacités émergentes

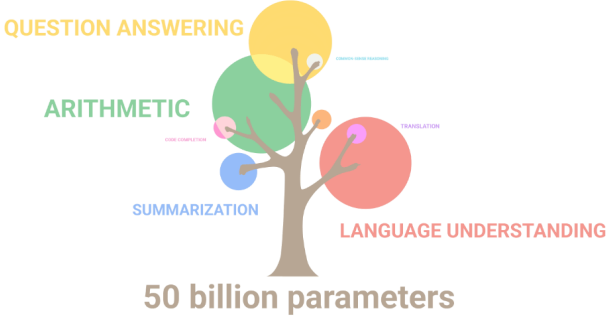


Les LLMs & la frugalité

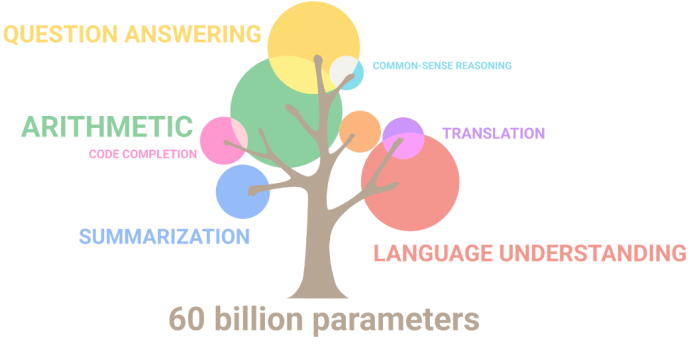
Capacités émergentes



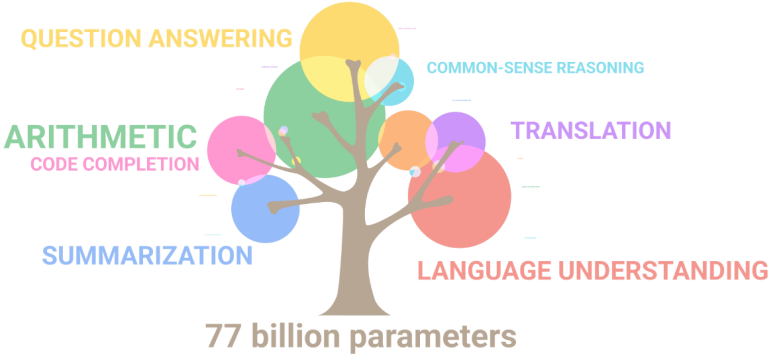
Capacités émergentes



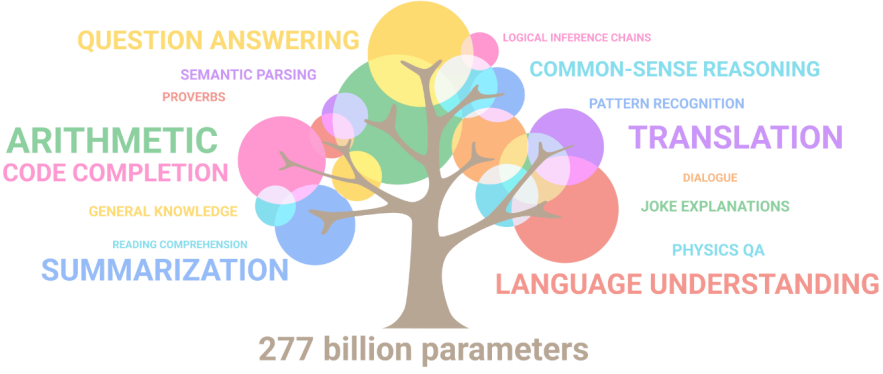
Capacités émergentes



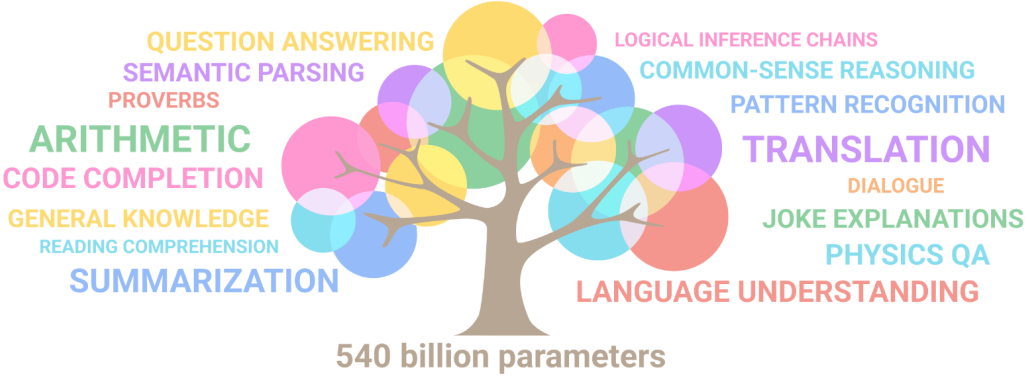
Capacités émergentes



Capacités émergentes

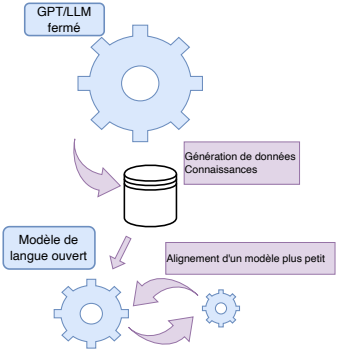


Capacités émergentes



Les LLMs & la frugalité

Distillation



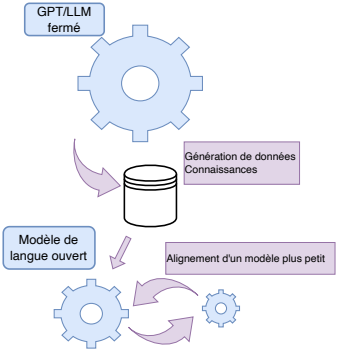
Elagage
Quantification

Mixture of experts

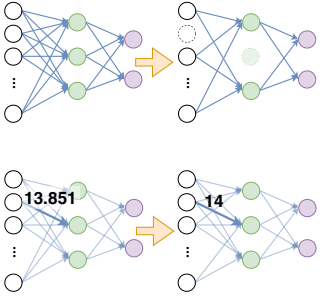
La frugalité... Taille modèle **x1000 en 3 ans**... Puis **opti. x1/100 en 2ans**

Les LLMs & la frugalité

Distillation



Elagage
Quantification



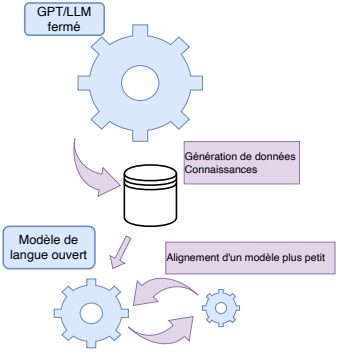
FP32 $\Rightarrow$ INT4

Mixture of experts

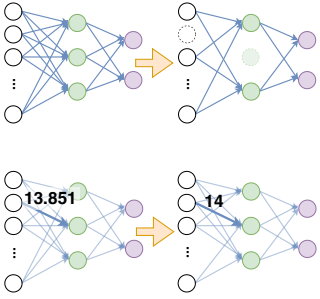
La frugalité... Taille modèle **x1000 en 3 ans**... Puis **opti. x1/100 en 2ans**

Les LLMs & la frugalité

Distillation

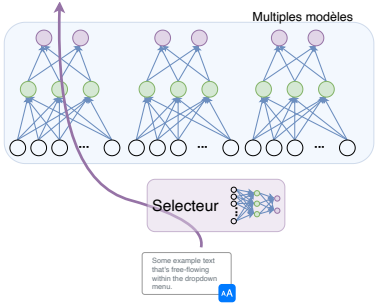


Elagage Quantification



FP32 $\Rightarrow$ INT4

Mixture of experts

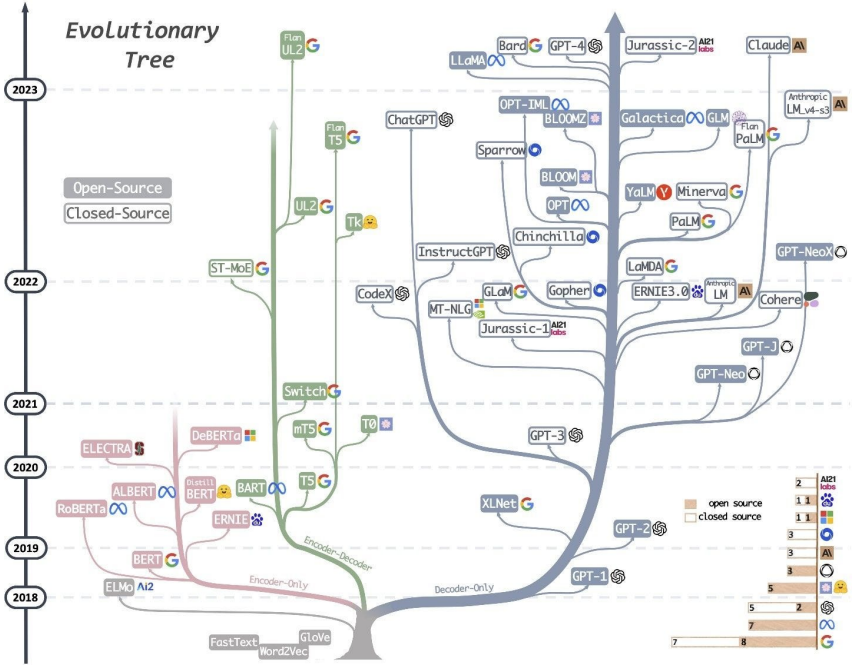


+ Industriualisation du code

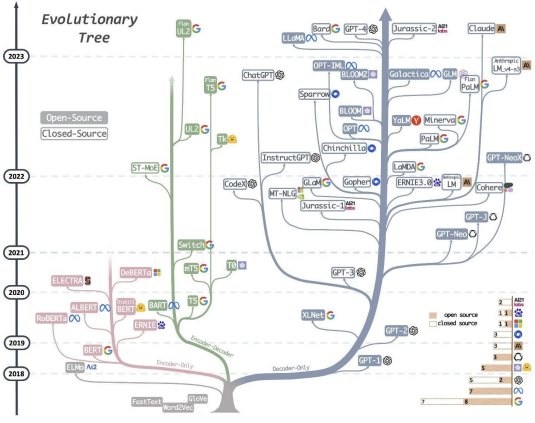
Mais le plus important: adéquation aux usages!

La frugalité... Taille modèle **x1000 en 3 ans**... Puis **opti. x1/100 en 2ans**

Les LLMs, la transparence & la privacy



Les LLMs, la transparence & la privacy



Transparence

- Est ce que je peux le modifier?
  Adaptation
- Quelles données d'entrainement?
  Data contamination / skills
- Quelle ligne éditoriale / censure?
  Accès à l'information

Privacy

- Fonctionnement local
  A quel coût?
- Quelles garanties sur les serveurs?
  Contrats $\Leftrightarrow$ US Patriot Act

Les LLMs, la transparence & la privacy

Des solutions disponibles:

- Clé en main: Ollama 
  - Nombreux modèles disponibles + MAJ
  - Accès API / console
- Sur mesure : HuggingFace 
  - Pytorch/TensorFlow
  - Accès modèles + fine-tuning
  - Quantif / optim

 Open-Weight $\neq$ Open-Source

Transparence

- Est ce que je peux le modifier?
  Adaptation
- Quelles données d'entrainement?
  Data contamination / skills
- Quelle ligne éditoriale / censure?
  Accès à l'information

Privacy

- Fonctionnement local
  A quel coût?
- Quelles garanties sur les serveurs?
  Contrats $\Leftrightarrow$ US Patriot Act

Les LLMs, la transparence & la privacy

Des solutions disponibles:

- Clé en main: Ollama
  - Nombreux modèles disponibles + MA
  - Acc
- Sur mes
  - Pyt
  - Acc
  - Qua

DeepSeek

- Toujours des Transformers / 600 Milliards de paramètres
- Mixture of experts (comme MixTral)
- 2000 GPU NVidia + 5M $/training (=chatGPT)
- ... Un critère d'apprentissage original
  + performant pour le raisonnement

Open-Weight $\neq$ Open-Source

Transparence

- Est ce que je peux le modifier?
  Adaptation
- Quelles données d'entrainement?
  Data contamination / skills
  re?
  nation
  coût?
  urs?
  Contrats $\Leftrightarrow$ US Patriot Act

Les enjeux à venir

- **Quid des hallucinations?**
  - Faut-il les réduire ou vivre avec?
  - Les LLM vont-ils progresser? Dans quelles directions?
  - Le LLM nous fait-il *perdre* le rapport à la vérité? à la vérification?
- **Faut-il des petits ou des grands modèles de langues?**
  - Combien ça coute? Est-ce soutenable?
  - Avec ou sans fine-tuning?
  - Qu'est ce que la frugalité dans l'univers des LLM?
- **Quand les autres s'en servent... Quel impact sur moi?**
  - Productivité (collègues chercheurs, codeurs, relecteurs, ...)
  - Pédagogie : gérer/former des étudiants *branchés*
- **Protection des données... Les miennes et celles des autres**
  - Est-il raisonnable d'entrainer les LLM sur github, wikipedia, les articles scientifiques, les journaux, ... ?
  - Quelle est l'importance de la privacy? Quels risques lorsque j'utilise un LLM?

Les enjeux à venir

- **Quid des hallucinations?**
  - Faut-il les réduire ou vivre avec?
  - Les LLM vont-ils progresser? Dans quelles directions?
  - Le LLM nous fait-il *perdre* le rapport à la vérité? à la vérification?
- **Faut-il des petits ou des grands modèles de langues?**
  - Combien ça coute? Est-ce soutenable?
  - A
  - Q
- **Quan**
  - Productivité (collègues chercheurs, codeurs, relecteurs, ...)
  - Pédagogie : gérer/former des étudiants *branchés*

Le smartphone a fait de moi un *humain-augmenté*...
Est ce que le LLM va faire de moi un *chercheur-augmenté*?
$\Rightarrow$ Jetez (quand même) un oeil à NotebookLM

- **Protection des données... Les miennes et celles des autres**
  - Est-il raisonnable d'entrainer les LLM sur github, wikipedia, les articles scientifiques, les journaux, ... ?
  - Quelle est l'importance de la privacy? Quels risques lorsque j'utilise un LLM?