

# Utilisation des IA génératives (LLM) en entreprise

Formation continue – 1 jour

# Introduction – Vue d'ensemble des modèles d'IA

## 1 – Introduction à la notion de modèle d'IA

1.1 – Qu'est-ce qu'un modèle ?

1.2 – Apprentissage Machine (*Machine Learning*), entraînement

## 2 – Explication du fonctionnement des LLM

2.1 – Modèle de fondation – génération de texte, *token*, *embeddings*

2.2 - Entraînement

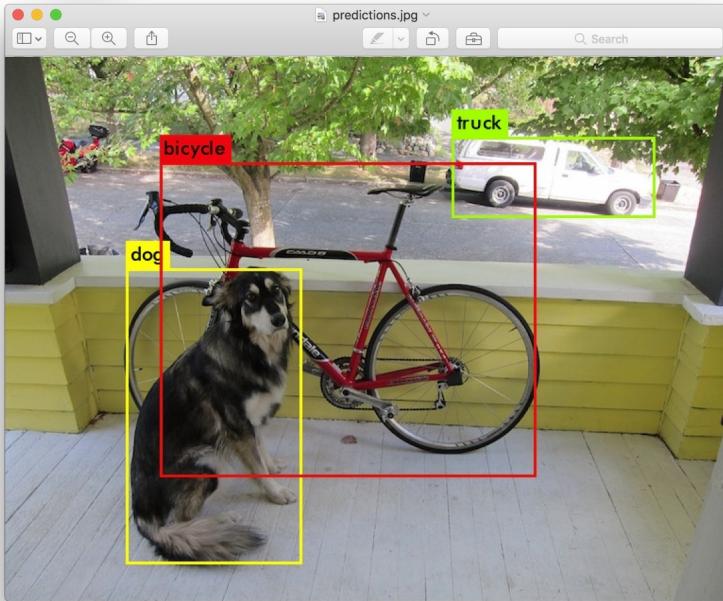
2.3 – Du modèle de fondation à l'agent conversationnel

## 3 – Écosystème et historique des LLM

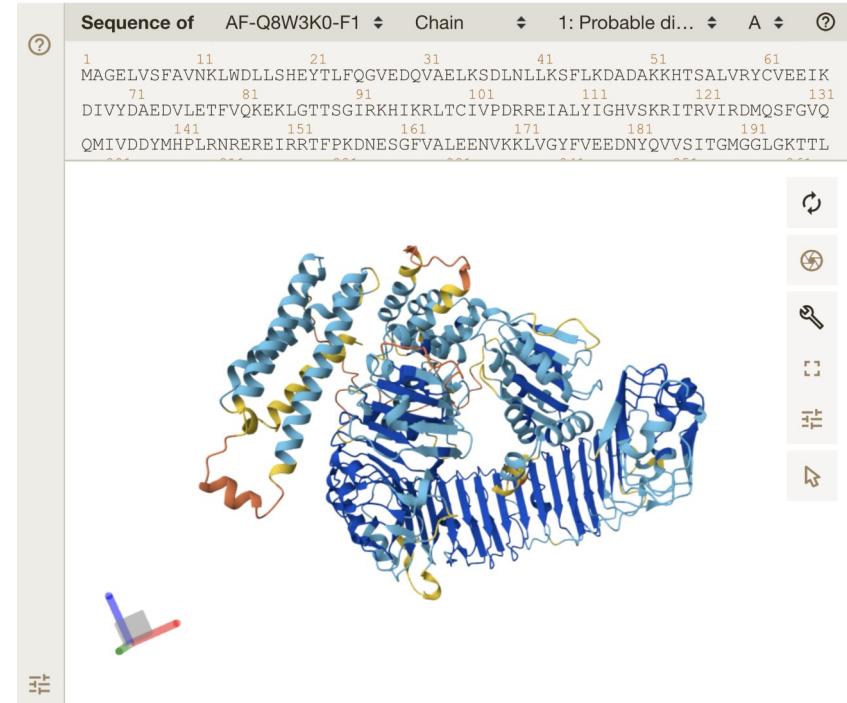
3.1 – Modèles propriétaires, ouverts, open-source

3.2 – Historique et chronologie de l'engouement

# 1 – Introduction à la notion de modèle d'IA



You Only Look Once (YOLO)  
<https://pjreddie.com/darknet/yolo/>



AlphaFold  
<https://alphafold.com/entry/Q8W3K0>

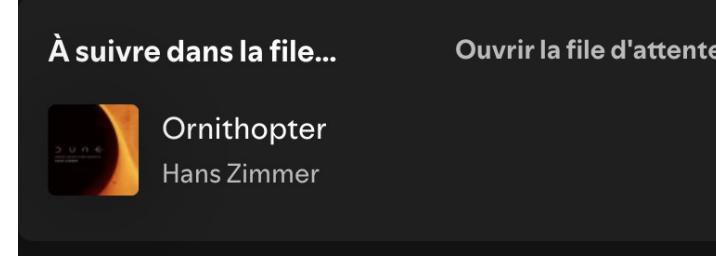
# 1 – Introduction à la notion de modèle d'IA



Prompt: scene of a giant ancient tortoise with a fantasy city built on its back. The tortoise's shell is covered in lush, dense forest with towering trees and a hidden, misty village nestled in the foliage. The city consists of intricately designed buildings that blend seamlessly with the natural environment, featuring rope bridges connecting different sections of the city.

Stable Diffusion

<https://stability.ai/news/introducing-stable-diffusion-3-5>



Algorithme de recommandation  
Spotify, YouTube,  
TikTok, ...

Itinéraire

La Poterie Rue de Vern, 215, Rennes, Fr...  
Université de Rennes - Campus Beaulieu

Partir maintenant

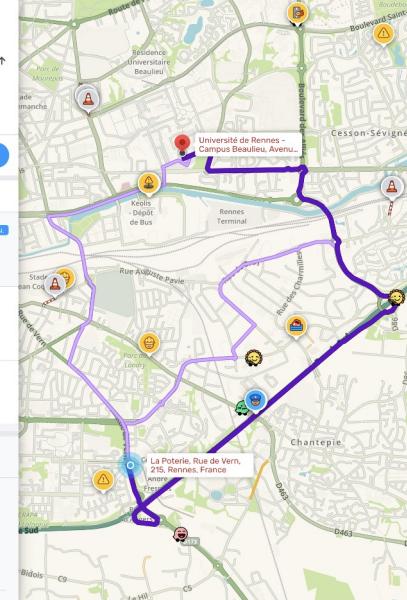
Enregistrer dans l'application

Itinéraires

- 11 min Arrivée à 2:41 PM Rocade Sud; Boulevard des Alliés Cesson-Sévigné 7.9 KM
- 14 min Arrivée à 2:44 PM Rue de Vern, Boulevard Villebois-Mareuil, Avenue François Château Rennes 5.2 KM
- 15 min Arrivée à 2:44 PM Rue de Bray Cesson-Sévigné 7.1 KM

Point de départ      Destination

Université de Rennes - Campus Beaulieu  
Avenue du Général Leclerc, 263, Rennes, Ille-et-Vilaine, France



Prédiction de temps de trajet - Waze

# 1 – Introduction à la notion de modèle d'IA



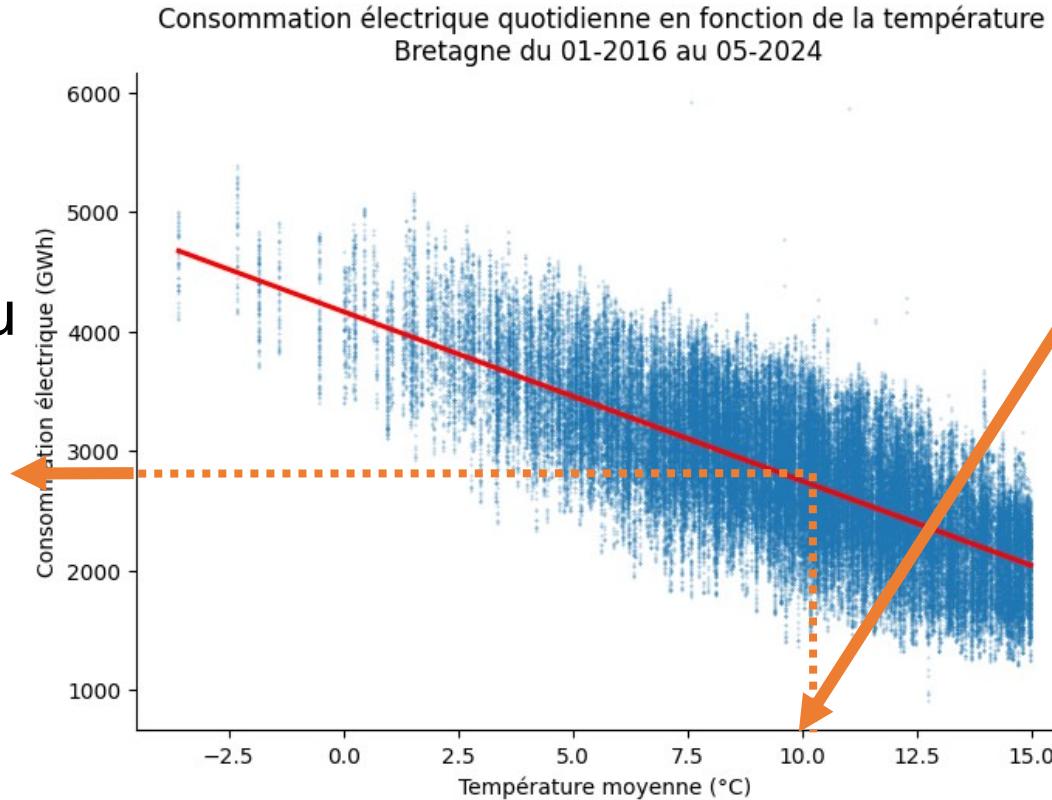
Stable Diffusion

<https://arxiv.org/pdf/2403.03206>

A whimsical and creative image depicting a hybrid creature that is a mix of a waffle and a hippopotamus. This imaginative creature features the distinctive, bulky body of a hippo, but with a texture and appearance resembling a golden-brown, crispy waffle. The creature might have elements like waffle squares across its skin and a syrup-like sheen. It's set in a surreal environment that playfully combines the natural water habitat of a hippo with elements of a breakfast table setting, possibly including oversized utensils or plates in the background. The image should evoke a sense of playful absurdity and culinary fantasy.

# 1.1 – Qu'est-ce qu'un modèle ?

Réponse du  
modèle :  
2800 GWh



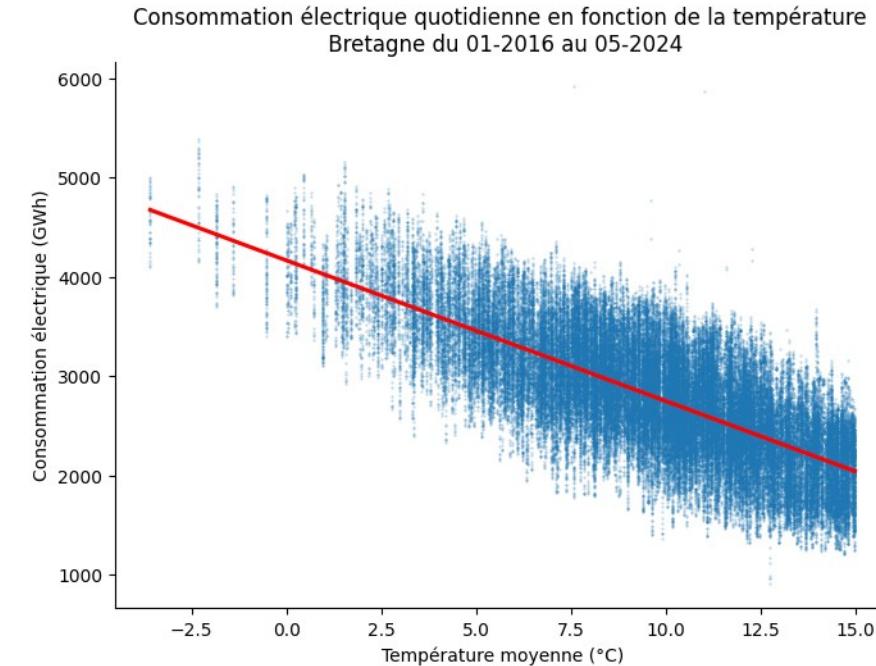
Demain, il fera en moyenne  $10^{\circ}\text{C}$ , quelle sera la consommation électrique ?

# 1.1 – Qu'est-ce qu'un modèle ?

## Qu'est-ce qu'un modèle ?

- on observe un phénomène
- on a besoin de le **reproduire**, de **mieux le comprendre**, de **l'automatiser** (*via* un programme informatique par exemple)

-> Un modèle apporte un support théorique pour résoudre ces problèmes



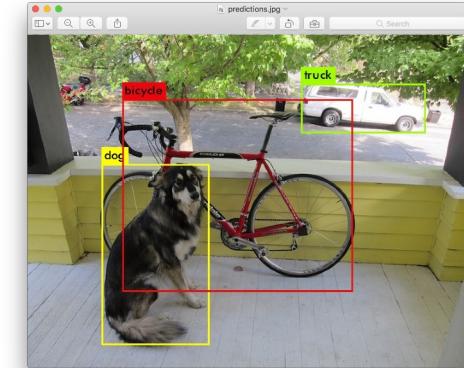
# 1.1 – Qu'est-ce qu'un modèle ?

Modèle  
« Entrée » → « Sortie »

10°C



2800 GWh



A whimsical and creative image depicting a hybrid creature that is a mix of a waffle and a hippopotamus...



A whimsical and creative image depicting a hybrid creature that is a mix of a waffle and a hippopotamus. This image is often used to demonstrate how neural networks can identify objects in a scene. It's a creative interpretation that playfully combines a real-world object (the hippo) with a food item (the waffle) to create a new entity.



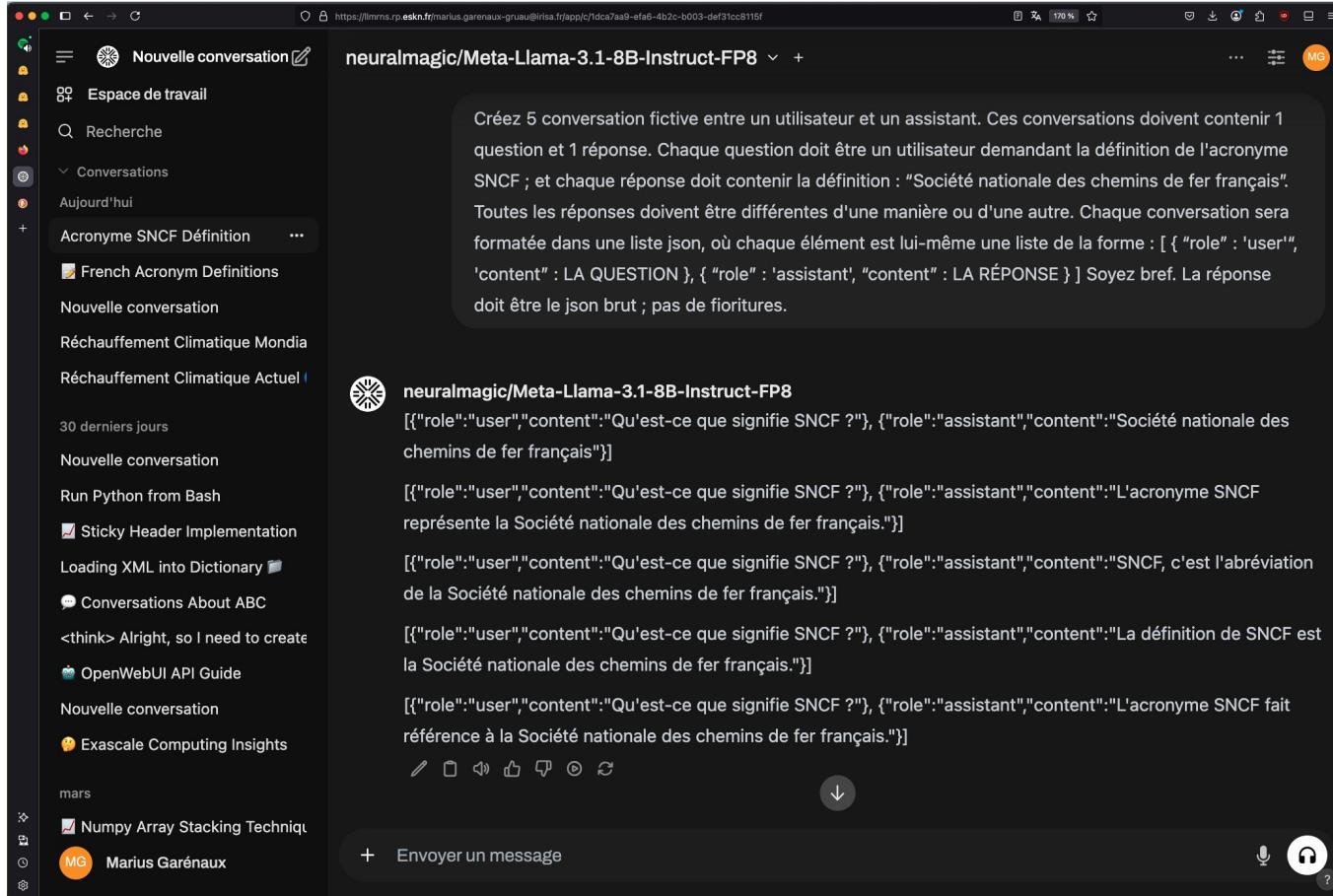
## 1.1 – Qu'est-ce qu'un modèle ?

Et les LLM ?

LLM = Large Language Model (Grand Modèle de Langage)

Ce sont des modèles qui **compriment le langage naturel**, qui sont capables de **produire et d'analyser** du **texte**.

# 1.1 – Qu'est-ce qu'un modèle ?



The screenshot shows a web-based interface for a large language model, specifically neuralmagic/Meta-Llama-3.1-8B-Instruct-FP8. The left sidebar contains a list of recent conversations and topics, such as "Acronyme SNCF Définition", "French Acronym Definitions", and "Réchauffement Climatique Mondia". The main area displays a conversation prompt and several generated responses.

**Prompt:**

Créez 5 conversation fictive entre un utilisateur et un assistant. Ces conversations doivent contenir 1 question et 1 réponse. Chaque question doit être un utilisateur demandant la définition de l'acronyme SNCF ; et chaque réponse doit contenir la définition : "Société nationale des chemins de fer français". Toutes les réponses doivent être différentes d'une manière ou d'une autre. Chaque conversation sera formatée dans une liste json, où chaque élément est lui-même une liste de la forme : [ { "role" : 'user', "content" : LA QUESTION }, { "role" : 'assistant', "content" : LA RÉPONSE } ] Soyez bref. La réponse doit être le json brut ; pas de fioritures.

**Responses (JSON format):**

- [{"role": "user", "content": "Qu'est-ce que signifie SNCF ?"}, {"role": "assistant", "content": "Société nationale des chemins de fer français"}]
- [{"role": "user", "content": "Qu'est-ce que signifie SNCF ?"}, {"role": "assistant", "content": "L'acronyme SNCF représente la Société nationale des chemins de fer français."}]
- [{"role": "user", "content": "Qu'est-ce que signifie SNCF ?"}, {"role": "assistant", "content": "SNCF, c'est l'abréviation de la Société nationale des chemins de fer français."}]
- [{"role": "user", "content": "Qu'est-ce que signifie SNCF ?"}, {"role": "assistant", "content": "La définition de SNCF est la Société nationale des chemins de fer français."}]
- [{"role": "user", "content": "Qu'est-ce que signifie SNCF ?"}, {"role": "assistant", "content": "L'acronyme SNCF fait référence à la Société nationale des chemins de fer français."}]

**Message Input:**

+ Envoyer un message

ChatGPT

Se connecter S'inscrire

Créez 5 conversations fictives entre un utilisateur et un assistant. Ces conversations doivent contenir 1 question et 1 réponse. Chaque question doit être une demande de définition de l'acronyme SNCF ; et chaque réponse doit contenir la définition : "Société nationale des chemins de fer français". Toutes les réponses doivent être différentes d'une manière ou d'une autre. Chaque conversation sera formatée dans une liste JSON, où chaque élément est lui-même une liste de la forme : [ { "role" : "user", "content" : LA QUESTION }, { "role" : "assistant", "content" : LA RÉPONSE } ] Soyez brefs. La réponse doit être le JSON brut ; pas de fioritures.

json

Copier Modifier

```
[  
  [  
    { "role": "user", "content": "C'est quoi, la SNCF ?"},  
    { "role": "assistant", "content": "SNCF signifie Société nationale des chemins de fer"},  
  ],  
  [  
    { "role": "user", "content": "Peux-tu me dire ce que veut dire l'acronyme SNCF ?"},  
    { "role": "assistant", "content": "Bie\u2022 ur, SNCF veut dire Société nationale des che"}]
```

Poser une question

Joindre Rechercher Raisonnez

ChatGPT peut faire des erreurs. Envisagez de vérifier les informations importantes.



## 1.1 – Qu'est-ce qu'un modèle ?

Comment modéliser quelque chose d'aussi complexe que le langage ?

« La **souris** est un **rongeur**. »

« La **souris** de mon **ordinateur** est cassée. »

« Hier soir, on a mangé de la **souris d'agneau**. »

-> polysémie, interprétation, contexte, ...

## 1.1 – Qu'est-ce qu'un modèle ?

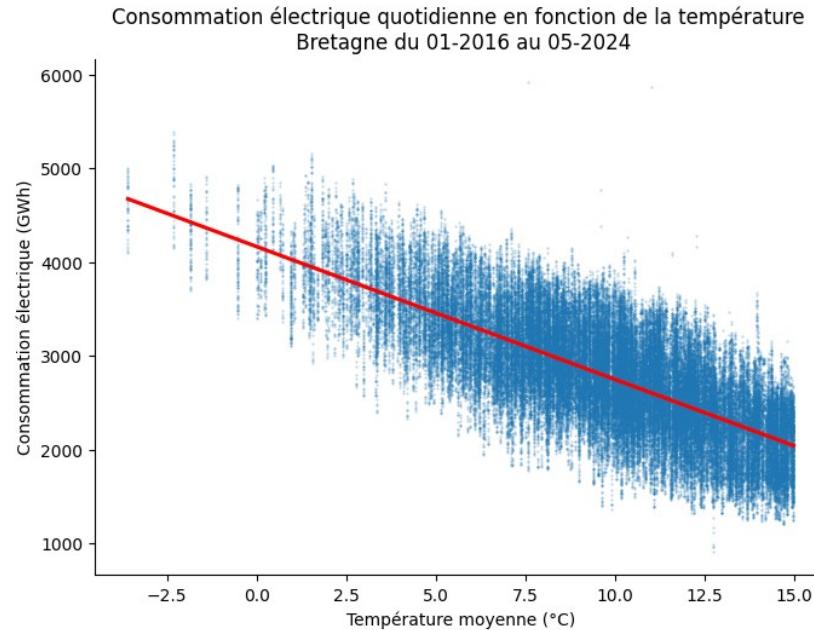
Comment modéliser quelque chose d'aussi complexe que le langage ?

On utilise énormément de données, et on construit des modèles qui « apprennent » avec ces données.

## 1.2 – Apprentissage Machine, entraînement

**Entraîner** = Trouver la **meilleure** hauteur et la meilleure pente pour notre droite

**Meilleure** ? La plus proche du nuage de point



## 1.2 – Apprentissage Machine, entraînement

### Notion de paramètre / poids :

Ce sont des **nombres** ; qui permettent d'ajuster un modèle.

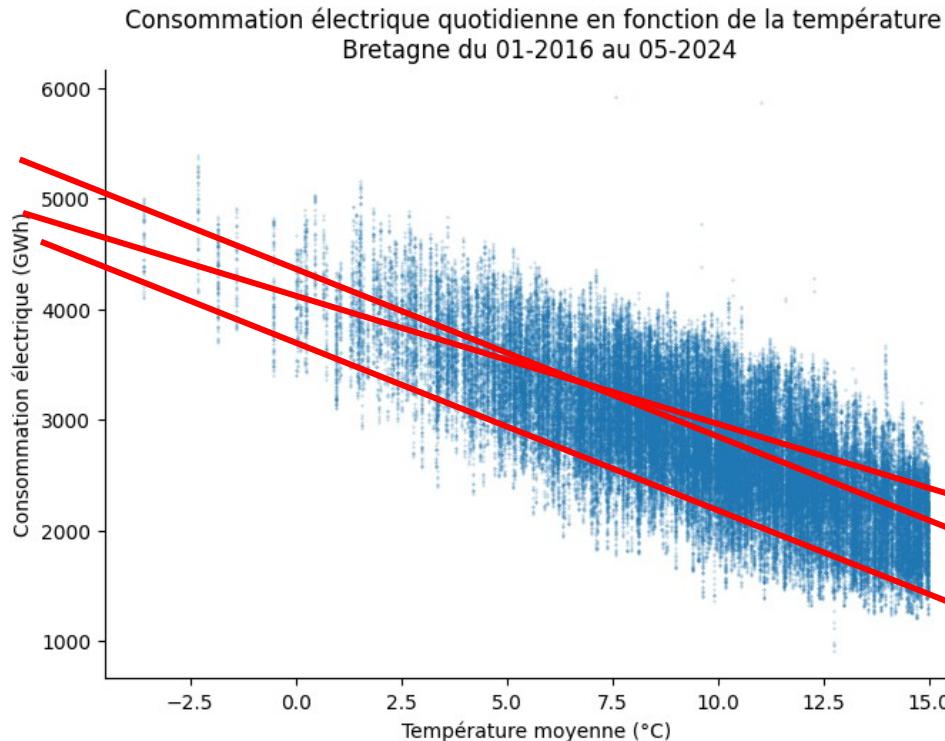
Par exemple, on peut ajuster la droite en jouant sur sa hauteur et sa pente (=2 paramètres).

Pour la droite on peut écrire le lien entre l'entrée et la sortie avec la formule :

**sortie = entrée \* pente + hauteur (ou  $y = a*x + b$ )**

# 1.2 – Apprentissage Machine, entraînement

Faire apprendre un programme informatique :



Pente : 160  
Hauteur : 4700

Pente : 200  
Hauteur : 5000

Pente : 200  
Hauteur : 4200

## 1.2 – Apprentissage Machine, entraînement

Droite = 2 paramètres (hauteur et pente)



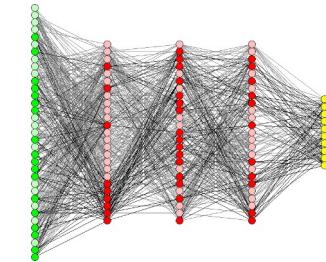
Plus de puissance de calcul

Dillon Marsh - Palabora Mine - 4.1 million tonnes of copper

LLM  $\sim 10\ 000\ 000\ 000$  paramètres

(plusieurs dizaines / centaines de milliards)

Des modèles avancés



Beaucoup de données



# 2 – Explication du fonctionnement des LLM

## 1 – Introduction à la notion de modèle d'IA

1.1 – Qu'est-ce qu'un modèle ?

1.2 – Apprentissage Machine (*Machine Learning*), entraînement

## 2 – Explication du fonctionnement des LLM

2.1 – Modèle de fondation – génération de texte, *token*, *embeddings*

2.2 - Entraînement

2.3 – Du modèle de fondation à l'agent conversationnel

## 3 – Écosystème et historique des LLM

3.1 – Modèles propriétaires, ouverts, open-source

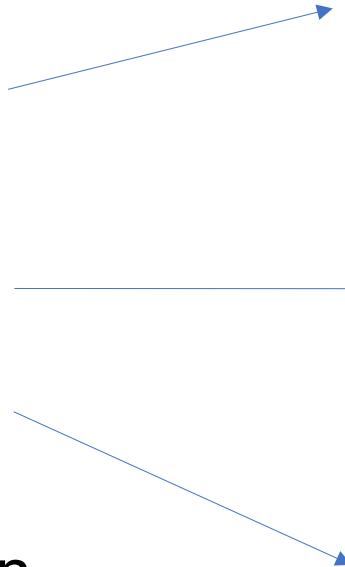
3.2 – Historique et chronologie de l'engouement

## 2.1 – Modèle de fondation

### Spécialisations



Modèle de fondation



## 2.1 – Modèle de fondation



- cœur des LLM
- comprend le langage naturel
- créé une représentation numérique des textes (*tokens, embeddings*)**
- peut être spécialisé pour diverses tâches

### Modèle de fondation

## 2.1 – Modèle de fondation

Un modèle de fondation peut être spécialisé pour différentes tâches :

- génération de texte (agent conversationnel, ...)
- résumé de texte,
- remplir des textes à trous,
- classification de texte (dire si une phrase est positive ou négative)
- ....

<https://huggingface.co/models>



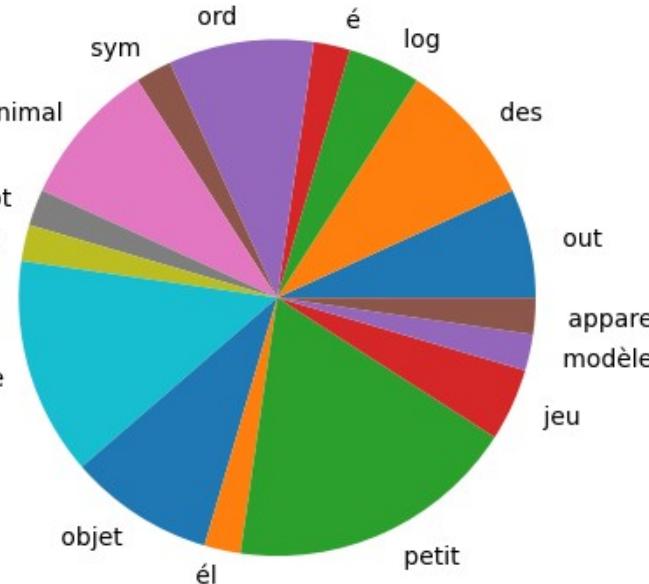
## 2.1 – Modèle de fondation – génération de texte

### Modèle de fondation

**Entrée :**

« La souris de l'ordinateur est un »

→  
programme  
robot instrument



**Sortie :**

« programme »

**Souvent, on appelle cette étape  
« l'inférence »**

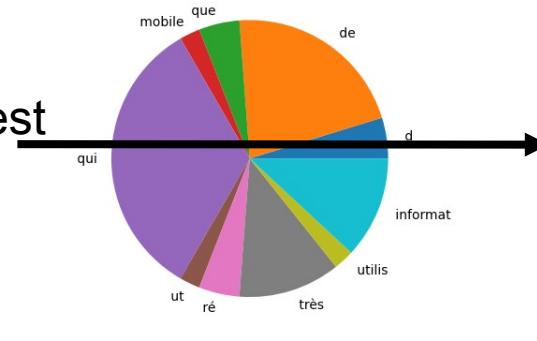
## 2.1 – Modèle de fondation – génération de texte

Aparté :

- Certains modèles ne complètent pas forcément le dernier token, mais plutôt un token au milieu de la phrase (ex : BERT)
- Il existe différentes méthodes pour tirer au hasard le prochain token à partir de la distribution de probabilité (voir document joint à cette présentation:  
<https://logistica.fenix.rudi-univ-rennes1.fr/media/download/22d8b25b-6ef6-475b-a5e6-a6dd41c62b4f> )

## 2.1 – Modèle de fondation - génération de texte

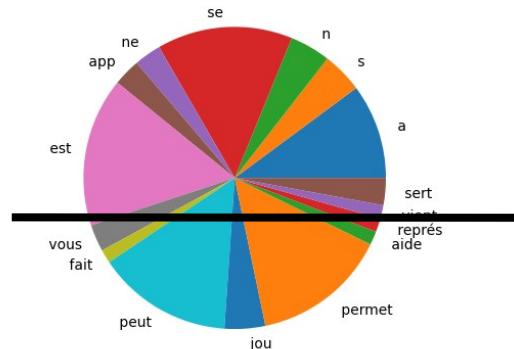
« La souris de l'ordinateur est un objet »



### Modèle de fondation

« qui »

« La souris de l'ordinateur est un objet qui »



« permet »

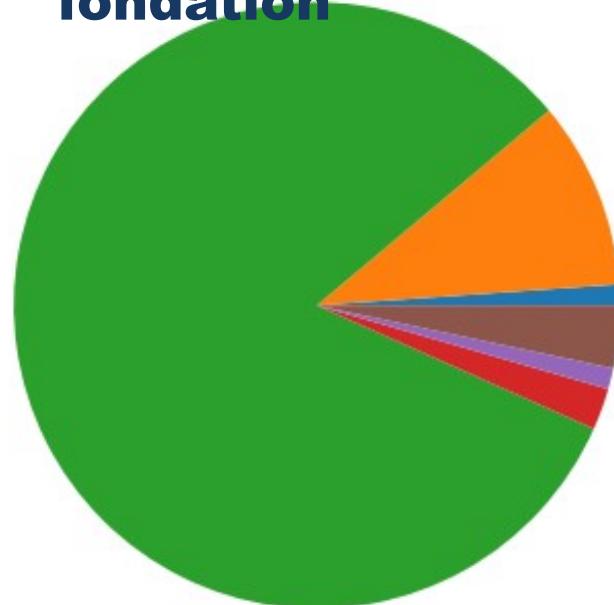
et on recommence....

## 2.1 – Modèle de fondation - génération de texte

### Modèle de fondation

#### Entrée :

« La souris de l'ordinateur est un objet qui permet de manipuler les informations sur un »



é

dis  
système  
terminal

PC



#### Sortie :



« La souris de l'ordinateur est un objet qui permet de manipuler les informations sur un ordinateur »

## 2.1 – Modèle de fondation – *tokens*

### Les « tokens »

La souris de l'ordinateur est un objet qui permet de manipuler les informations sur un ordinateur.

Encoding: o200k\_base (GPT-4o) ▾

```
je mange,  
tu manges,  
il mange,  
nous mangeons,  
vous mangez,  
elles mangent.
```

Clear

je mange,<newline>tu manges,<newline>il mange,<newline>nous mangeons,<newline>vous mangez,<newline>elles mangent.

## 2.1 – Modèle de fondation – *tokens*

Chaque *token* a un *id*, qui est simplement son rang dans le vocabulaire (un nombre entier de 0 à la taille du vocabulaire -1).

token_id	token
0	!
1	"
2	#
3	\$
4	%
...	...
128251	< reserved_special_token_243 >
128252	< reserved_special_token_244 >
128253	< reserved_special_token_245 >
128254	< reserved_special_token_246 >
128255	< reserved_special_token_247 >
128256 rows × 1 columns	

Vocabulaire du modèle  
Llama 3.2 1B  
→ le modèle est  
capable de  
comprendre 128256  
*tokens*.

## 2.1 – Modèle de fondation – *tokens*

Lors de l'inférence (prédiction du prochain *token* pour les modèles génératifs) ; le modèle se voit donner en entrée la liste des *id* des *tokens* de la phrase qu'il doit compléter.

La souris de l'ordinateur est un objet

4579, 155080, 334, 305, 6, 197185, 893, 537, 13839



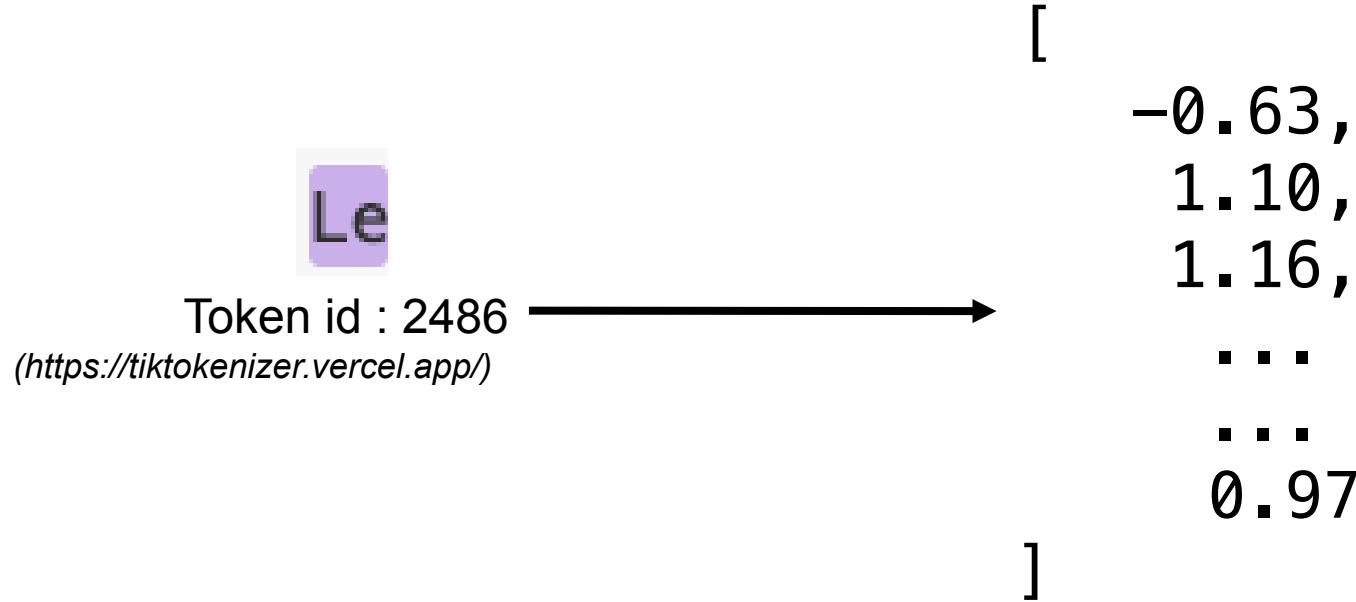
modèle

4579, 155080, 334, 305, 6, 197185, 893, 537, 13839, 22771

La souris de l'ordinateur est un objet qui

## 2.1 – Modèle de fondation – *embeddings*

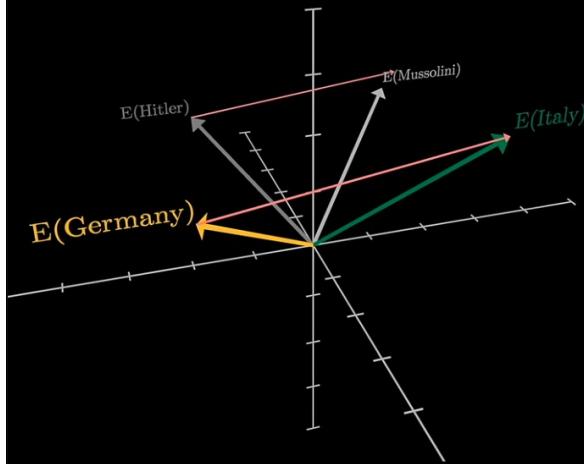
Pour prédire le camembert, le modèle utilise une représentation numérique des *tokens* : **les embeddings**  
**Ce sont des vecteurs (=liste de nombres).**



## 2.1 – Modèle de fondation – *embeddings*

$$E(\text{Hitler}) + E(\text{Italy}) - E(\text{Germany})$$

$$\approx E(\text{Mussolini})$$



Les **vecteurs** encodent alors la sémantique

2 vecteurs « **proches** » ont une signification proche.

## 2.1 – Modèle de fondation – embeddings

Exemple (un peu ancien) : cemantix, pedantix

→ jeu qui consiste à trouver un mot mystère par essai // erreur, en disant la proximité d'un mot par rapport à un mot mystère

Jour n°1157

Trouvé par 18187 personnes

8 emoji icons representing various categories like currency, temperature, and progress.

%	🌡️	°C
1000	💰	100,00
999	🥶	66,19
990	🔥	51,45
900	😱	40,78
1	😎	27,66
	🤖	0,00
	🏧	-100,00

Le mot d'hier était **pollution**

[cemantile](#)

**cemantix**

Contextual information around the word: contexte, nom pluriel, orthographe, lexical, antonyme, signifier, synonyme, secret, singulier, français, masculin, adjetif, jeu.

Trouvez le mot secret !

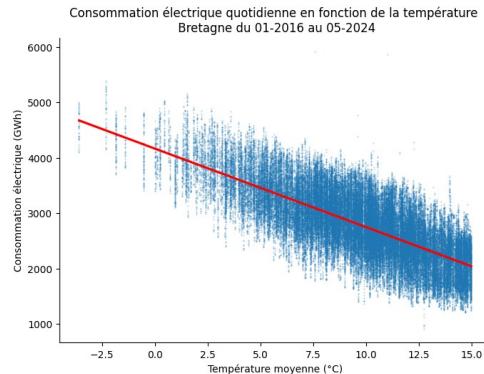
politique

Envoyer

Nº Mot	°C	% Progression
5 <b>politique</b>	39,32 😎	862
4 nature	2,47 🤪	
2 chaise	-4,48 🚫	
1 table	-11,05 🚫	
3 maison	-12,75 🚫	

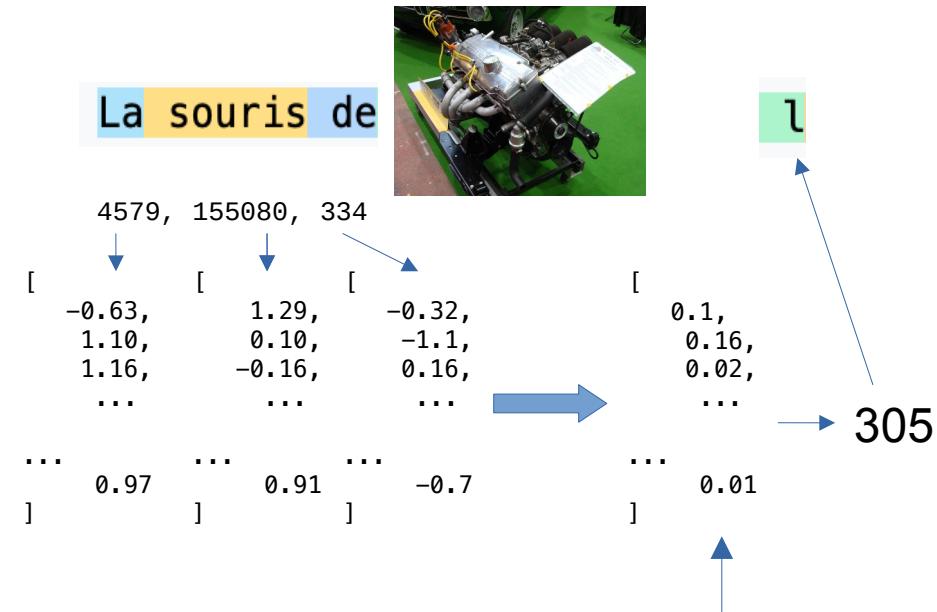
## 2.1 – Modèle de fondation – *embeddings*

La partie « prédition d'un prochain *token* » consiste en pratique à des **opérations numériques** sur les *embeddings* des *tokens* donnés en entrée.



**sortie = entrée \* pente + hauteur**

$$2800 \text{ GWh} = 10^\circ \times (-190) + 4700$$



\* En pratique, la sortie du modèle contient autant de distributions de probabilités que de tokens en entrée (voir partie 'entraînement')

Distribution de probabilité du prochain token

## 2.1 – Modèle de fondation – *embeddings*

Les *embeddings* sont utiles pour d'autres tâches que la génération de texte :

- RAG (Retrieval Augmented Generation)
- Recherche dans une base de données par similarité de sens
- Cluster (regroupement de morceaux de textes en catégories)
- Une fois un dictionnaire *token-vecteur* calculé, la traduction est très rapide (*même si on perd des informations sur le contexte*)
- Choix de la dimension (nombre de valeurs stockées dans les vecteurs)  
-> Permet de choisir l'équilibre entre les performances (temps de conversion des tokens en vecteurs), et la qualité de la réponse (des tokens de plus grande dimension donnent des réponses de meilleure qualité)

## 2.2 – Modèle de fondation – l'entraînement

La souris de l'ordinateur est un objet qui permet de manipuler les informations sur

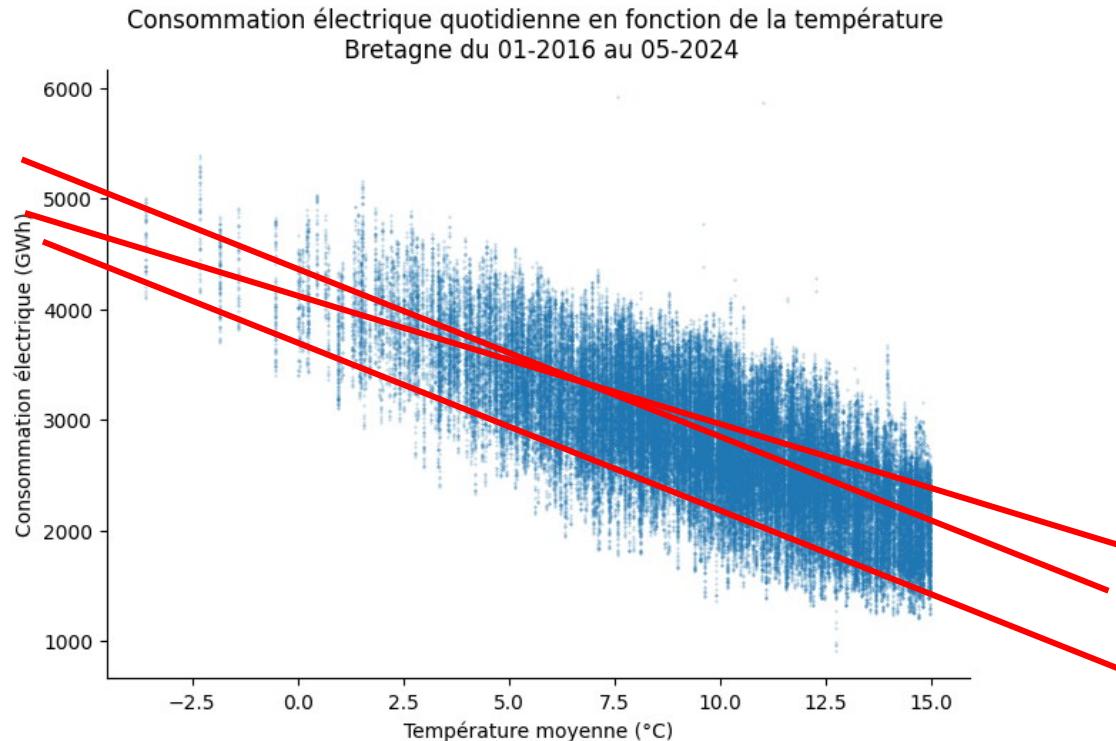


unuroyat  
microscopyaryana\_existingroc  
ketTurn\$', cardi  
microscopyintentrocketeners(  
movie Peanut:description тръ  
withObject\_translate Volt

Exemple de génération de texte sur le modèle  
Llama 3.2 1B; avec des poids réinitialisés

## 2.2 – Modèle de fondation – l'entraînement

Faire apprendre un programme informatique :



## 2.2 – Modèle de fondation – l'entraînement

Pour que le modèle apprenne à prédire du texte,  
il nous faut des exemples de textes !



≡  WIKIPÉDIA L'encyclopédie libre Rechercher sur Wikipédia Rechercher Faire un don Créer un compte Se connecter ...

 Wiki Loves FOLKLORE  Photographiez votre culture locale, aidez Wikipédia et gagnez des prix !

### Gratin dauphinois

Sommaire masquer Article Discussion Lire Modifier Modifier le code Voir l'historique Outils Apparence masquer Taille du texte Petite Standard Grande Largeur Standard Large Couleur (bêta) Automatique Clair Sombre

Début Histoire Ingrédients Record Notes et références Voir aussi Articles connexes Liens externes

Pour les articles homonymes, voir [Gratin et Dauphinois](#).

Le **gratin dauphinois** ou **pommes de terre à la dauphinoise** est un plat gratiné traditionnel de la cuisine française, d'origine dauphinoise (Sud-Est de la France), à base de pommes de terre et de lait. Ce plat est connu en Amérique du Nord comme « *au gratin style potatoos* » (États-Unis et Canada anglophone) ou « pommes de terre au gratin » (Canada francophone).

**Histoire** [ modifier | modifier le code ]

Le gratin dauphinois est officiellement mentionné pour la première fois le 12 juillet 1788 (à la veille de la Révolution française) à l'occasion d'un dîner offert aux officiers municipaux de la ville de Gap dans le Dauphiné, par le duc Jules Charles Henri de Clermont-Tonnerre, alors lieutenant

**Gratin dauphinois**   
Gratin dauphinois.

Lieu d'origine Dauphiné Place dans le Accompagnement,

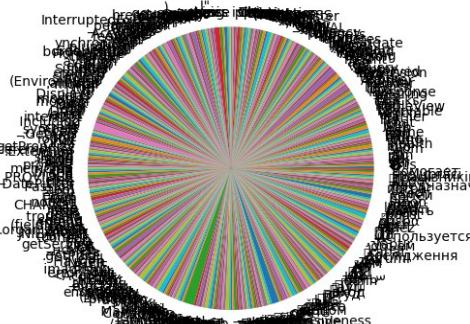
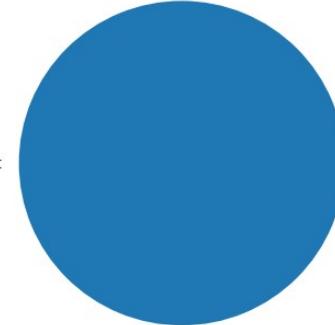
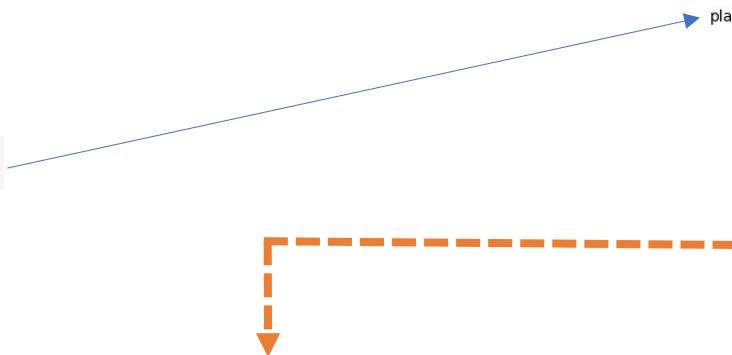
<https://commoncrawl.org>

<https://huggingface.co/datasets/HuggingFaceFW/fineweb>

## 2.2 – Modèle de fondation – l'entraînement



Le gratin dauphinois est un plat



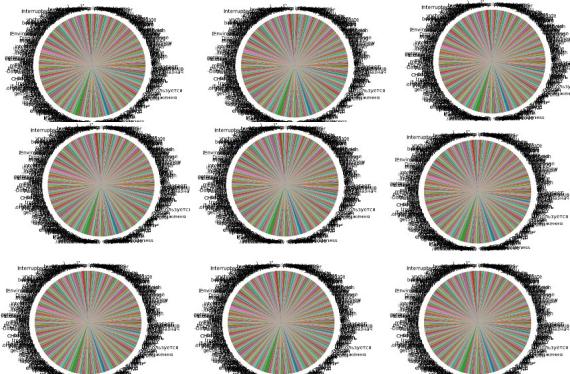
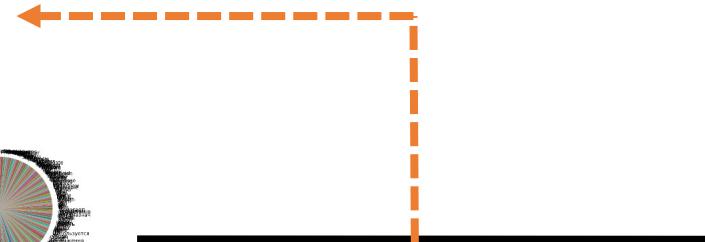
Le gratin dauphinois est un



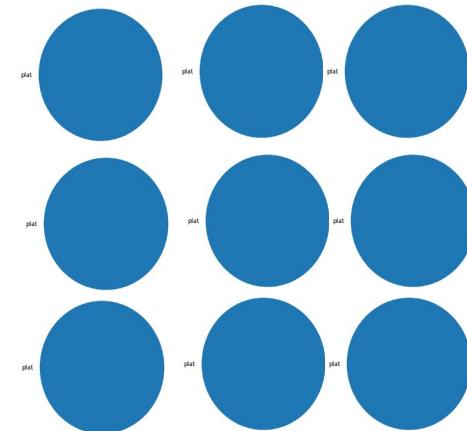
## 2.2 – Modèle de fondation – l'entraînement

En pratique, toutes les distributions de probabilités possibles sont calculée en une passe :

Le gratin dauphinois est un plat



→ un camembert par token de la séquence d'entrée



## 2.2 – Modèle de fondation – l'entraînement



Le gratin dauphinois ou pommes de terre à la dauphinoise est un plat gratiné traditionnel de la cuisine française, d'origine dauphinoise (Sud-Est de la France), à base de pommes de terre et de lait.  
Ce plat est connu en Amérique du Nord comme « au gratin style potatoes »

```
class RudiNodeManagerConnector(Connector):<newline>    def __init__(self,  
server_url: str, auth: RudiNodeAuth | dict, headers_user_agent: str = _DEFAULT  
_USER_AGENT, keep_connection: bool =False):<newline>        self.server_url = server_url  
<newline>        super().__init__(server_url=server_url, keep_connection=keep  
_connection)<newline>
```

We introduce LLaMA, a collection of foundation language models ranging from 7B to 65B parameters. We train our models on trillions of tokens, and show that it is possible to train state-of-the-art models using publicly available datasets exclusively, without resorting to proprietary and inaccessible datasets. In particular, LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla 70B and PaLM-540B. We release all our models to the research community.

<https://commoncrawl.org>

<https://huggingface.co/datasets/HuggingFaceFW/fineweb>

## 2.2 – Modèle de fondation – l'entraînement

```
Le gratin dauphinois ou pommes de terre à la dauphinoise est un plat gratiné traditionnel de la cuisine française, d'origine dauphinoise (Sud-Est de la France), à base de pommes de terre et de lait.<newline><newline>Ce plat est connu en Amérique du Nord comme « au gratin style potatoes ».
```

```
class RudiNodeManagerConnector(Connector):<newline>    def __init__(self, server_url: str, auth: RudiNodeAuth | dict, headers_user_agent: str = _DEFAULT_USER_AGENT, keep_connection: bool = False):<newline>        self.server_url = server_url<newline>        super().__init__(server_url=server_url, keep_connection=keep_connection)<newline>
```

We introduce LLaMA, a collection of foundation language models ranging from 7B to 65B parameters. We train our models on trillions of tokens, and show that it is possible to train state-of-the-art models using publicly available datasets exclusively, without resorting to proprietary and inaccessible datasets. In particular, LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla 70B and PaLM-540B. We release all our models to the research community.



Entraînement **très long**, et **très coûteux** :

- Plusieurs **mois**
- Beaucoup de matériel informatiques (cartes graphiques – **GPU**)
- Beaucoup **d'énergie**,
- De **l'eau** pour refroidir les composants informatiques.

## 2.2 – Modèle de fondation – l'entraînement

Quelques paramètres pour la génération de *tokens* :

**Température** : un nombre positif, en pratique entre 0 et 1 (jusqu'à 2 sur certains modèles). Représente la créativité du modèle

**Top p** : Permet de générer le prochain token à partir d'une portion du vocabulaire. Ne sont gardés que les tokens les plus probables (dont la somme des probabilités est supérieure ou égale à p)

**Top k** : Comme top p, mais on garde cette fois ci les « k » tokens les plus probables avant de générer le suivant.

## 2.2 – Modèle de fondation – l’entraînement

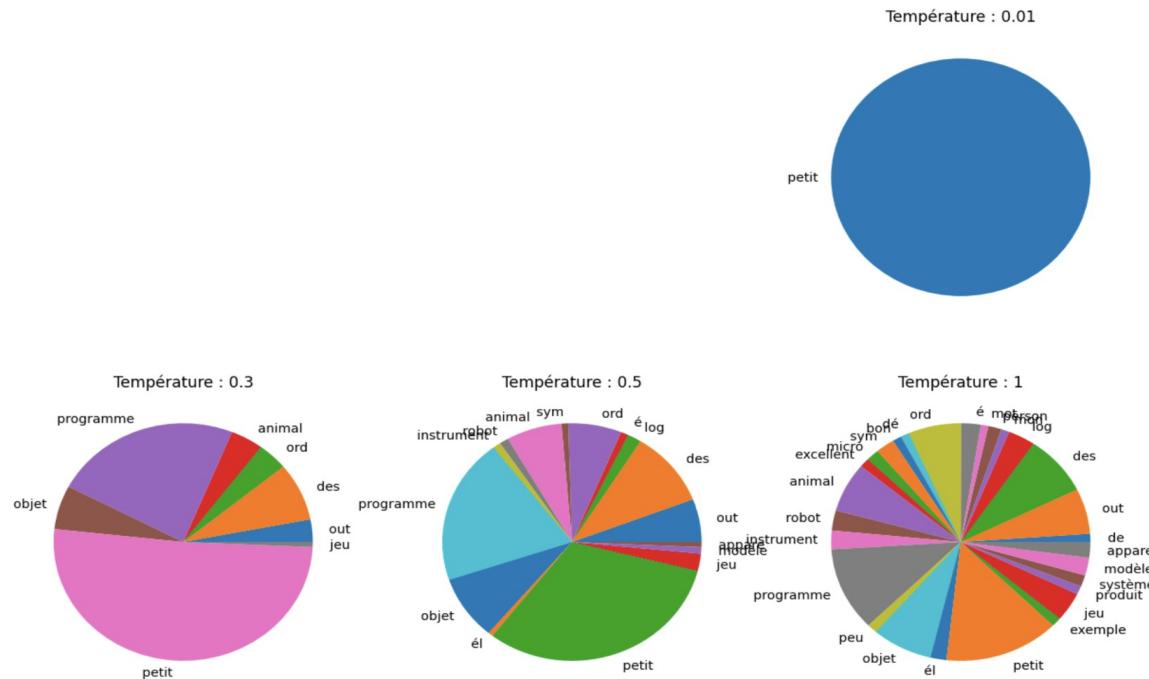


Figure 9 – Suite de la phrase : "La souris de l'ordinateur est un", avec différentes températures - Llama 3.2  
1B



# 2 – Explication du fonctionnement des LLM

## 1 – Introduction à la notion de modèle d'IA

1.1 – Qu'est-ce qu'un modèle ?

1.2 – Apprentissage Machine (*Machine Learning*), entraînement

## 2 – Explication du fonctionnement des LLM

2.1 – Modèle de fondation – génération de texte, *token*, *embeddings*

2.2 - Entraînement

2.3 – Du modèle de fondation à l'agent conversationnel

## 3 – Écosystème et historique des LLM

3.1 – Modèles propriétaires, ouverts, open-source

3.2 – Historique et chronologie de l'engouement

## 2.3 – Du modèle de fondation à l'agent conversationnel



### Modèle de fondation

« Je suis contente de pouvoir cuisiner! J'ai un petit gars qui a besoin d'attention et de nourriture. Je vais le nourrir. »

« J'ai dans mon placard des tomates pelées, du riz et du poulet. Qu'est ce que je peux cuisiner ? »

?



### Chatbot

« C'est une question classique!

Voici quelques idées de plats que vous pouvez préparer avec ces ingrédients :

**Riz et poulet aux tomates :** Faites chauffer du riz et du poulet dans une poêle, ajoutez des tomates pelées coupées en dés et du sel. Servez chaud. »

## 2.3 – Du modèle de fondation à l'agent conversationnel

Comment passer d'un modèle de fondation à un agent conversationnel ?

- On continue l'entraînement, mais avec des techniques différentes qui permettent au modèle de se comporter comme un humain qui répond aux questions

## 2.3 – Du modèle de fondation à l'agent conversationnel

### **ATTENTION**

Les méthodes utilisées pour faire évoluer un modèle de fondation en agent conversationnel sont multiples, et parfois inconnues (modèles propriétaires).

On en présente 2, qui peuvent être complémentaires; mais elles ne sont pas exhaustives. **Tous les modèles n'utilisent pas forcément les mêmes méthodes.**

## 2.3 – Du modèle de fondation à l'agent conversationnel

### 1<sup>ère</sup> méthode (**Supervised Fine Tuning**) :

-> On utilise des exemples de conversations pour que le modèle apprenne à se comporter comme un agent conversationnel

<b>text</b> string	<b>role</b> string	<b>lang</b> string
Combien de livres Hunger Games existe-t-il ? ça fait environ combien de pages ?	prompter	fr
Il existe 4 livres Hunger Games: -Hunger Games (384 pages) -Hunger Games: L'embrasement (400 pages) -...	assistant	fr
Il y a 3 ouvrages principaux à la série Hunger Games, avec un préquel. En version française, ça...	assistant	fr
Il y a 3 livres principaux (1 préquel) et ça fait beaucoup de pages : Livre 1 : 399 Livre 2 :400 Livr...	assistant	fr

## 2.3 – Du modèle de fondation à l'agent conversationnel

**1<sup>ère</sup> méthode (Supervised Fine Tuning) :**  
 -> Cela permet de conditionner le modèle

text	role	lang
string	string	string

Qui de l'oeuf ou de la poule est apparu en premier ?	prompter	fr
Cette question n'a pas vraiment de réponse.	assistant	fr
...		

text	role	lang
string	string	string
Mon amie est considérée comme "harceleuse" dans sa communication avec une soignante. Comment puis-je réagir pour l'aider ?...	prompter	fr
Il est important de comprendre d'où peut venir ce comportement afin de mieux comprendre comment y réagir. Les comportements harcelants peuvent être le résultat de facteurs tels que le stress, la frustration, l'anxiété, la colère o...	assistant	fr
Le harcèlement est grave et il est important de prendre au sérieux les comportements de harcèlement, qu'ils soient commis en personne ou en ligne, car ils peuvent causer un préjudice émotionnel et psychologique considérable aux victimes....	assistant	fr
Avant tout , je dois vous rappeler que je ne suis qu'un chatbot , sans aucune qualification pour des questions d'ordre vital , médical , légal ou autres...	assistant	fr

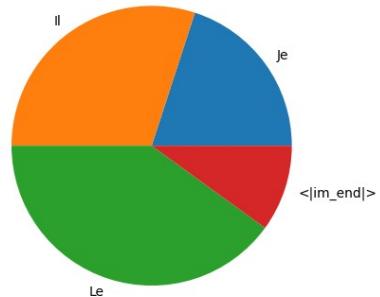
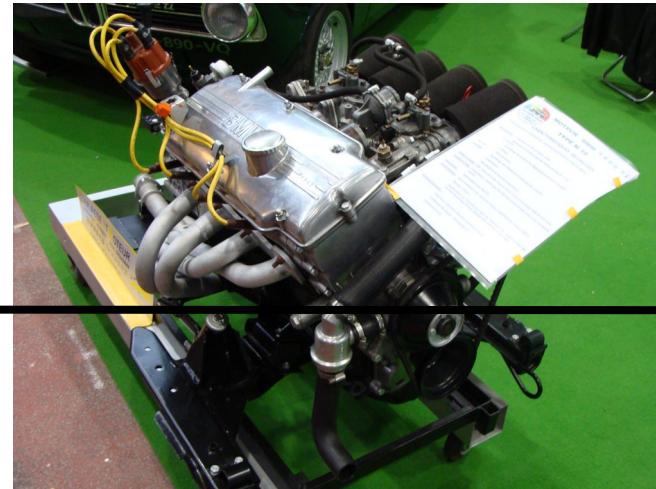
## 2.3 – Du modèle de fondation à l'agent conversationnel

**1<sup>ère</sup> méthode (Supervised Fine Tuning) :**

```
<|im_start|>user<|im_sep|>Combien de livres Hunger Games existe-t-il ? ça fait environ combien de pages ?<|im_end|><|im_start|>assistant<|im_sep|>Il existe 4 livres Hunger Games : -Hunger Games (384 pages) -Hunger Games: L'embrasement (400 pages) -Hunger Games: La Révolte (339 pages) -Hunger Games: La Ballade du Serpent de l'Oiseau Chanteur (1712 pages) 384+400+339+1712= 2835 pages (Le nombre de page est calculé à partir des éditions originales de ces livres, et non des versions françaises)<|im_end|><|im_start|>assistant<|im_sep|>
```

Il

```
<|im_start|>system<|im_sep|>Combien de livres Hunger Games existe-t-il ? ça fait environ combien de pages ?<|im_end|><|im_start|>assistant<|im_sep|>
```

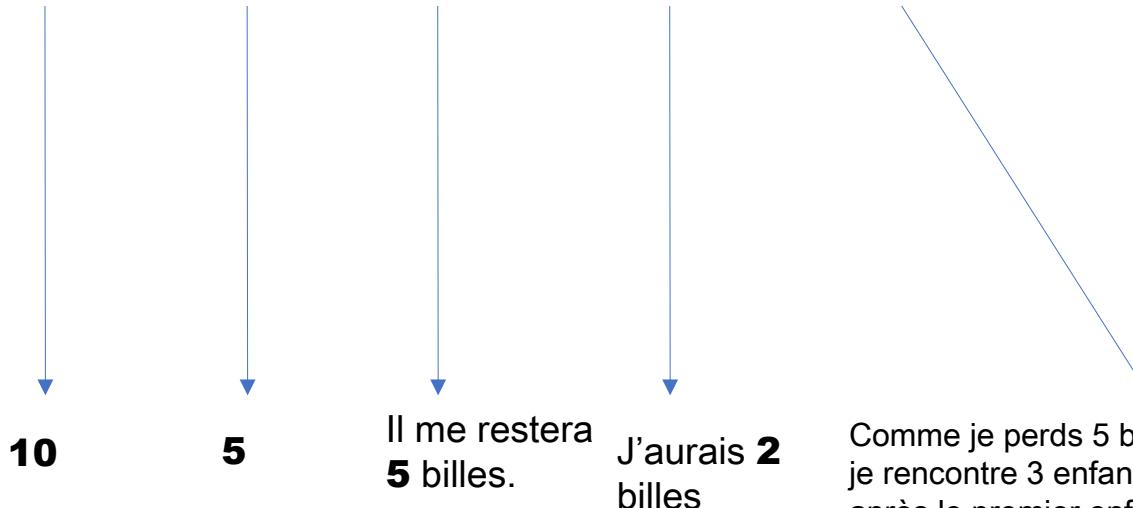


## 2.3 – Du modèle de fondation à l'agent conversationnel

### 2<sup>ème</sup> méthode (**Reinforcement Learning**) :

-> On laisse le modèle présenter ses réponses, et on l'ajuste en fonction

Si tu as 20 billes et tu en donnes 5 à chaque enfant que tu rencontres sur ton chemin, combien en auras-tu après avoir rencontré 3 enfants ?



Il me restera  
**5** billes.  
J'aurais **2**  
billes

Comme je perds 5 billes par enfant et que je rencontre 3 enfants, cela fait  $20 - 5$  après le premier enfant, soit 15, puis  $15 - 5$  pour le second, soit 10 et enfin  $10 - 5$  après le dernier. Il me restera **5** billes.

## 2.3 – Du modèle de fondation à l'agent conversationnel

Le processus d'entraînement peut être customisé, afin de créer des modèles qui **s'adaptent** à des besoins précis.

- > Données d'entraînement spécialisées dans un domaine
- > On parle de **fine-tuning**

## 2.4 – Utilisation d'un modèle

Une fois le modèle entraîné, plusieurs manières de l'utiliser :

- **localement** : tout se passe sur votre ordinateur (même sans connexion internet le modèle fonctionnera)
- **distant** : les calculs (l'inférence) faits par le modèle sont effectués sur une autre machine

Les calculs nécessaires à l'inférence sont assez importants ; même s'ils dépendent du modèle.

## 2.4 – Utilisation d'un modèle

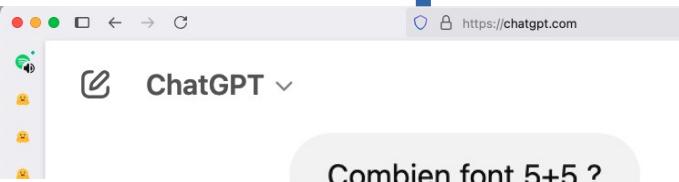
### Localement :

- via une application qui crée une interface entre vous et le modèle : Ollama, OpenWebUI, ...
- directement en codant en Python, avec des bibliothèques de codes disponibles gratuitement (*transformer* d'hugging face par exemple - <https://huggingface.co/docs/transformers/index>)

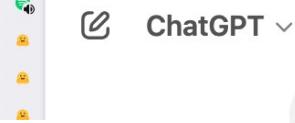
### Distant :

- quelqu'un d'autre utilise une solution locale sur une machine, et la rend accessible *via* un réseau (intra/internet) :
  - ChatGPT, RagARenn, HuggingFace Spaces, ...

## 2.4 – Utilisation d'un modèle



```
[  
  {  
    "role": "user",  
    "content": "Combien font 5+5 ?"  
  }  
]
```

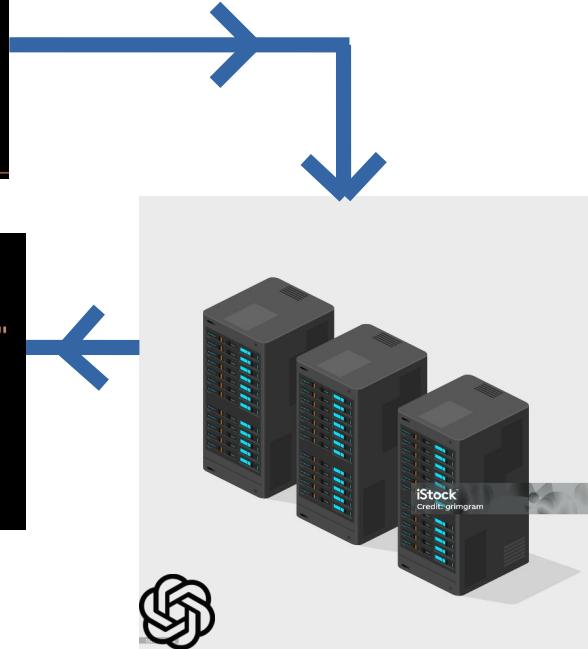


Combien font 5+5 ?

5 + 5 font 10.



```
[  
  {  
    "role": "user",  
    "content": "Combien font 5+5 ?"  
  },  
  {  
    "role": "assistant",  
    "content": "5+5 font 10."  
  }  
]
```





# 3 – Écosystème et historique des LLM

## 1 – Introduction à la notion de modèle d'IA

1.1 – Qu'est-ce qu'un modèle ?

1.2 – Apprentissage Machine (*Machine Learning*), entraînement

## 2 – Explication du fonctionnement des LLM

2.1 – Modèle de fondation – génération de texte, *token*, *embeddings*

2.2 - Entraînement

2.3 – Du modèle de fondation à l'agent conversationnel

## 3 – Écosystème et historique des LLM

3.1 – Modèles propriétaires, ouverts, open-source

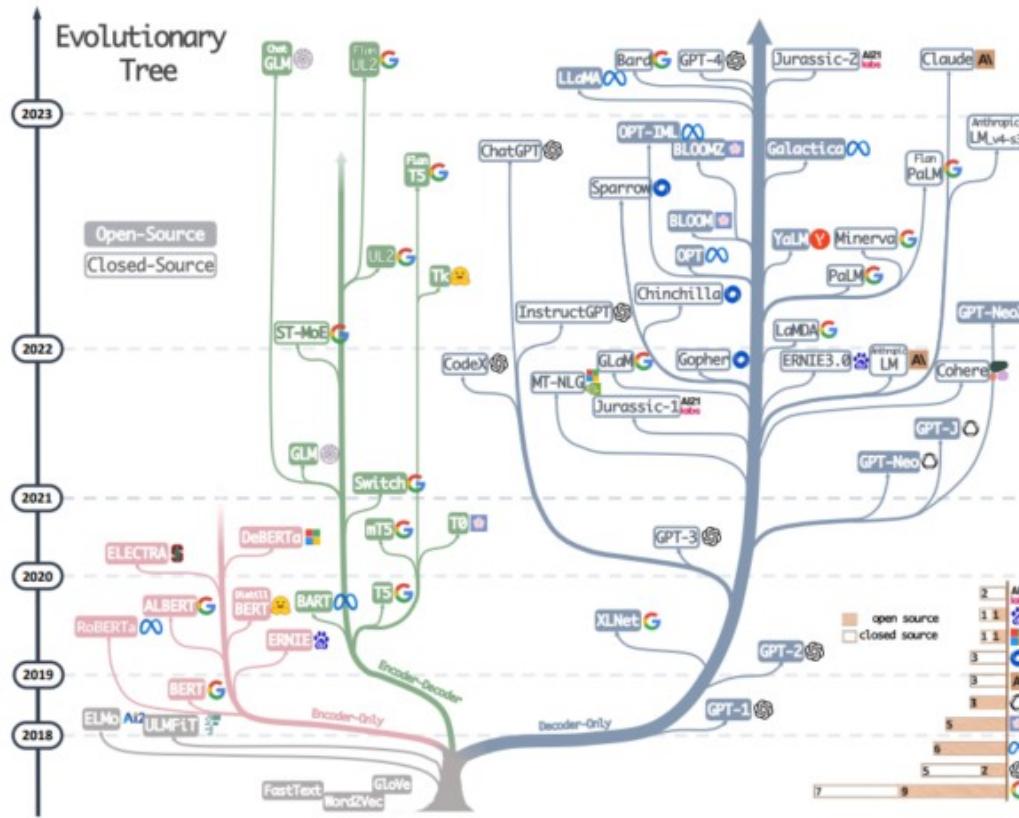
3.2 – Historique et chronologie de l'engouement

## 3.1 – Écosystème – Chronologie

- 
- 2017: Transformer - analyse contextuelle efficace, base technique de tous les modèles
  - 2018: BERT - comprendre le contexte avant/après chaque mot  
Amélioration pour la recherche et l'analyse de texte
  - 2020: GPT-3  
Génération de texte cohérent, premier grand modèle de compréhension de texte généraliste
  - 2022: ChatGPT - Passage à l'échelle massive (175B)  
Première interface conversationnelle grand public.

À suivre...

## 3.1 – Écosystème – Chronologie

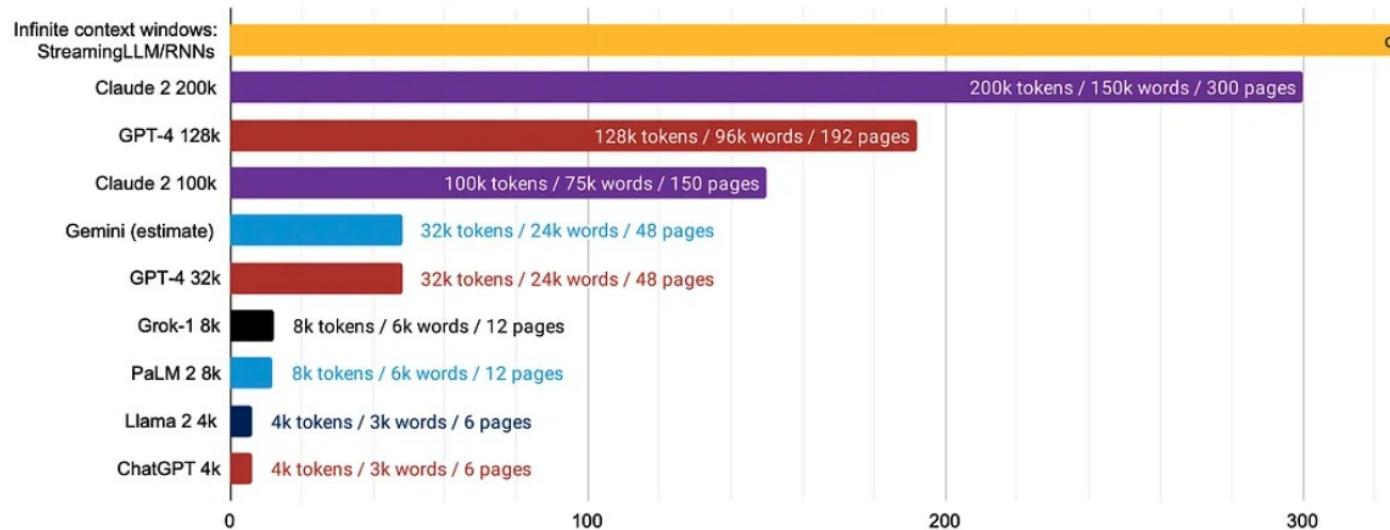


## 3.1 – Écosystème – Différencier les LLMs

- **Accès:** modèles propriétaires et ouverts
  - \* Tendance fermée: licences restrictives, absence d'accès au code/poids, API... (ex: GPT, Claude, Gemini)
  - \* Tendance ouverte: licences permissives, accès libre au code et aux poids, auto-hébergement (ex: Llama, Mistral, DeepSeek)
- **Taille des paramètres** (petit: ~1B, 7B, moyen: ~32B, grand: 100B à + de 1T paramètres)
- **Taille du contexte** (4k tokens, 128k, 1M+)
- **Cas d'usages / spécialisation** (portabilité, raisonnement, ...)
- **Formats:** ajustés, quantisés, distillés...

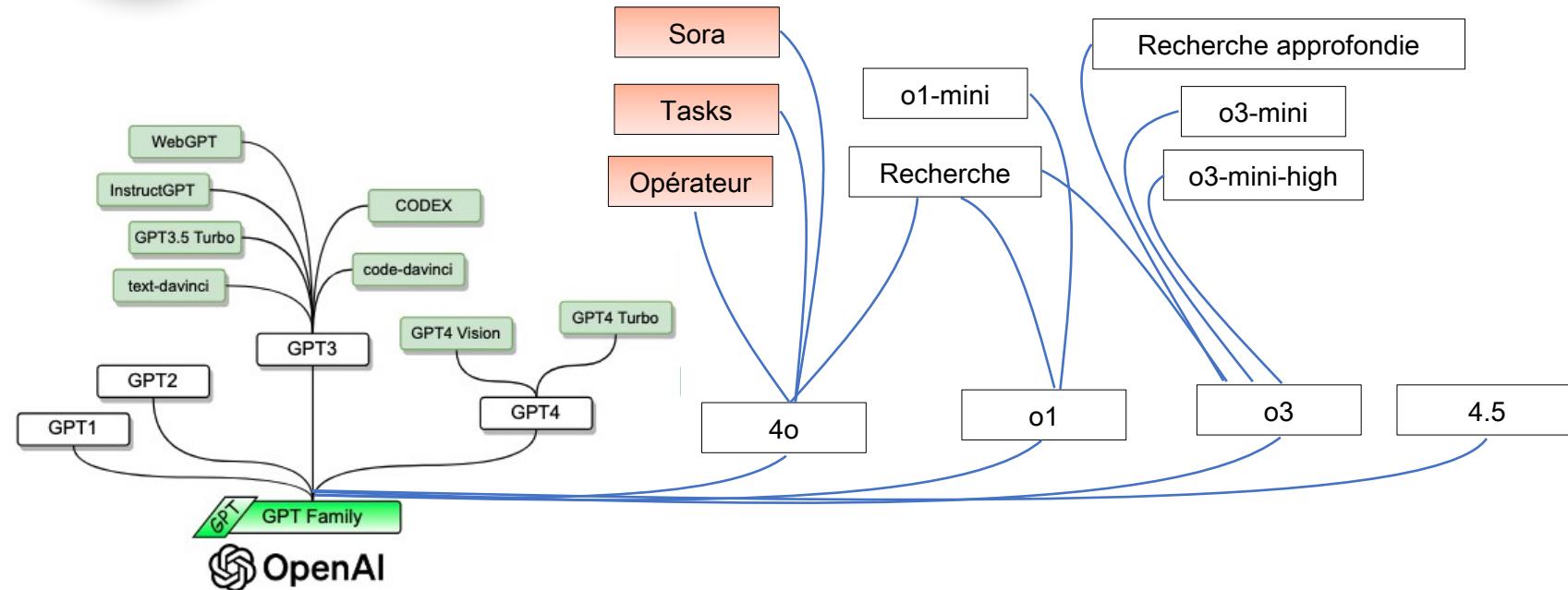
## 3.1 – Écosystème – Un point de comparaison

### Un point de comparaison: le contexte



Source (2024): Alan D. Thompson (LifeArchitect.ai)

## 3.1 – Écosystème – Exemple de la famille GPT



# 3.1 – Écosystème – Exemple : les modèles Mistral

Des modèles selon les **cas d'usages**

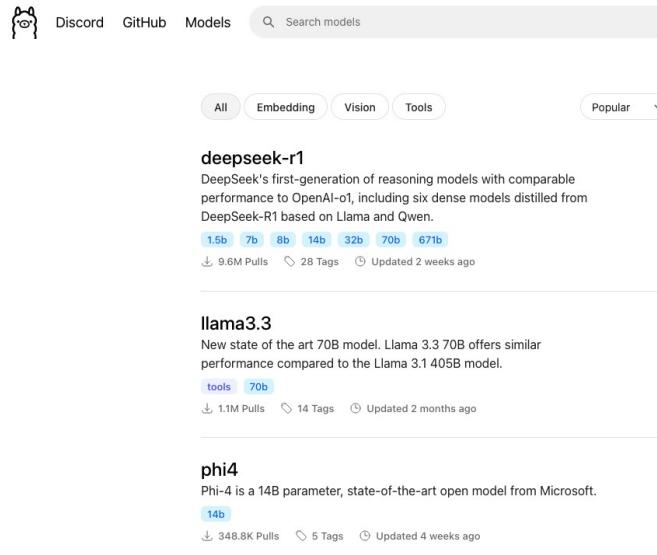


Source: Mistral

...Mistral Edge (Minstral 3B et Minstral 8B)

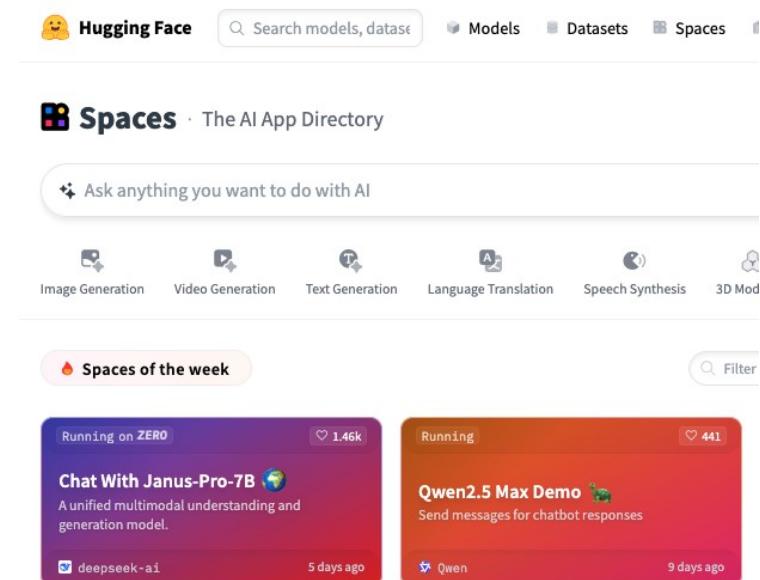
## 3.2 – L'exploration des modèles

Essayez les modèles avec Ollama (`ollama run <model>`)  
ou directement sur HuggingFace Spaces



The screenshot shows the Ollama web interface. At the top, there are links for Discord, GitHub, Models, and a search bar. Below the search bar are buttons for "Sign in" and "Download". A "Popular" dropdown menu is open. Underneath, there are three model cards:

- deepseek-r1**: DeepSeek's first-generation of reasoning models with comparable performance to OpenAI-o1, including six dense models distilled from DeepSeek-R1 based on Llama and Qwen. It has 1.5b, 7b, 8b, 14b, 32b, 70b, and 671b variants. It has 9.6M Pulls, 28 Tags, and was updated 2 weeks ago.
- llama3.3**: New state of the art 70B model. Llama 3.3 70B offers similar performance compared to the Llama 3.1 405B model. It has tools and 70b variants. It has 1.1M Pulls, 14 Tags, and was updated 2 months ago.
- phi4**: phi4 is a 14B parameter, state-of-the-art open model from Microsoft. It has 14b variants. It has 348.8K Pulls, 5 Tags, and was updated 4 weeks ago.



The screenshot shows the HuggingFace Spaces interface. At the top, there are links for Models, Datasets, and Spaces. Below the search bar, there are sections for "Spaces · The AI App Directory" and "Ask anything you want to do with AI". There are also buttons for various AI tasks: Image Generation, Video Generation, Text Generation, Language Translation, Speech Synthesis, and 3D Mod. A "Spaces of the week" section highlights two spaces:

- Running on ZERO** by Chat With Janus-Pro-7B (1.46k stars)
- Running** by Qwen2.5 Max Demo (441 stars)

At the bottom, there are links for "deepseek-ai" (5 days ago) and "Qwen" (9 days ago).

## FOLLOW US!

 [www.irisa.fr](http://www.irisa.fr)

 [irisa-lab](#)

 [@irisa\\_lab](#)

