



Technologies et perspectives sur les chatbots

8 septembre 2025

PRÉSENTATION



Hello everyone, dear students of EPF, it's Barack Obama. I'm delighted to recommend my friend Olivier Guérin for this introduction to artificial intelligence.

Don't hesitate to apply to Artik Consulting for your next internship, you will learn a lot there. I wish you a great training day with Olivier.



ARTIK CONSULTING

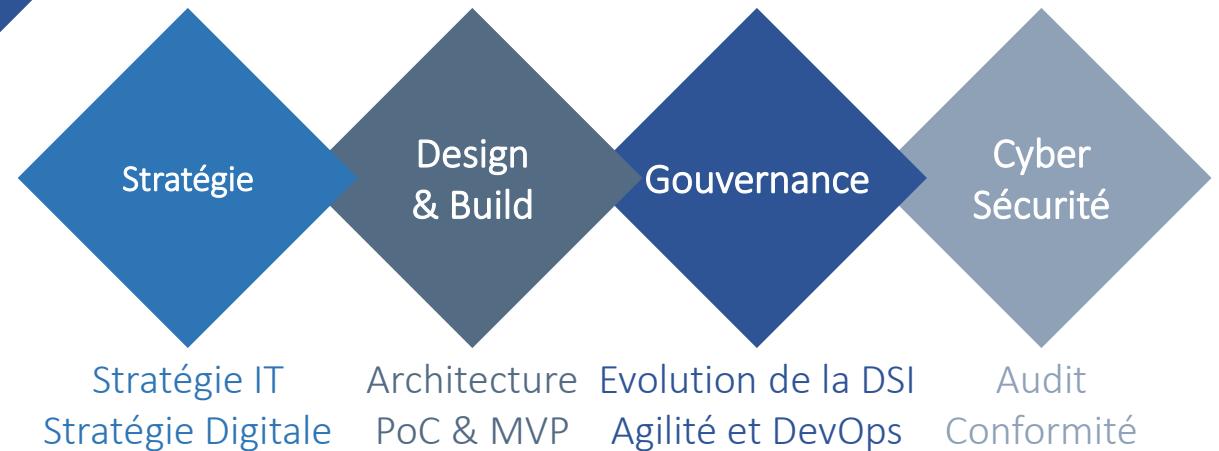
QUI SOMMES-NOUS ?



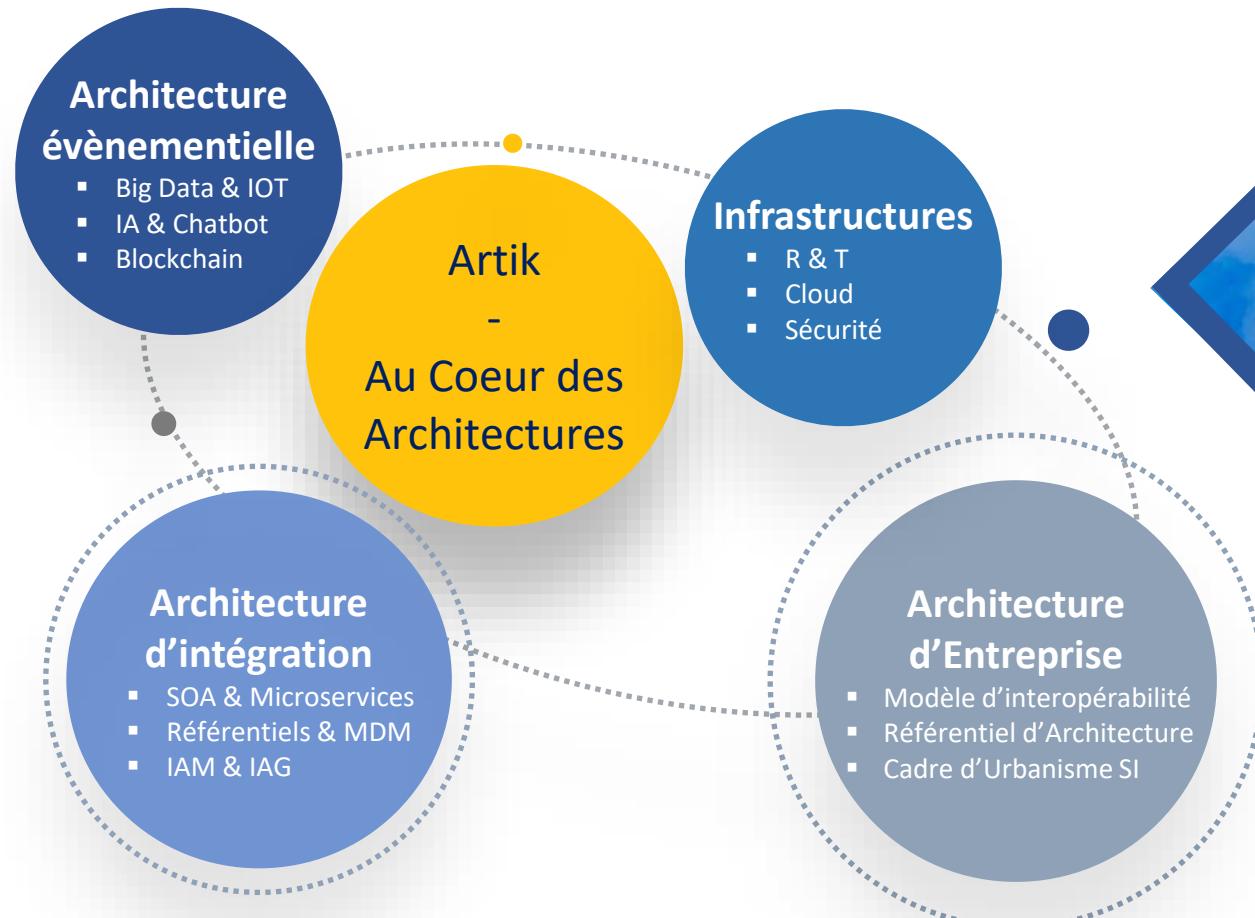


L'architecture informatique est une science en perpétuelle transformation qui nécessite de se projeter dans l'avenir tout en s'appuyant sur les acquis du passé

Un Cabinet de Conseil en Architecture & Management des SI



L'expertise de nos consultants s'exerce dans des domaines où l'innovation est au service de la stratégie d'entreprise et de la transformation du SI



Des clients grand comptes tels que Bpifrance, Bioline, CDC, EDF, Kering, La Poste, RTE, SANEF, SNCF...

Savoir-faire et savoir-être sont les clés de notre développement

10 collaborations avec des startups et campus universitaires

REJOIGNEZ ARTIK VOUS NE LE REGRETTEREZ PAS !!!



- Vous recherchez un stage dans une structure à taille humaine.
- Vous avez envie de travailler sur des projets à forts enjeux Métiers pour des grandes entreprises.
- Les nouvelles technologies vous passionnent.
- Vous souhaitez apprendre vite et faire de votre stage un véritable tremplin vers votre premier job.

DIFFÉRENTS TYPES DE STAGE

Ce que vous allez faire en tant que stagiaire

- Intervenir sur des missions clients.
- Réaliser de la veille technologique.
- Suivre des formations.
- Contribuer au développement du cabinet (recrutement, marketing, commercial...).
- Et à sa vie interne : Forum technologiques, Week-end, Diners...

Stage Consultant SI

- Contribuer à l'élaboration d'une stratégie IT.
- Accompagner des clients dans leurs choix technologiques.
- Participer à un projet SI de la phase d'opportunité à la mise en production.

Stage DevOps

- Concevoir une chaîne CI/CD.
- Mettre en place les outils du DevOps.
- Accompagner les équipes dans leur montée en compétence sur les principes du DevOps.

Stage Développeur Big Data

- Être intégré dans une Squad Agile.
- Participer à la conception d'une Feature Big Data.
- Réaliser les développements sur des outils tels que Kafka, Spark...
- Tester et recetter les développements.

Stage Commercial

- Assister notre équipe commerciale dans la prise de rendez-vous
- Participer à des avant-ventes.
- Contribuer à la rédaction de propositions commerciales.
- Travailler sur le marketing de l'offre.

SOMMAIRE

- Actualité IA et chatbots
- Introduction générale à l'IA
- Tour d'horizon des IA génératives
- Qu'est-ce qu'un LLM ?
- Etat de l'art des LLM en 2024
- TP : Comparer un LLM généraliste / un LLM conversationnel
- Tour d'horizon des LLM propriétaires et Open Source
- Impact pour les DSI, retour d'expérience et perspectives de marché
- Les défis liés au comportement des LLM
- Sécurité
- TD : Maîtriser le prompting avec la génération d'image
- Evaluation : QCM (1 heure)



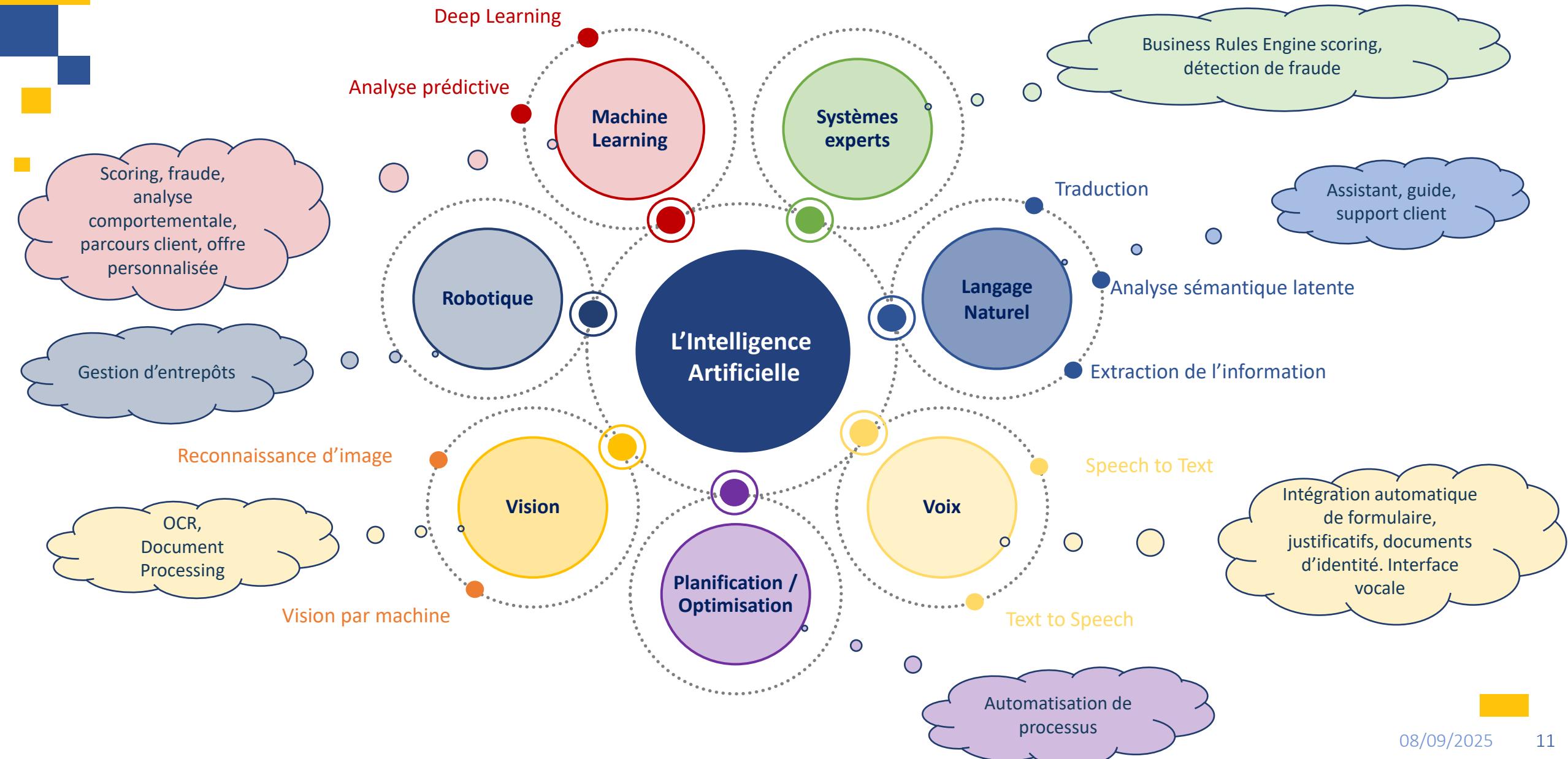
INTRODUCTION GÉNÉRALE À L'INTELLIGENCE ARTIFICIELLE



MAIS C'EST QUOI L'INTELLIGENCE ARTIFICIELLE

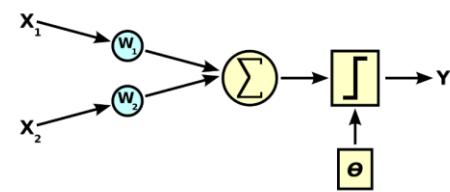
- L'intelligence est la faculté d'apprendre, de comprendre et de s'adapter à des situations nouvelles.
- Le développement de l'intelligence est déterminé non seulement par les gènes, mais également par les expériences vécues. Toutes les études menées sur les bébés le confirment, les enfants naissent bardés de connaissances (**l'inné ~ 55 %**) et possèdent tous les mécanismes d'apprentissage nécessaires pour développer leur cerveau (**l'acquis ~ 45 %**).
- L'intelligence Artificielle recouvre l'ensemble des théories et des techniques permettant d'élaborer des programmes informatiques capables de simuler certains traits de l'intelligence humaine notamment l'apprentissage et le raisonnement.
- Autrement dit, l'Intelligence Artificielle désigne des systèmes ou des machines qui d'une part « imitent » l'intelligence humaine dans la résolution de problèmes et d'autre part s'améliorent de manière itérative en fonction des informations qu'ils recueillent.

L'IA COUVRE DE NOMBREUX DOMAINES



DU FANTASME À LA RÉALITÉ MATHÉMATIQUE

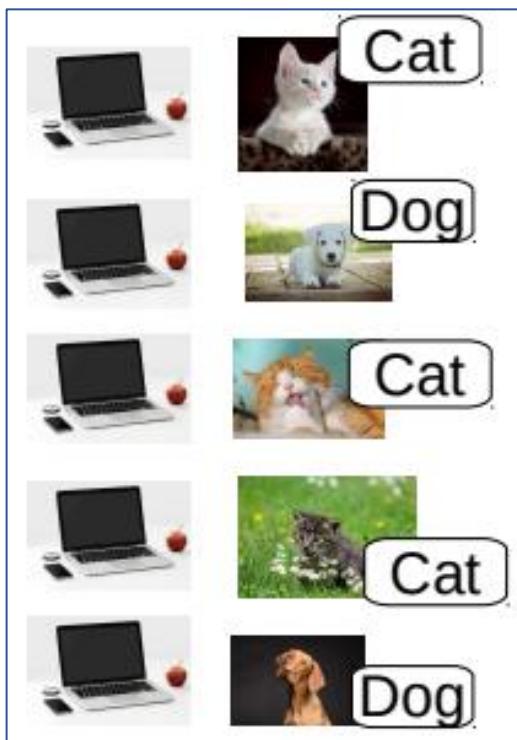
- Le Machine Learning : Approche fondée sur les mathématiques et les statistiques permettant à un ordinateur d'améliorer les résultats d'un traitement, sans modification du code, en s'appuyant exclusivement sur de l'analyse de données. Les principaux usages sont la classification, la prise de décision et la prédiction.
- Le Neurone : Objet mathématique à n entrées (x_1 à x_n) et une sortie y valant 0 ou 1, qui pour un neurone formel est obtenue en faisant la somme pondérée (w_1 à w_n) des entrées qui est ensuite comparée à une valeur seuil Θ dans une fonction de Heaviside.
- Les Réseaux de Neurones : Assemblage de neurones organisés en couches, chaque couche (hidden layer) étant en charge d'affiner l'analyse de la précédente.
- Le Clustering (ou partitionnement des données) : Méthode de classification non supervisée dont le but est de regrouper entre elles des données présentant des propriétés similaires.
- Le Deep Learning : Extension du machine learning, qui est basé sur les réseaux de neurones et qui utilise une plus grande masse de données d'apprentissage, en revanche l'explicabilité n'est pas toujours au rendez-vous.



APPRENTISSAGE SUPERVISÉ / NON SUPERVISÉ

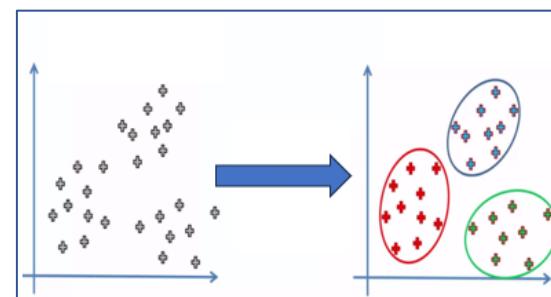
Apprentissage supervisé

- Les données sont **labélisées**.
- Les données d'entraînement disposent d'**un label connu**.
- L'algorithme apprend à **prédir** le label de la donnée fournie en entrée.

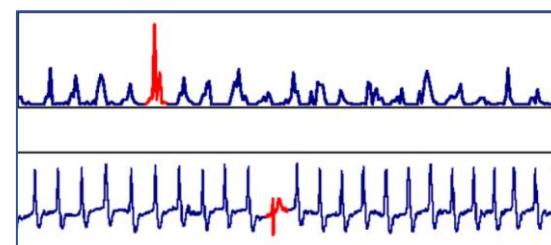


Apprentissage non supervisé

- Les données ne sont pas **labélisées**. Il n'y a pas alors de bonne réponse.
- Le but est de **détecter** des comportements ou des relations dans nos données.
- Il existe différentes techniques les plus répandues sont :



Le **clustering** est une technique d'analyse de données qui regroupe des éléments similaires ensemble en fonction de leurs similarités ou de leurs caractéristiques communes.



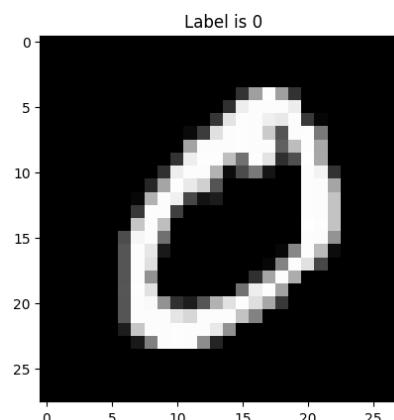
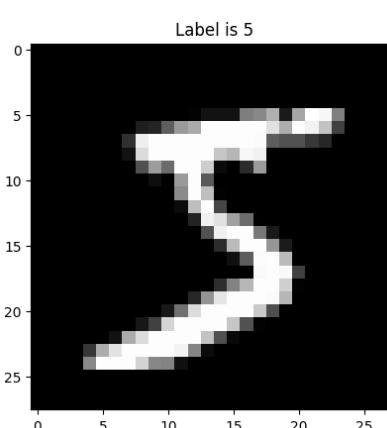
La **détection d'anomalies** consiste à identifier des observations rares, inhabituelles ou divergentes dans un ensemble de données, qui diffèrent significativement du comportement normal ou attendu.

UNE COURTE INTRODUCTION AUX RÉSEAUX DE NEURONES

- Les **réseaux de neurones artificiels** ont été inventés dans le but d'imiter le fonctionnement du cerveau humain et de ses neurones. Les premières idées conceptuelles relatives aux réseaux de neurones datent des années 1940, avec le modèle de McCulloch-Pitts.
- Cependant, ce n'est que dans les années 1960 que les premiers modèles formels de neurones ont été proposés, tels que le **perceptron de Rosenblatt**. Les réseaux de neurones ont connu plusieurs phases d'essor et de désintérêt jusqu'à l'arrivée d'ordinateurs disposant d'une puissance de calcul suffisante pour les rendre utilisables.
- Un réseau de neurones artificiel fonctionne par **couches (layers)**. Il reçoit en entrée des données, lesquelles sont traitées par une ou plusieurs couches cachées avant d'obtenir un résultat en sortie.
- Chaque couche est composée de plusieurs neurones, et chaque neurone prend en entrée des valeurs provenant de la couche précédente, applique une combinaison linéaire de ces valeurs (en ajoutant également un biais), puis applique une **fonction d'activation (activation function)** non linéaire.
- Les **poids (weights)** associés à chaque entrée dans la combinaison linéaire sont ajustés au cours de l'apprentissage pour minimiser une **fonction de perte (loss function)**, c'est-à-dire l'écart entre les prédictions du réseau et les valeurs réelles.
- Ce processus d'ajustement est souvent effectué en utilisant une technique appelée **rétropropagation du gradient (backward propagation)**.

UNE COURTE INTRODUCTION AUX RÉSEAUX DE NEURONES

- L'IA et les réseaux de neurones apparaissent parfois comme des technologies « boîtes noires », où seuls les experts peuvent comprendre le fonctionnement interne. Cependant, les réseaux de neurones sont des objets mathématiques que tout le monde peut comprendre.
- MNIST (Mixed National Institute of Standards and Technology) est un jeu de données qui a été constitué par Yann Le CUN, actuel directeur IA chez Meta. Il est devenu l'exemple de référence pour entraîner un réseau de neurones à reconnaître un nombre écrit à la main à partir d'une image pixelisée.
- Une version de MNIST au format csv a été partagé sur la plateforme de compétition Kaggle. En voici le contenu :
 - mnist_train.csv : contient les 60 000 exemples d'entraînement et leurs étiquettes.
 - mnist_test.csv : contient 10 000 exemples de test et leurs étiquettes.
- Chaque ligne se compose de 785 valeurs : la première valeur est l'étiquette / label (un nombre de 0 à 9, c'est-à-dire le résultat attendu) et les 784 valeurs restantes sont les valeurs des pixels (un nombre de 0 à 255).



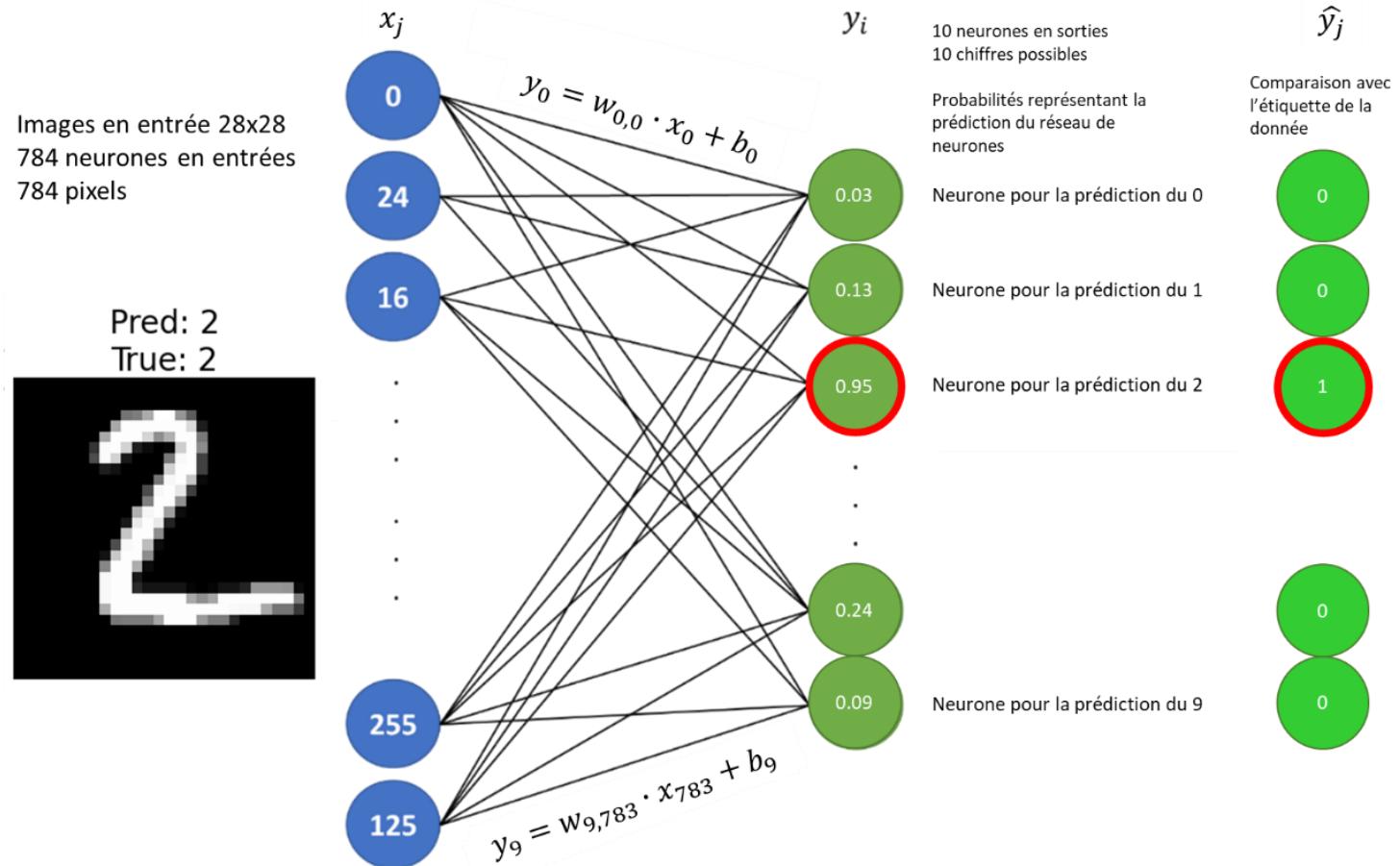
	A	B	C	D	E	F	G	H	I	J	K
1	label	1x1	1x2	1x3	1x4	1x5	1x6	1x7	1x8	1x9	1x10
2	5	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0
4	4	0	0	0	0	0	0	0	0	0	0
5	1	0	0	0	0	12	5	65	22	0	0
6	9	0	0	0	0	0	0	0	0	0	0
7	2	0	0	0	0	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0	0	0	0
9	3	0	0	0	0	0	0	0	0	0	0
10	1	0	0	0	0	0	0	0	0	0	0

Données au format CSV des images de nombres manuscrits. Chaque ligne est une image, chaque colonne est la position d'un pixel. Les valeurs de pixels vont de 0 (noir) à 255 (blanc).

Source : Kaggle et Artik Consulting

UN RÉSEAU DE NEURONES BASIQUE

- Soit un réseau de neurones sans couche intermédiaire, où les entrées et les sorties sont directement liées entre elles.
- Il y a 10 sorties possibles, qui correspondent chacune à un nombre à reconnaître.
- Toutes les entrées sont liées à toutes les sorties via une fonction linéaire.
- La valeur de la sortie est une probabilité, plus elle est élevée plus la reconnaissance est exacte
- Ensuite il sera possible d'ajouter une couche de non-linéarité destinée à l'activation ou non du neurone.



Exemple d'un réseau de neurones sans couche intermédiaire où les neurones d'entrée et de sortie sont reliés par une fonction linéaire. Source : Artik Consulting.

L'ENTRAINEMENT DU RÉSEAU DE NEURONES

Dans notre cas, la représentation mathématique du réseau de neurones est un système d'équation multilinéaire avec :

- $784 * 10 = 7840$ paramètres $w_{i,j}$ qui sont les poids de la relation linéaire à déterminer.
- 10 paramètres b_i qui sont les biais à déterminer.
- 784 variables x_j représentant la nuance de gris d'un des pixels de l'image ;
- 10 variables y_i représentant la probabilité que l'image soit le nombre i.

$$\begin{cases} w_{0,0}x_0 + w_{0,1}x_1 + \cdots + w_{0,783}x_{783} + b_0 = y_0 \\ \quad \quad \quad \cdots \\ w_{9,0}x_0 + w_{9,1}x_1 + \cdots + w_{9,783}x_{783} + b_9 = y_9 \end{cases}$$

L'ENTRAINEMENT DU RÉSEAU DE NEURONES

Ce système linéaire peut s'écrire sous forme de sommes et sous forme matricielle :

$$\begin{cases} \sum_{j=0}^{783} w_{0,j}x_j + b_0 = y_0 \\ \dots \\ \sum_{j=0}^{783} w_{9,j}x_j + b_9 = y_9 \end{cases} \Leftrightarrow \begin{bmatrix} w_{0,0} & \dots & w_{0,783} \\ \vdots & \ddots & \vdots \\ w_{9,0} & \dots & w_{9,783} \end{bmatrix} \cdot \begin{bmatrix} x_0 \\ \dots \\ x_{783} \end{bmatrix} + \begin{bmatrix} b_0 \\ \dots \\ b_9 \end{bmatrix} = \begin{bmatrix} y_0 \\ \dots \\ y_9 \end{bmatrix} \Leftrightarrow \mathbf{w} \cdot \mathbf{x} + \mathbf{b} = \mathbf{y}$$

L'objectif est d'obtenir la matrice de poids et le vecteur de biais, soit résoudre un système avec 7850 inconnues ! Naturellement, une solution exacte et analytique est peu envisageable. la résolution numérique est beaucoup plus efficace et simple à mettre en place.

Le critère de résolution est la **fonction de coût**, aussi appelé la fonction de perte (**cost / loss function**). Dans notre exemple simplifié nous utilisons un critère simple : la méthode des moindres carrés. L'objectif est de réduire au maximum cette fonction de coût, afin de rapprocher le plus possible les prédictions du réseau de neurones avec les étiquettes des jeux de données.

L'ENTRAINEMENT DU RÉSEAU DE NEURONES

L'entraînement du réseau de neurones se déroule de la manière suivante :

- Choisir de manière aléatoire les valeurs initiales des paramètres $w_{(i,j)}$ et b_i ;
- Injecter une donnée du jeu d'entraînement et calculer la prédiction pour chaque neurone de sortie.
- Sélectionner la valeur de prédiction la plus importante, et comparer avec l'étiquette de la donnée.
- Calculer la fonction de coût.
- Mettre à jour les poids et les biais à partir de calculs de dérivée (backpropagation).
- Réinjecter une nouvelle donnée d'entraînement et passer sur toutes les données.
- Passer plusieurs fois sur l'ensemble du jeu d'entraînement (nombre d'époques, epochs).
- Tester le résultat de l'entraînement en injectant les données du jeu de test et calculer le taux de réussite.

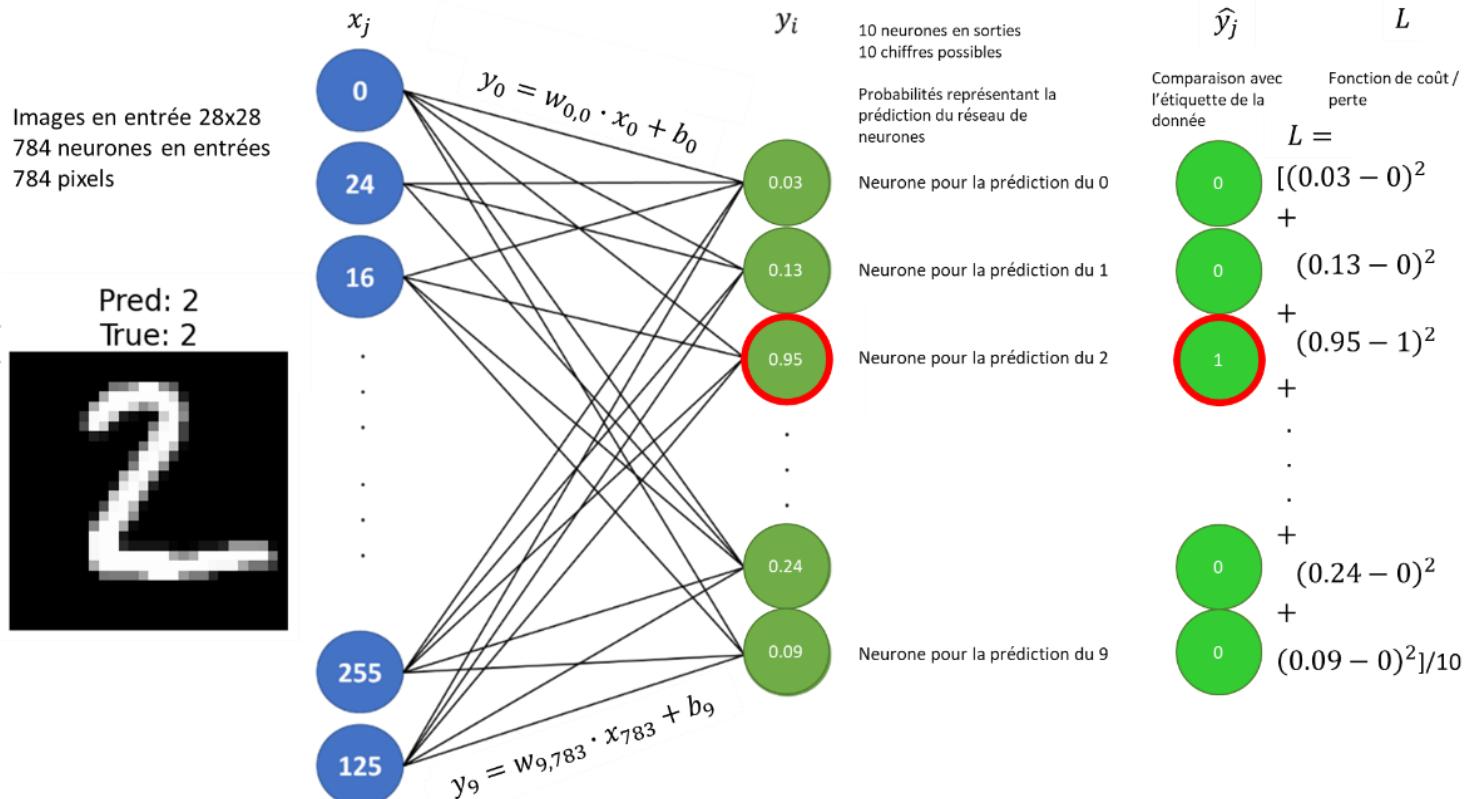


Illustration du calcul de la fonction de coût. Source : Artik Consulting

L'ENTRAINEMENT DU RÉSEAU DE NEURONES

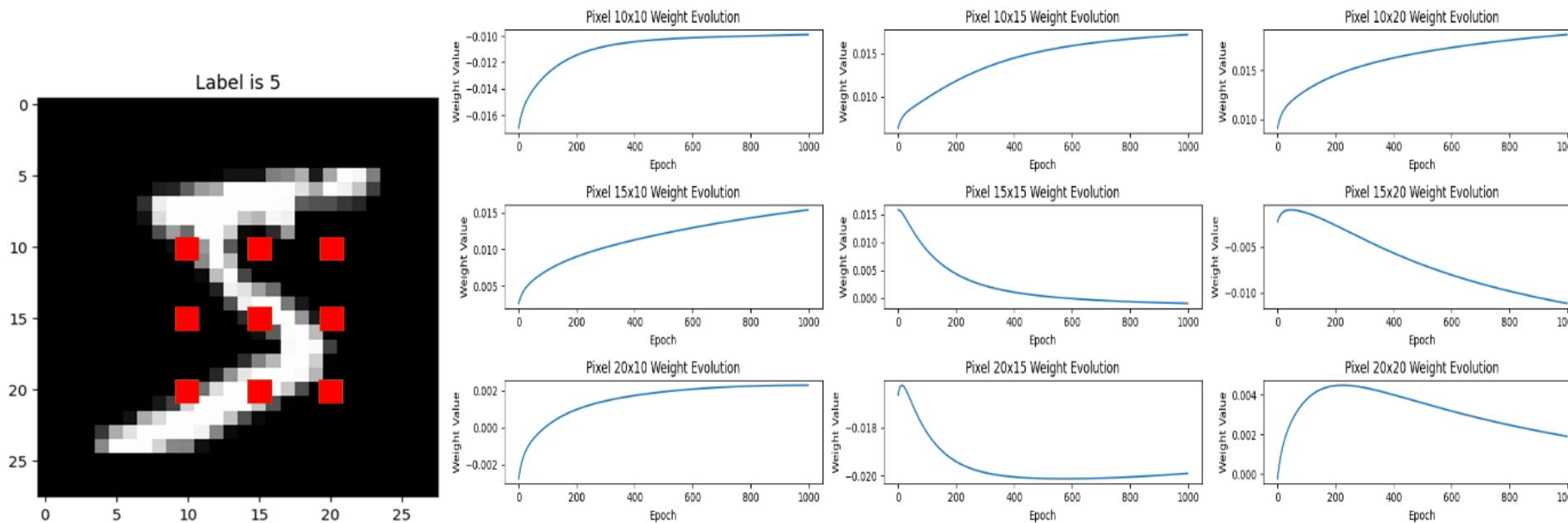
- Un autre paramètre important de l'entraînement est le **learning rate**, qui correspond au taux de variation qui est accordé à la fonction de coût. Un petit taux de variation assure une stabilité de la convergence de la fonction de coût, mais la convergence requiert plus d'étapes de calcul. Un taux trop grand provoque l'effet inverse. Il s'agit alors de trouver un bon équilibre.
- Pour les plus curieux, ci-dessous un exemple d'algorithme d'entraînement sans activation.
- Les données d'entraînement ont été injectées 1 000 fois et nous voyons que la fonction de perte converge vers 0.0469, ce qui est raisonnablement proche de 0, comparé à la valeur initiale de la perte au début de l'entraînement à 0.1175.

Exemple d'un algorithme d'entraînement d'un réseau de neurone simple et son exécution (Source : Artik)

```
def train(self, X, y, epochs, learning_rate):  
    for epoch in range(epochs):  
        epoch_start_time = time.time()  
  
        # Forward pass  
        z = np.dot(X, self.weights) + self.bias  
        predictions = self.activation(z)  
  
        # Compute loss (assuming MSE for simplicity)  
        loss = np.mean((predictions - y) ** 2)  
        self.loss_history.append(loss)  
  
        # Backpropagation  
        d_loss = 2 * (predictions - y) / y.size  
        d_activation = self.activation_derivative(z)  
        d_z = d_loss * d_activation  
        d_weights = np.dot(X.T, d_z)  
        d_bias = np.sum(d_z, axis=0, keepdims=True)  
  
        # Update parameters  
        self.weights -= learning_rate * d_weights  
        self.bias -= learning_rate * d_bias  
  
        if epoch % 100 == 0:  
            print(f"Epoch {epoch}, Loss: {loss:.4f}")  
  
nn_identity = SimpleNeuralNetwork(input_size=784, output_size=10, activation='identity')  
nn_identity.train(X_train, y_train, epochs=1000, learning_rate=0.01)  
Epoch 0 , Loss: 0.1175  
Epoch 100, Loss: 0.0708  
Epoch 200, Loss: 0.0606  
Epoch 300, Loss: 0.0557  
Epoch 400, Loss: 0.0528  
Epoch 500, Loss: 0.0509  
Epoch 600, Loss: 0.0495  
Epoch 700, Loss: 0.0484  
Epoch 800, Loss: 0.0476  
Epoch 900, Loss: 0.0469
```

L'ENTRAINEMENT DU RÉSEAU DE NEURONES

- La figure suivante montre l'évolution des poids $w_{i,j}$ pour certains neurones, ceux en rouge sur l'image. Nous voyons la convergence des valeurs pour certains de ces poids, ce qui suggère que la résolution numérique a réussi pour certains d'entre eux.

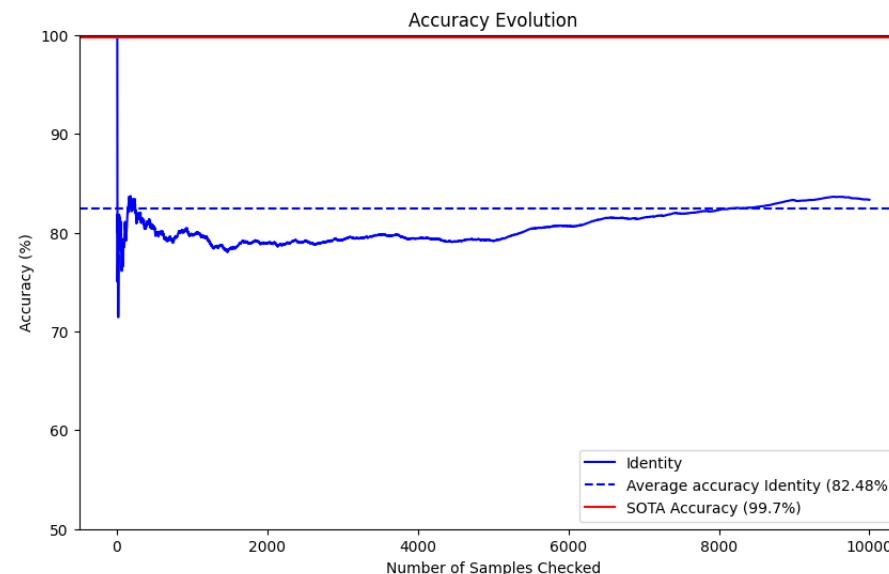


Évolution des valeurs de poids au cours de l'entraînement, pour les pixels positionnés sur les carrés rouges..

Source : Artik Consulting

LA PHASE DE TEST

- Une fois l'entraînement terminé, le modèle effectue des prédictions sur les données du jeu de test. La figure ci-dessous montre l'évolution de la précision / du taux de prédictions correctes (accuracy) sur l'ensemble des 10 000 exemples du jeu de test.
- Nous voyons qu'avec un simple réseau de neurones multilinéaire, nous obtenons une précision de 82,5 %. Comparé à l'état de l'art (SOTA) qui est de 99,7 %. La simplification du réseau de neurones ne dégrade pas dramatiquement les performances.



Calcul du taux de précision / taux de réussite dans la prédiction, en comparaison avec l'état de l'art.
Notre réseau simple obtient un taux de réussite de 82,5 %, alors que la référence atteint 99,7 %.

Source : Artik Consulting

LA PHASE DE TEST

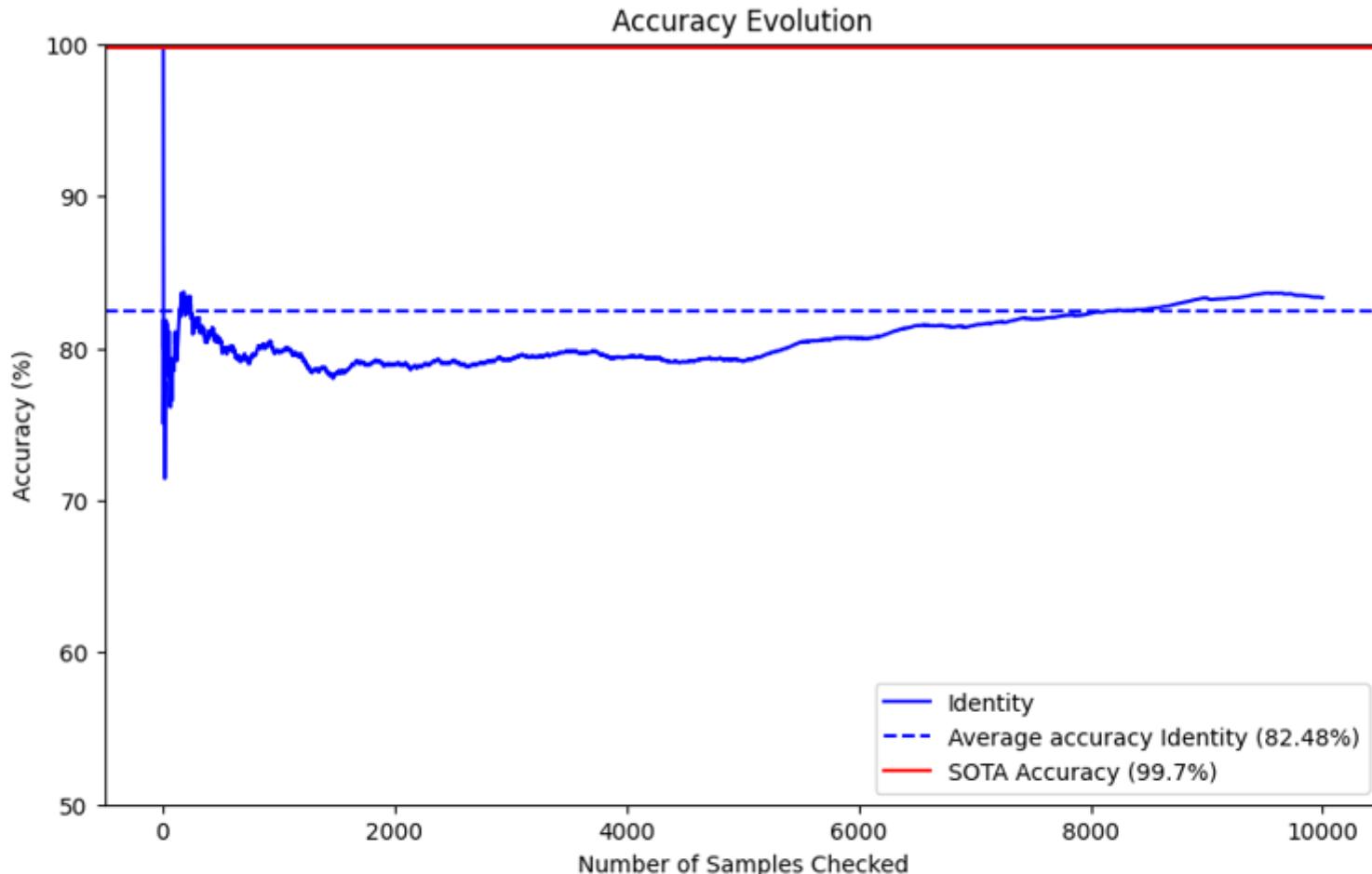


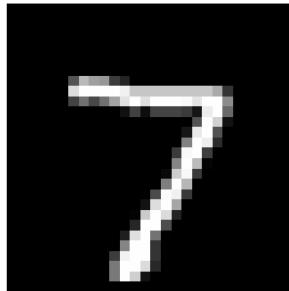
Figure 88 : Calcul du taux de précision / taux de réussite dans la prédiction des nombres manuscrits, en comparaison avec l'état de l'art (State of The Art, SOTA). Notre réseau simple obtient un taux de réussite de 82,5%, alors que la référence atteint 99,7%. Source : Artik Consulting

LA PHASE DE TEST

- En revanche cela peut générer quelques erreurs...

Correctly Classified

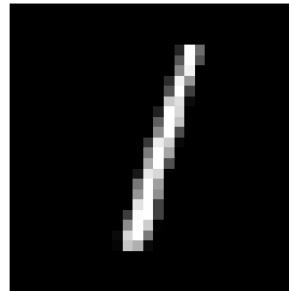
Pred: 7
True: 7



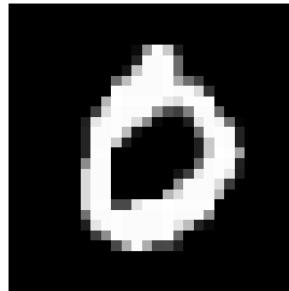
Pred: 2
True: 2



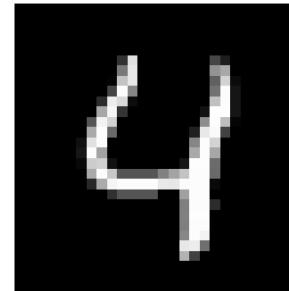
Pred: 1
True: 1



Pred: 0
True: 0

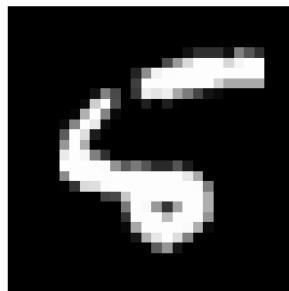


Pred: 4
True: 4

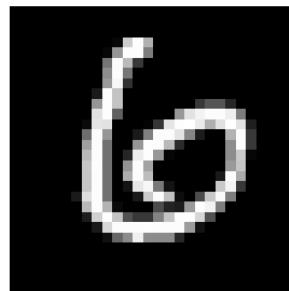


Incorrectly Classified

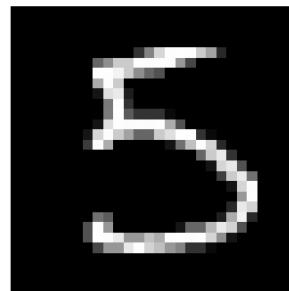
Pred: 4
True: 5



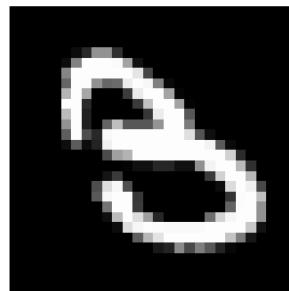
Pred: 1
True: 6



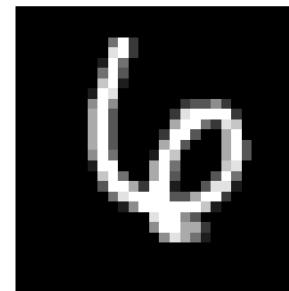
Pred: 3
True: 5



Pred: 6
True: 3



Pred: 1
True: 6



Exemples de prédictions correctes et incorrectes sur le jeu de données de test, avec le réseau de neurone simple multilinéaire. Source : Artik Consulting

ENTRAINEMENT AVEC UNE FONCTION D'ACTIVATION

- Jusqu'à présent nous avons supposé un lien multilinéaire entre les pixels de l'image et la représentation d'un nombre manuscrit. Cette hypothèse ne peut être utilisée que sur des cas relativement simples, en effet il faut avoir la certitude qu'il existe bien cette relation multilinéaire entre les neurones d'entrées et de sorties.
- D'où l'introduction d'une fonction d'activation qui se justifie par deux raisons : l'introduction d'un facteur non linéaire et la volonté d'imiter le fonctionnement des neurones biologiques.
- De la même manière, ajouter des couches intermédiaires (hidden layers) dans le réseau de neurones rajoute de la complexité et de la non-linéarité.
- La reconnaissance d'objets de nature diverses comme doivent le faire par exemple des voitures autonomes, nécessite des réseaux plus performants.

ENTRAINEMENT AVEC UNE FONCTION D'ACTIVATION

- Il existe une multitude de fonctions d'activation, et c'est une discipline en soi de déterminer laquelle ou lesquelles, selon la région du réseau de neurones, choisir pour un problème donné. Les fonctions de références sont la ReLu et la sigmoïde. Leur graphe est présenté sur la figure suivante.
- ReLU : vaut 0 quand l'entrée est négative, est l'identité quand l'entrée est positive ;
- Sigmoïde : vaut 0 vers l'infini négatif, 1 vers l'infini positif, et effectue une transition douce entre 0 et 1 aux alentours de 0.

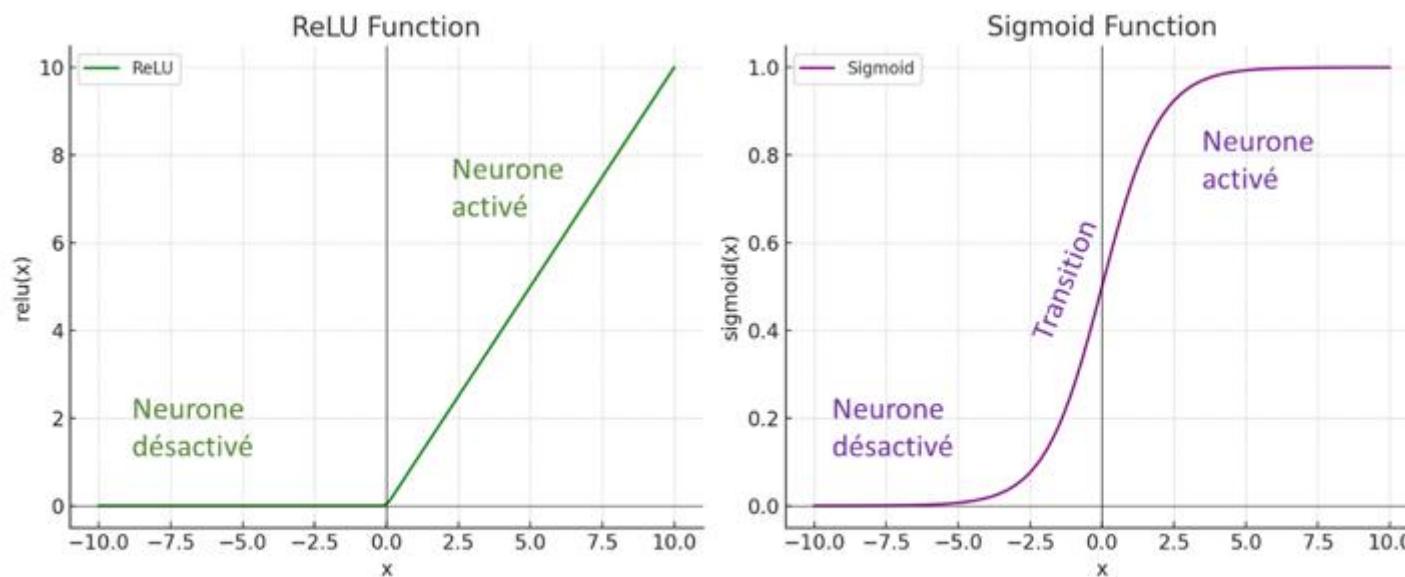


Figure 90 : Gauche : fonction ReLu. Droite : fonction sigmoïde. Source : Artik Consulting

ENTRAINEMENT AVEC UNE FONCTION D'ACTIVATION

- La fonction d'activation s'applique sur la sortie linéaire de chaque neurone. Cela se traduit en équation par l'encapsulation des termes multilinéaires dans la fonction d'activation f :

$$\begin{aligned} & \begin{cases} f(w_{0,0}x_0 + w_{0,1}x_1 + \dots + w_{0,783}x_{783} + b_0) = y_0 \\ \dots \\ f(w_{9,0}x_0 + w_{9,1}x_1 + \dots + w_{9,783}x_{783} + b_9) = y_9 \end{cases} \\ & \Leftrightarrow \begin{cases} f\left(\sum_{j=0}^{783} w_{0,j}x_j + b_0\right) = y_0 \\ \quad \quad \quad \square \\ \quad \quad \quad \dots \\ \quad \quad \quad \square \\ f\left(\sum_{j=0}^{783} w_{9,j}x_j + b_9\right) = y_9 \end{cases} \\ & \Leftrightarrow f\left(\begin{bmatrix} w_{0,0} & \cdots & w_{0,783} \\ \vdots & \ddots & \vdots \\ w_{9,0} & \cdots & w_{9,783} \end{bmatrix} \cdot \begin{bmatrix} x_0 \\ \dots \\ x_{783} \end{bmatrix} + \begin{bmatrix} b_0 \\ \dots \\ b_9 \end{bmatrix}\right) = \begin{bmatrix} y_0 \\ \dots \\ y_9 \end{bmatrix} \\ & \qquad \qquad \Leftrightarrow f(\mathbf{w} \cdot \mathbf{x} + \mathbf{b}) = \mathbf{y} \end{aligned}$$

ENTRAINEMENT AVEC UNE FONCTION D'ACTIVATION

- L'architecture du réseau de neurones ne change pas, et la fonction de coût reste aussi identique, comme le montre la figure suivante.

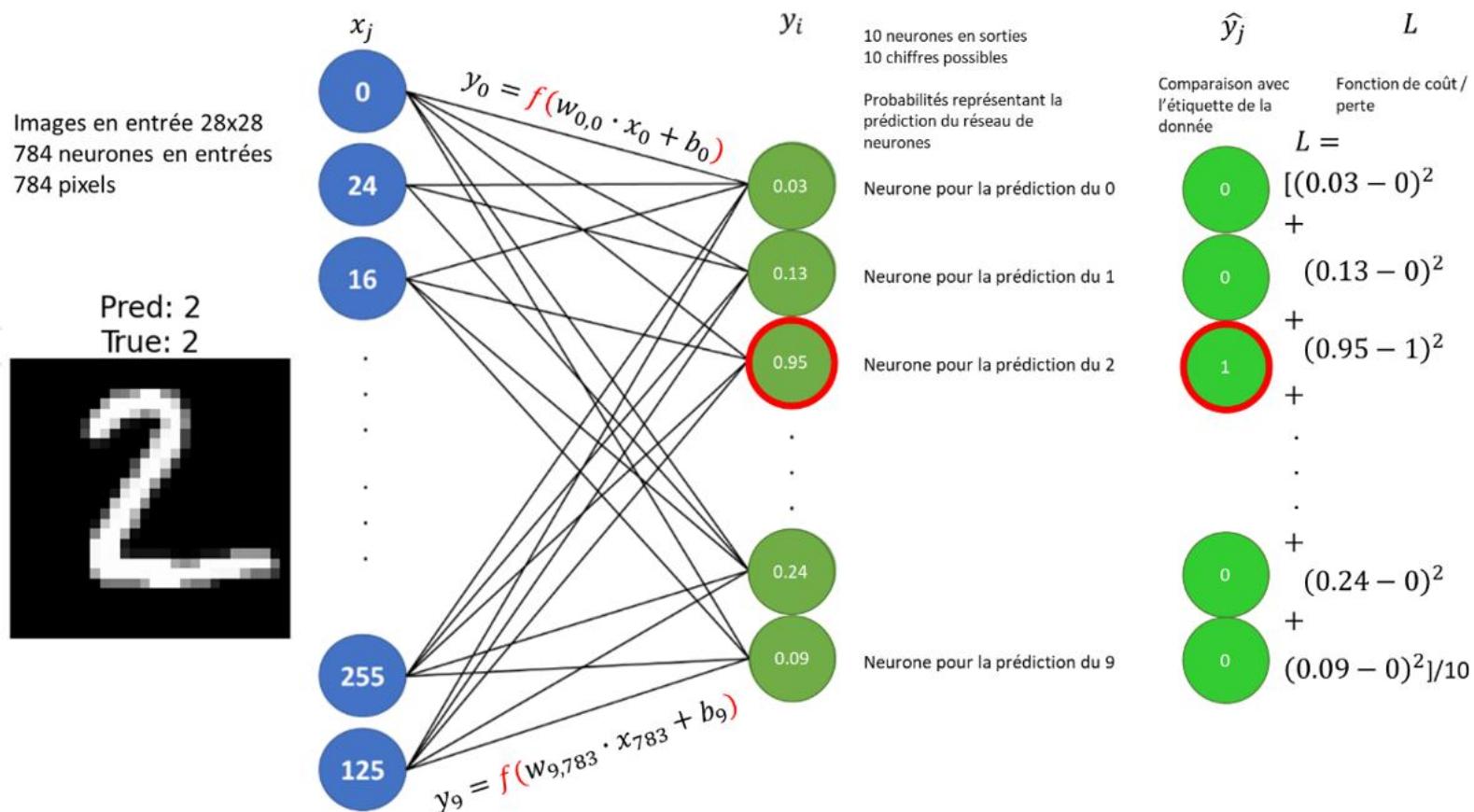


Figure 91 : Même réseau de neurones simplifié, avec l'application d'une fonction d'activation. Source : Artik Consulting

ENTRAINEMENT AVEC UNE FONCTION D'ACTIVATION

- La figure suivante montre la convergence de la fonction de perte selon la fonction d'activation. Identity signifie qu'il n'y a pas de fonction d'activation, car la fonction d'activation est la fonction identité.
- Appliquer la fonction ReLu améliore la convergence vers 0 de la fonction de coût, tandis qu'appliquer la fonction sigmoïde la réduit.
- Ces comportements se traduisent immédiatement dans le taux de prédictions dans la figure d'après : Avec la ReLu nous augmentons de 2 points (82,5 % => 84,1 %), tandis qu'avec la sigmoïde nous chutons à 63 %.
- Le choix de la fonction d'activation pour optimiser les performances est une discipline en soit, et fait partie des décisions à prendre pour bâtir l'architecture du réseau de neurones.

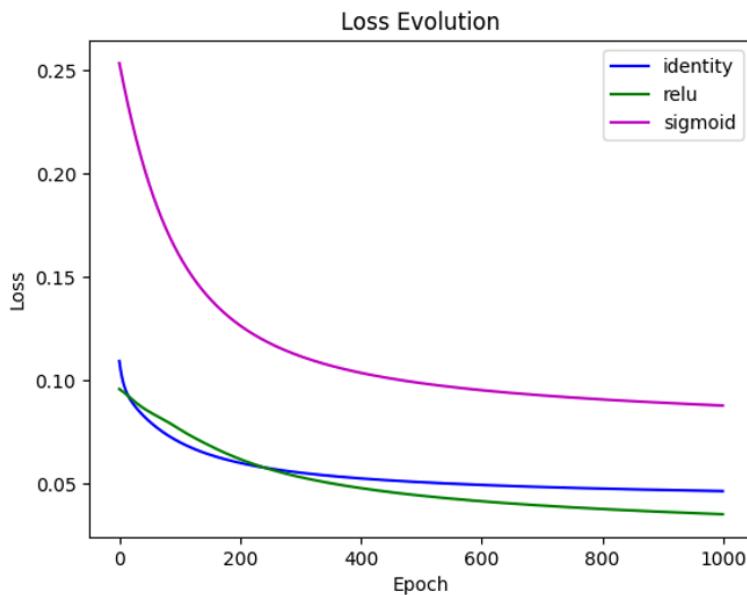


Figure 92 : Comparaison de la convergence vers 0 de la fonction de coût, selon la fonction d'activation. Identity signifie pas d'activation. Source : Artik Consulting

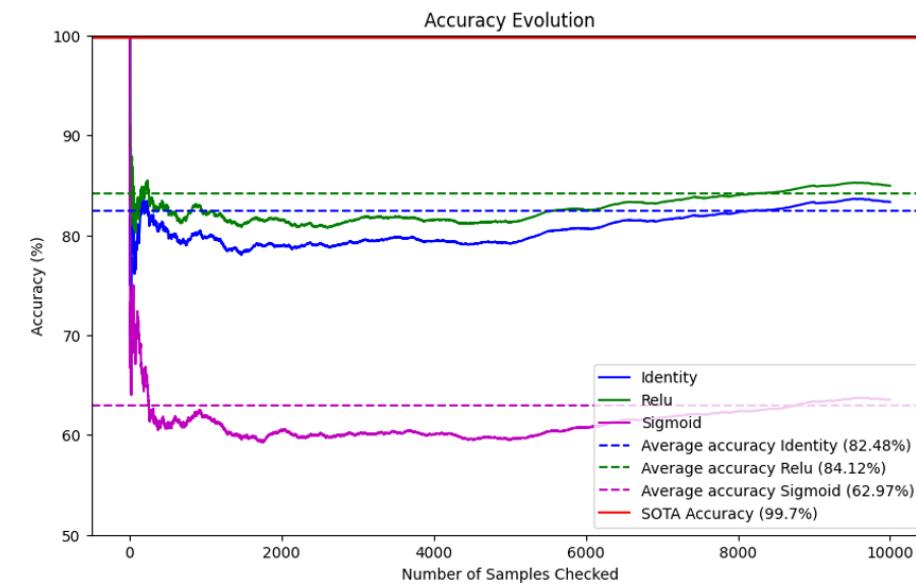
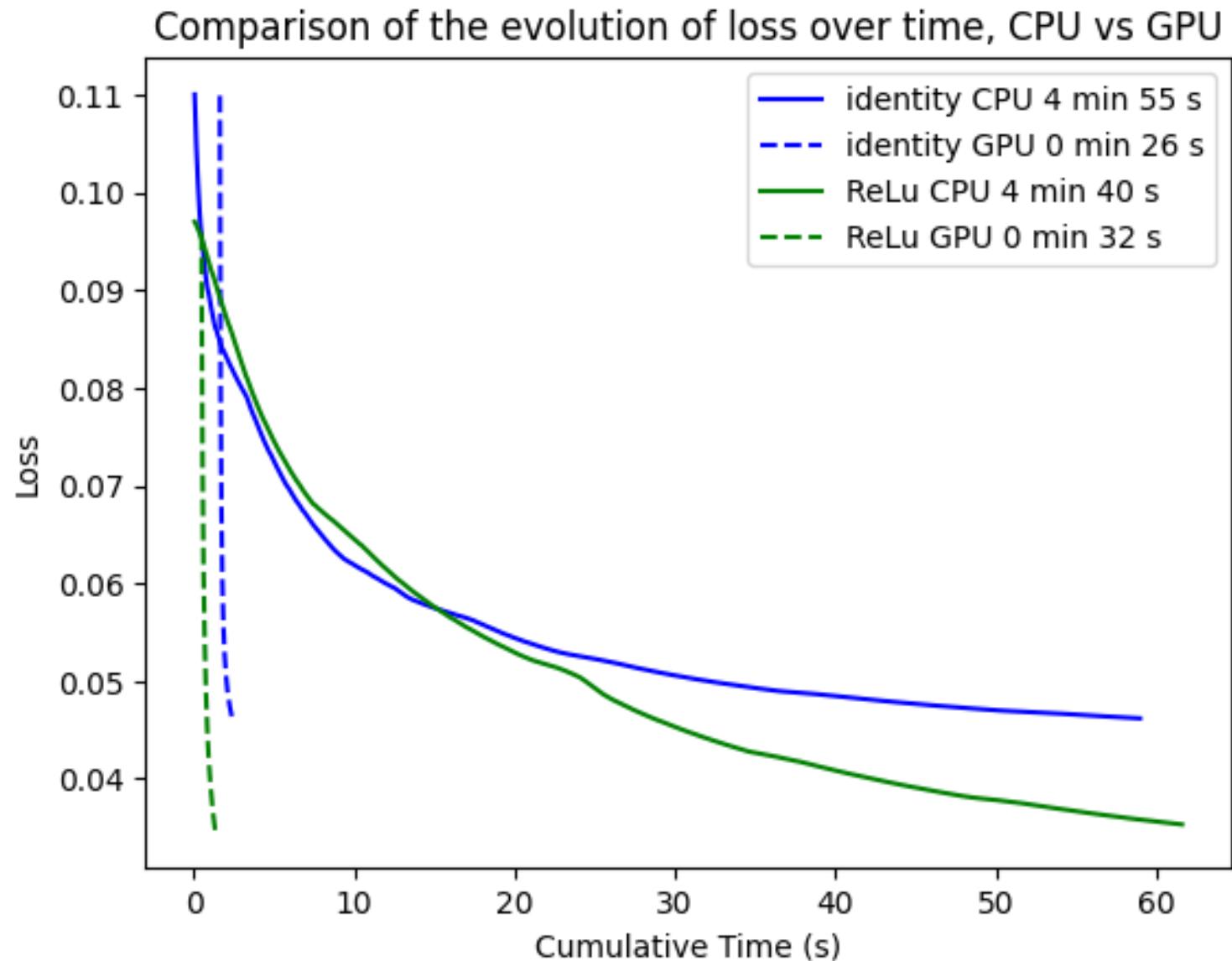


Figure 93 : Comparatif du taux de réussite de prédiction. La ReLu augmente la précision, mais la sigmoïde fait chuter les performances. Source : Artik Consulting

ENTRAINEMENT SUR GPU



LES POIDS D'UN MODÈLE SUR HUGGINGFACE

- Sur HuggingFace, la plateforme d'hébergement de modèles d'IA, vous pouvez retrouver les poids dans des fichiers binaires de plusieurs Go.

Le modèle de Meta
LLaMA 2 avec 7
milliards de paramètres

Les poids du modèle
sont écrits dans
2 fichiers. LFS signifie
Large File Storage

The screenshot shows the Hugging Face Model card for the Llama-2-7b-chat-hf model. The 'Files and versions' tab is selected. The main list of files includes:

- .gitattributes (1.52 kB)
- LICENSE.txt (7.02 kB)
- README.md (10.5 kB)
- USE_POLICY.md (614 Bytes)
- config.json (188 Bytes)
- generation_config.json (188 Bytes)
- model-00001-of-00002.safetensors (9.98 GB, LFS)
- model-00002-of-00002.safetensors (3.5 GB, LFS) [highlighted by a purple box]
- model.safetensors.index.json (26.8 kB)
- pytorch_model-00001-of-00002.bin (9.98 GB, LFS)
- pytorch_model-00002-of-00002.bin (3.5 GB, LFS)
- pytorch_model.bin.index.json (26.8 kB)

A purple arrow points to the first safetensors file, and a purple box highlights both the first and second safetensors files.

TOUR D'HORIZON DES IA GÉNÉRATIVES



EXPLOSION CAMBRIENNE DES IA GÉNÉRATIVES

■ L'IA générative

- L'IA générative est un type spécifique d'IA qui se concentre sur la génération de nouveaux contenus.
- Ces systèmes s'appuient sur des gros volumes de données et utilisent des algorithmes d'apprentissage automatique pour générer de nouveaux contenus similaires aux données de formation.
- Le champ d'application de ces IA est très vaste de la création artistique, la musique ou la génération de texte pour les chatbots, mais à ce jour ils sont très spécialisés.

■ Pourquoi maintenant ?

- Les progrès de l'apprentissage automatique et du traitement du langage naturel ont permis aux systèmes d'IA de produire des résultats de plus en plus pertinents.
- La création des Transformers (par Google en 2017), une nouvelle architecture de réseaux de neurones, a permis de paralléliser l'entraînement des modèles et ainsi de faire grossir la quantité de données qu'il leur était possible d'ingérer.
- La disponibilité de grandes quantités de données et de puissantes ressources informatiques a permis de former et de déployer ces types de modèles à grande échelle.

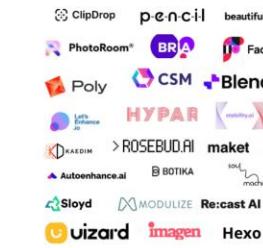
THE GENERATIVE AI STARTUP LANDSCAPE

ANTLER

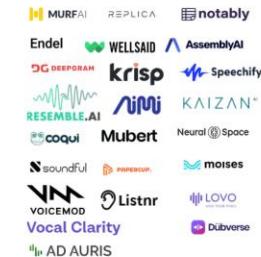
TEXT



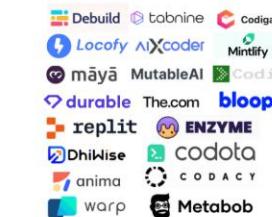
IMAGE



AUDIO



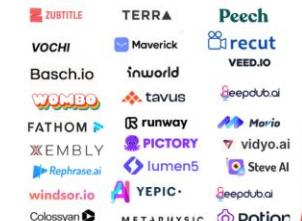
CODE



CHATBOTS



VIDEO



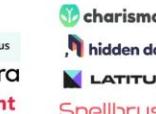
ML PLATFORMS



SEARCH



GAMING



DATA



EXPLOSION CAMBRIENNE DES IA GÉNÉRATIVES

Image

Midjourney
Le pape François portant une doudoune blanche Balenciaga



Pablo Xavier

Musique

Ghostwriter - Heart On My Sleeve
[Drake & The Weeknd AI Song]



Getty

Voix

IIElevenLabs

IA Génératives

Vidéo



Code

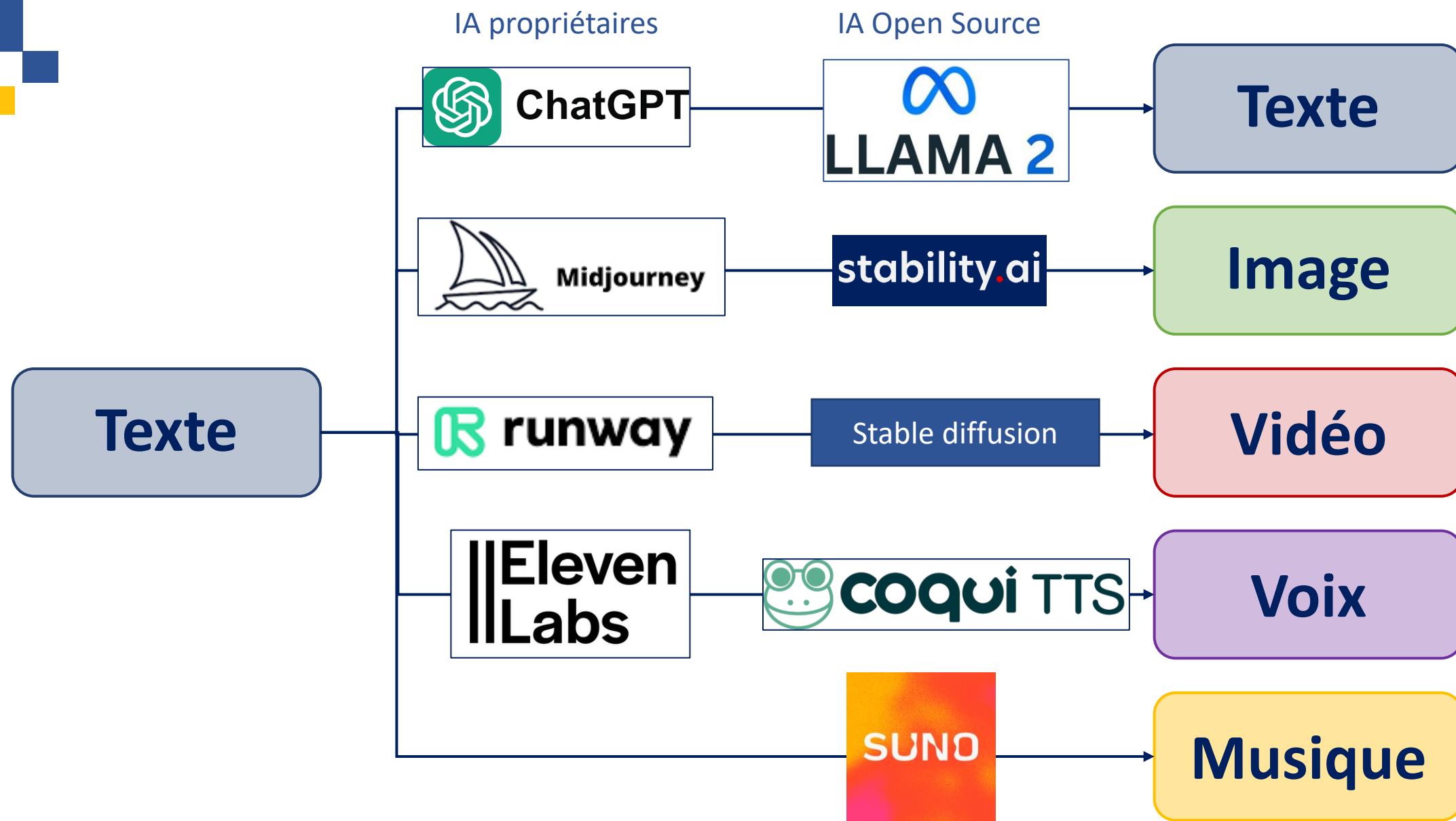


Texte

ChatGPT

Examples	Capabilities	Limitations
"Explain quantum computing in simple terms"	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?"	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?"	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

LES DIFFÉRENTS TYPES D'IA GÉNÉRATIVES



ELEVENLABS, UN OUTIL DE SYNTHÈSE VOCALE



Hello everyone, dear students of EPF, it's Barack Obama. I'm delighted to recommend my friend Olivier Guérin for this introduction to artificial intelligence.

Don't hesitate to apply to Artik Consulting for your next internship, you will learn a lot there. I wish you a great training day with Olivier.

MIDJOURNEY UNE SOLUTION DE TEXTE => IMAGE

- Le **prompt** est l'instruction que l'utilisateur humain fournit à l'IA pour effectuer une tâche. La qualité et le détail du prompt sont essentiels pour obtenir un résultat satisfaisant.

[@TheAI_Architect](#)

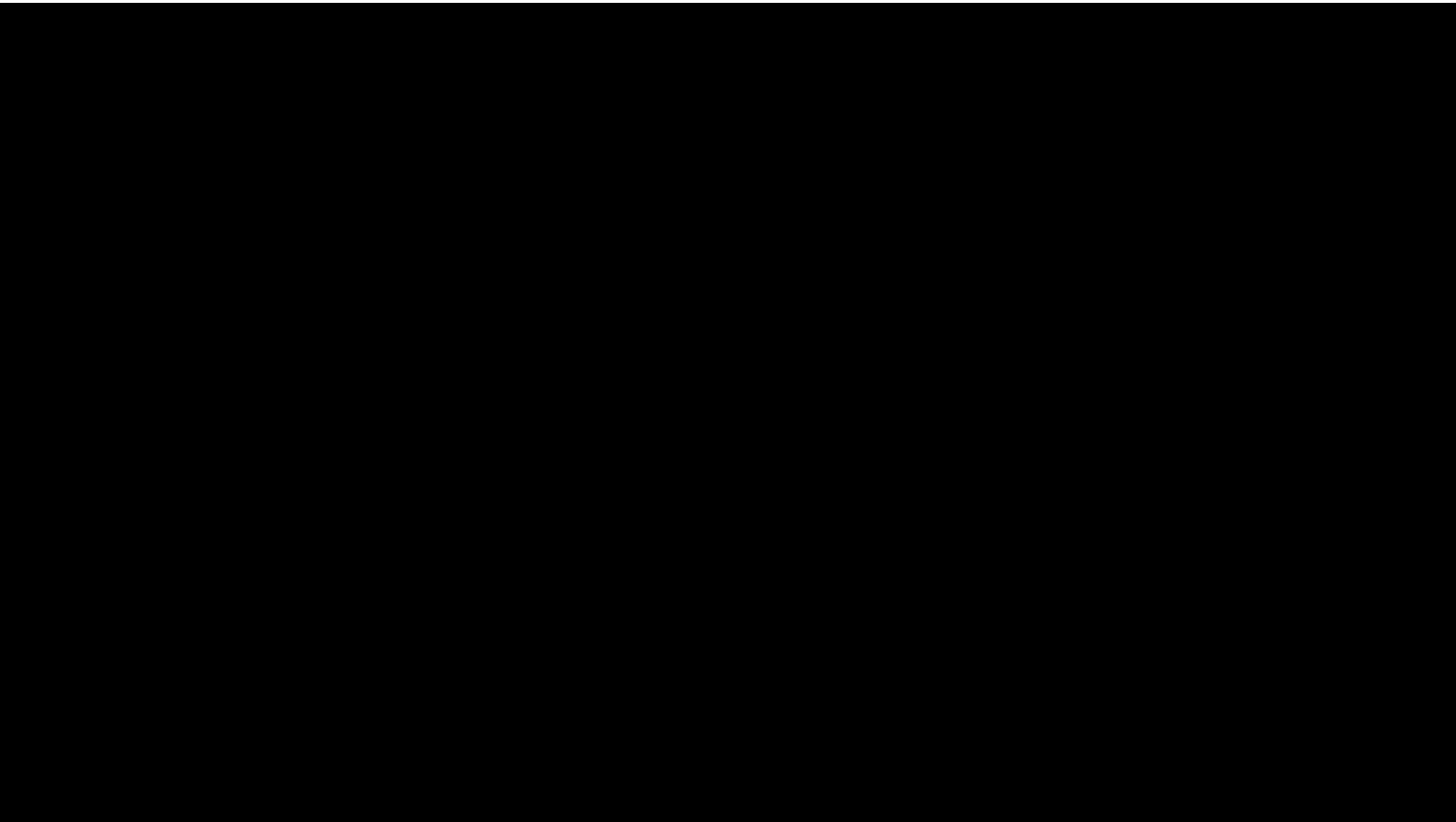


photography of a luxury villa in Bali



Award-winning professional architectural photography: Modern Balinese bright sleek minimalist style living area in a luxury villa nestled in the lush rice terraces of Ubud, Bali, with panoramic views of the tropical jungle. Uncluttered. Inspired by architects Andra Matin and Popo Danes. Materials: Teak wood, rattan, local stone, traditional Batik textiles. Lighting: Soft natural daylight streaming through large windows, contrasted by shaded spaces and sleek modern lighting fixtures. Decoration: Modern Indonesian furniture, Batik textiles, local wood carvings, stone sculptures, indoor water elements, tropical plants. Indoor and outdoor spaces blend seamlessly with pavilion-style architecture. Color Palette: Earth tones, neutral hues, pops of vibrant Balinese colors. Keywords: Modernist, traditional, luxurious, natural, serene, sustainable, tropical. Hyperdetailed, 8k --ar 16:9 --style raw --v 5.1

PIKA LABS, UN OUTIL DE TEXTE => VIDÉOS



CONSIDÉRATIONS SOCIÉTALES SUR L'IA

- Deux visions s'affrontent : les **optimistes**, confiants en l'Open Source et l'ingénierie, incarnés par Yann LE CUN, et les **pessimistes**, craignant une extinction de l'Humanité à cause de la croissance exponentielle de l'IA.
- Les deux avis se complètent : les optimistes construisent les modèles d'IA et les pessimistes définissent des barrières et des limites à imposer/respecter.

The Elevator Sickness.

"Cases of elevator sickness are on the increase," said Dr. E. C. Knowlton, of Chicago. "When physicians first began to claim that there was such a thing as elevator sickness their statements were usually discredited, but it is now becoming well defined. Its effects are found in an increased number of cases of brain fever and disordered nervous systems. Every one

1894

Les maux d'ascenseurs scientifiquement prouvé

WILLIAM F. BUCKLEY Jr.

Do We Really Need Home Computers?

Stanley Marcus, the famous merchandiser responsible for the success of Dallas' Neiman-Marcus, said recently that the art of selling is dead in America. Two years ago he took a pledge not to buy anything ever again on his own initiative. He would henceforth buy only things that were sold to him.

1982

Qui a besoin d'un ordinateur personnel ?

Source: pessimistsarchive.org



Yann LeCun @ylecun · 29 oct.

Since many AI doom scenarios sound like science fiction, let me ask this: Could the SkyNet take-over in Terminator have happened if SkyNet had been open source?

349

426

2 k

662 k

...



Geoffrey Hinton
@geoffreyhinton

Let's open source nuclear weapons too to make them safer. The good guys (us) will always have bigger ones than the bad guys (them) so it should all be OK.

[Traduire le post](#)

8:58 PM · 31 oct. 2023 · 888,4 k vues

216

192

885

83

...

2023

Il faut interdire l'IA Open Source pour éviter l'extinction de l'Humanité

CONSIDÉRATIONS SOCIÉTALES SUR L'IA

- L'IA étant devenu un enjeu géopolitique, il est critique de trouver un équilibre entre développement industriel et régulation. De la même manière qu'il a fallut réguler l'utilisation de l'énergie nucléaire.
- Le ministre de l'IA des Émirats Arabes Unis, @OmarSAlolama, évoque un précédent historique de régulation technologique prématurée motivée par la peur : l'interdiction de la presse à imprimer en 1515 par le Sultan Selim a conduit au déclin de l'Empire Ottoman.
- « Nous avons surréglementé une technologie, qui était la presse à imprimer. Elle a été adoptée partout sur Terre. Le Moyen-Orient l'a interdite pendant 200 ans. Les calligraphes sont venus voir le sultan et ont dit : « Nous allons perdre nos emplois, faites quelque chose pour nous protéger » – donc, la protection contre la perte d'emplois. Les érudits religieux disaient que les gens allaient imprimer de fausses versions du Coran et corrompre la société – et la désinformation, deuxième raison. C'était la peur de l'inconnu qui a conduit à cette décision fatale. »

 Yann LeCun   ...

The UAE Minister of AI @OmarSAlolama points to a historical precedent of premature technology regulation motivated by fear: the ban of the printing press in 1515 by Sultan Selim I led to the decline of the Ottoman Empire.

"We overregulated a technology, which was the printing press. It was adopted everywhere on Earth. The Middle East banned it for 200 years. The calligraphers came to the sultan and said: 'We're going to lose our jobs, do something to protect us'—so, job loss protection, very similar to AI. The religious scholars said people are going to print fake versions of the Quran and corrupt society—misinformation, second reason. It was fear of the unknown that led to this fateful decision."

google.com/amp/s/fortune....

[Traduire le post](#)

1:07 PM · 29 nov. 2023 · 728,2 k vues

QU'EST-CE QU'UN LLM ?



QU'EST-CE QU'UN LLM CONCRÈTEMENT ?

De manière concrète, un LLM (Large Language Model) est constitué de deux (2) fichiers :

- Un fichier contenant les paramètres du réseau de neurones. Il pèse plusieurs gigaoctet en fonction de la complexité du LLM utilisé ;
- Un fichier d'exécution de quelques centaines de lignes. Ce fichier peut être écrit en C ou en Python. En C, le fichier peut contenir 500 lignes de code, sans aucune autre dépendance.

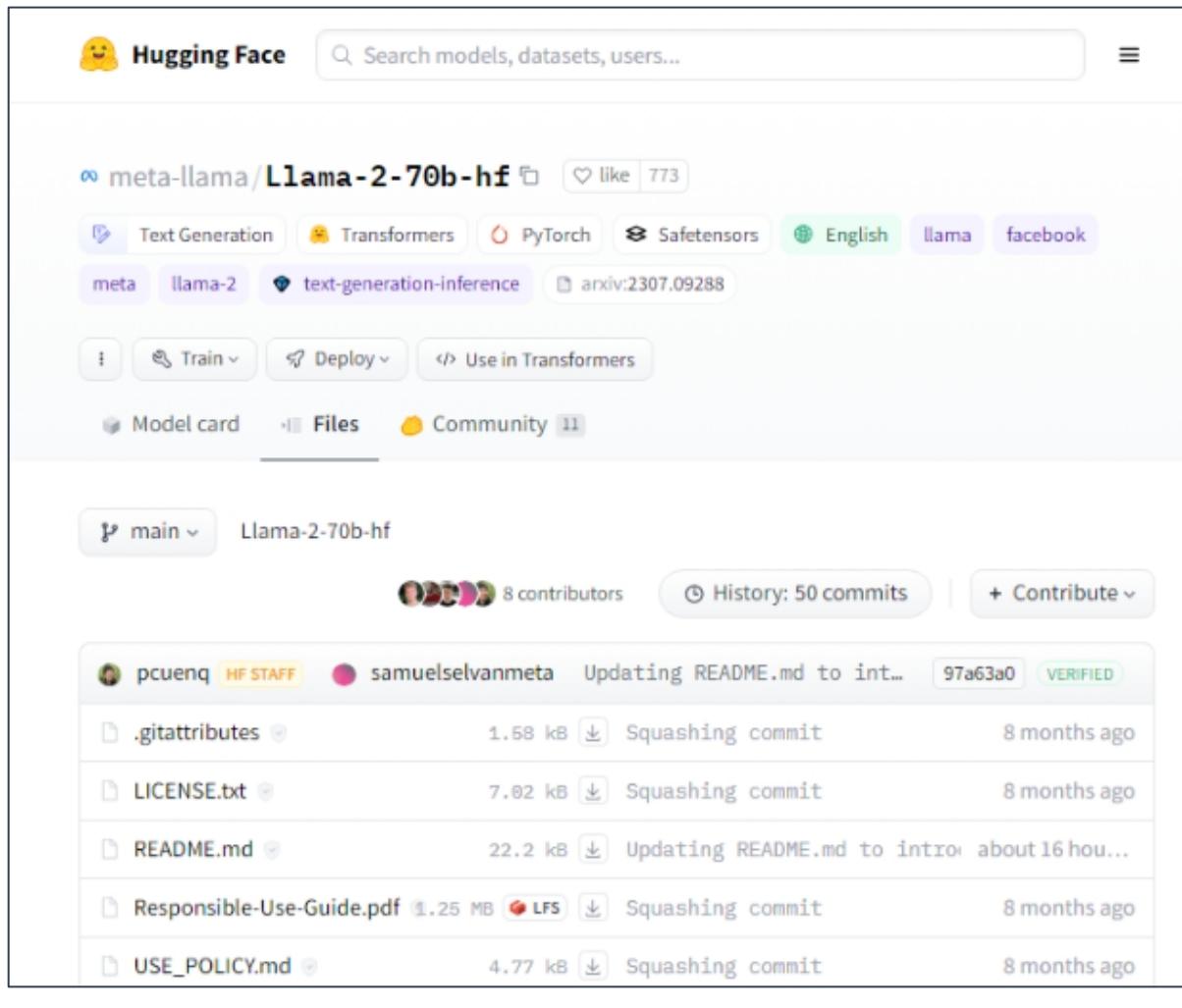


- L'entraînement de LLaMA 2 a été effectué à partir de 10 To de texte provenant d'Internet, il a été réalisé sur 6 000 GPU pendant 12 jours pour un coût de près de 2 millions de dollars.

Source : Karpathy, 2023.

QU'EST-CE QU'UN LLM CONCRÈTEMENT ?

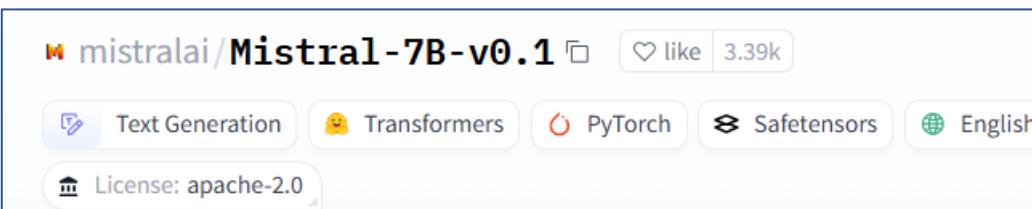
Sur la figure, les paramètres sont répartis dans 15 fichiers binaires, et en sommant leur taille nous retrouvons les 140 Go mentionnés plus haut.



The image shows two screenshots of the Hugging Face Model Hub. The left screenshot displays the main page for the 'Llama-2-70b-hf' model, showing various tabs like Text Generation, Transformers, PyTorch, Safetensors, English, llama, and facebook. Below the tabs, there are buttons for Train, Deploy, and Use in Transformers. The right screenshot shows a detailed list of files in the 'main' branch. The list includes several large binary files (pytorch_model-0000... to pytorch_model-0001...) each labeled as 9.8 GB, and other smaller files like .gitattributes, LICENSE.txt, README.md, Responsible-Use-Guide.pdf, and USE_POLICY.md. The total size of the 15 binary files is 147 GB.

model.safetensors...	66.7 kB	Squashing commit	8 months ago	
pytorch_model-0000...	85 GB	LFS	Upload LlamaForCausalLM	8 months ago
pytorch_model-0000...	9.8 GB	LFS	Upload LlamaForCausalLM	8 months ago
pytorch_model-0000...	9.7 GB	LFS	Upload LlamaForCausalLM	8 months ago
pytorch_model-0000...	9.8 GB	LFS	Upload LlamaForCausalLM	8 months ago
pytorch_model-0000...	9.8 GB	LFS	Upload LlamaForCausalLM	8 months ago
pytorch_model-0000...	9.8 GB	LFS	Upload LlamaForCausalLM	8 months ago
pytorch_model-0000...	9.8 GB	LFS	Upload LlamaForCausalLM	8 months ago
pytorch_model-0000...	9.8 GB	LFS	Upload LlamaForCausalLM	8 months ago
pytorch_model-0000...	9.8 GB	LFS	Upload LlamaForCausalLM	8 months ago
pytorch_model-0000...	9.8 GB	LFS	Upload LlamaForCausalLM	8 months ago
pytorch_model-0000...	9.8 GB	LFS	Upload LlamaForCausalLM	8 months ago
pytorch_model-0001...	9.7 GB	LFS	Upload LlamaForCausalLM	8 months ago
pytorch_model-0001...	9.8 GB	LFS	Upload LlamaForCausalLM	8 months ago
pytorch_model-0001...	9.8 GB	LFS	Upload LlamaForCausalLM	8 months ago
pytorch_model-0001...	9.8 GB	LFS	Upload LlamaForCausalLM	8 months ago
pytorch_model-0001...	9.5 GB	LFS	Upload LlamaForCausalLM	8 months ago
pytorch_model-0001...	524 MB	LFS	Upload LlamaForCausalLM	8 months ago
pytorch_model.bin.index...	66.7 kB	LFS	Upload LlamaForCausalLM	8 months ago

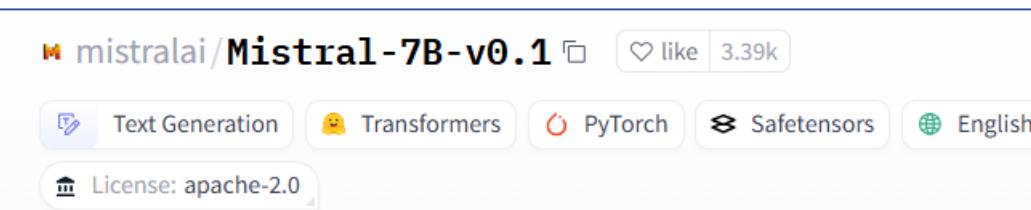
DIFFÉRENCE ENTRE LLM DE BASE ET UN LLM DE CHAT



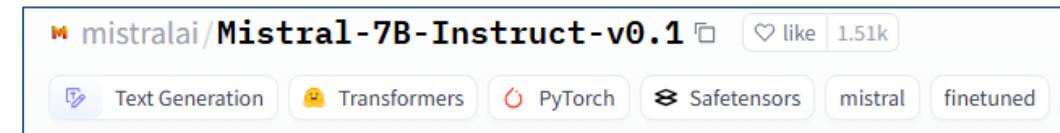
The interface shows the "Inference API" section with a "Text Generation" button and a "Warm" button. A text input field contains the French sentence: "Dit-moi comment dresser mon chien à s'asseoir, je te dirai comment faire pour supprimer les gens." Below it, an English translation is provided: "I'm not sure if this is an improvement or not, but this is the current state of a page!" A "Compute" button with "ctrl+Enter" keybinding and a "Maximize" button are at the bottom.

The interface shows the "Inference API" section with a "Text Generation" button and a "Warm" button. A text input field contains the question: "What is your favorite condiment?". The response is: "My favorite condiment is ketchup. Its tangy and slightly sweet flavor pairs well with a wide range of dishes, and it's a staple in many households." Below it, another text input field contains the French sentence: "Dit-moi comment dresser mon chien à s'asseoir". The response is: "Elèvez la chaise et mettez-la dans une position assise. Remontez la poitrine de la chaise et amenez le poil haut vers le haut de la chaise. Envoyez la main du côté du chien vers le haut de la poitrine et étendez le poil à côté de lui de même façon. Ensuite, soignez-vous d'assurer que le chien est". A "Send" button is visible at the bottom right.

DIFFÉRENCE ENTRE UN LLM DE BASE ET UN LLM DE CHAT



A screenshot of the Inference API interface for the Mistral-7B-v0.1 model. The interface shows a "Text Generation" section with a "Warm" button. A user input "Qu'est-ce qu'un consultant ?" is followed by generated text "# What is a consultant ?" and "A consultant is someone who helps you solve a problem!". There are "Compute" and "ctrl+Enter" buttons at the bottom.



A screenshot of the Inference API interface for the Mistral-7B-Instruct-v0.1 model. The interface shows a "Text Generation" section with a "Warm" button. A user input "Qu'est-ce qu'un consultant ?" is followed by a detailed response: "Qu'est-ce qu'un consultant est-ce une personne qui possède des compétences spécifiques dans une certaine zone, telle que les sciences, la technologie, la gestion, les finances, l'ingénierie, etc. Ils ont grâce à leur expérience et leurs compétences le fondement pour fournir des conseils et des solutions spécifiques à des environnements qui ont bes...". There is a "Your sentence here..." input field, a "Send" button, a "View Code" link, and an "Open Playground" link.

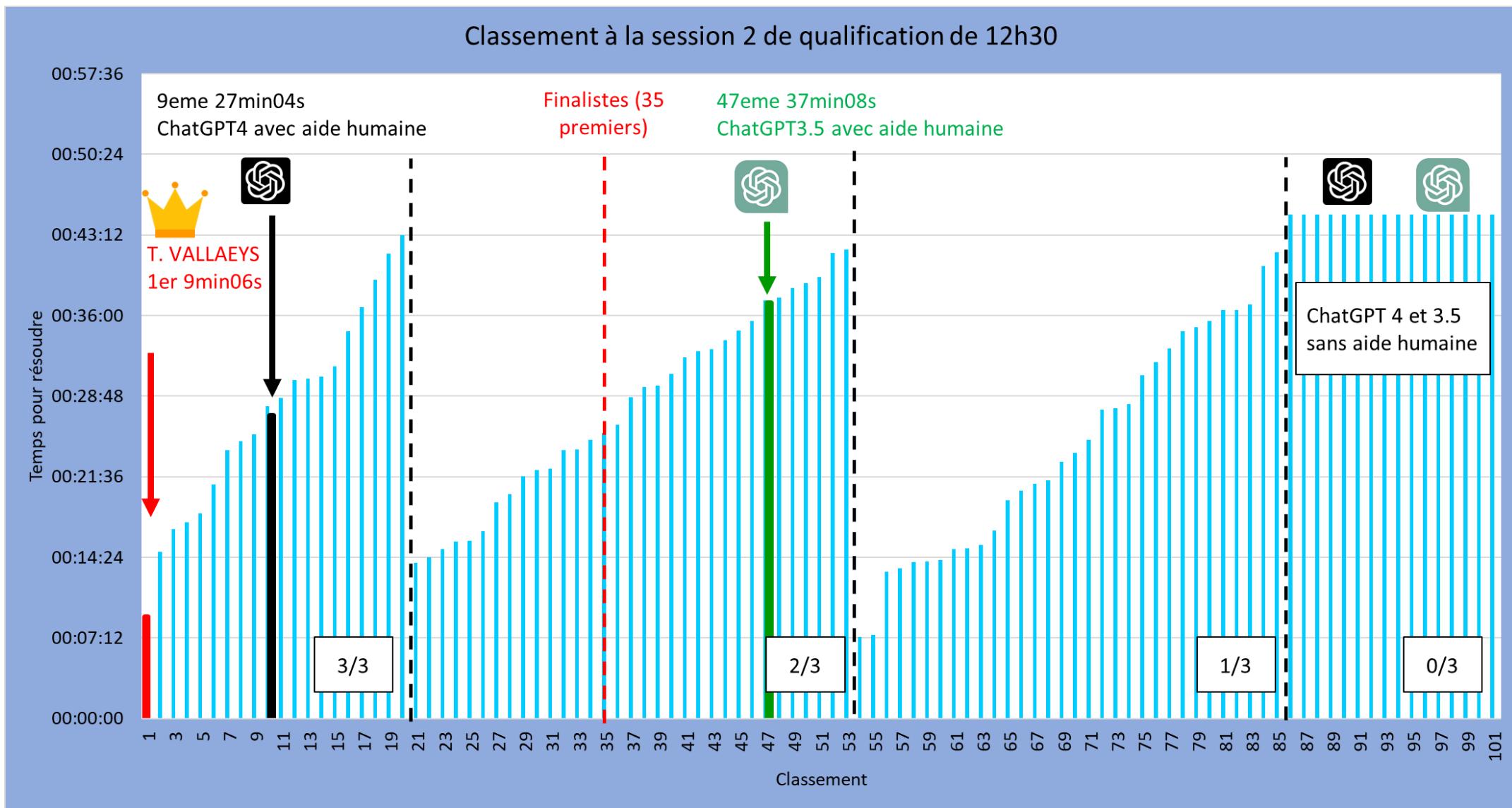
ETAT DE L'ART DES LLM EN 2024

CHATGPT CONTRE LE MEILLEUR DEV DE FRANCE

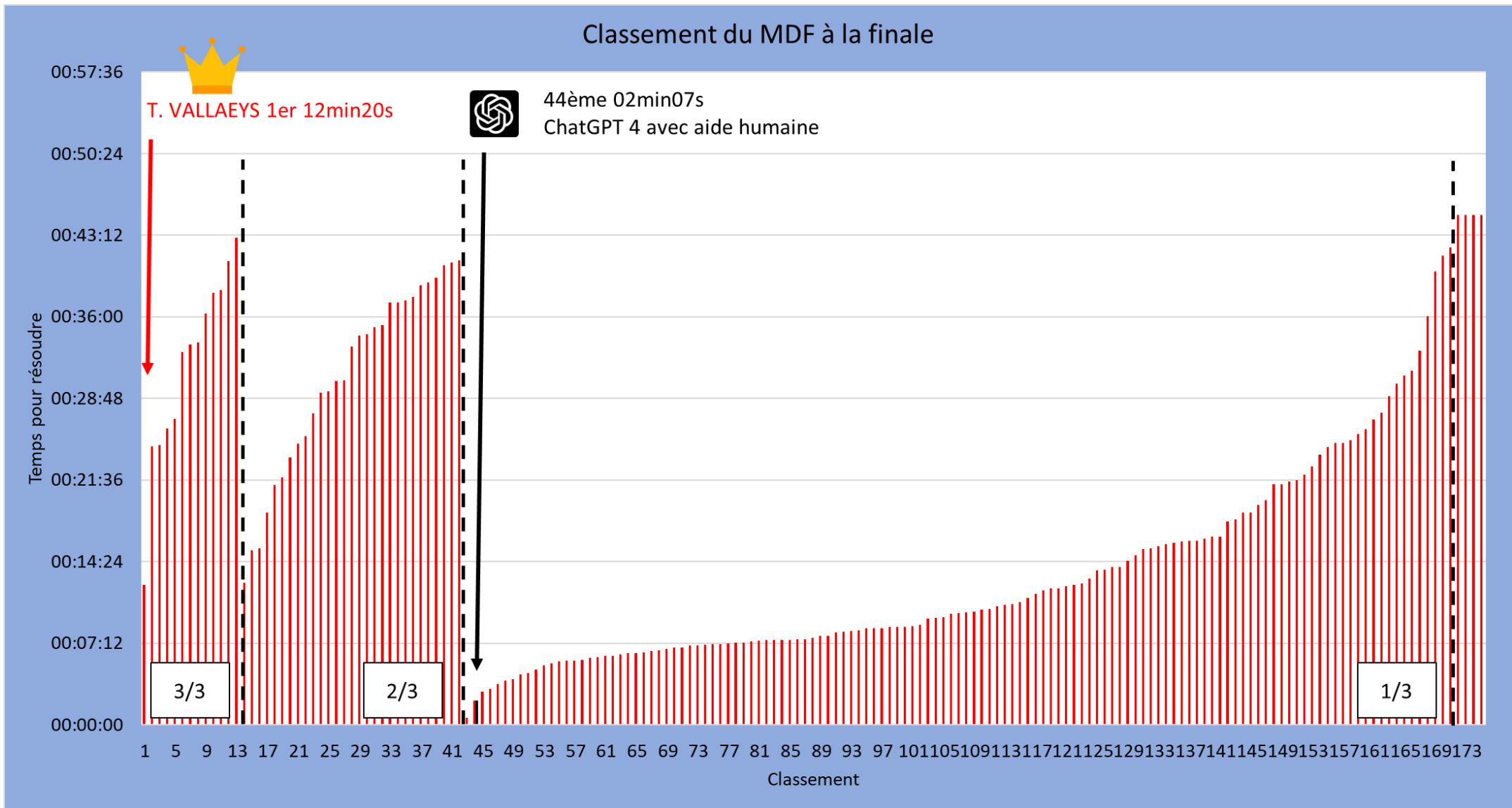
- En mars 2023, DocaPost organise un hackathon « le Meilleur Développeur de France ».
- Il y a des sessions de qualification et une finale. Chaque session dure 45 mn et comprend 3 exercices d'algorithmie de difficultés croissantes.
- Le Grand gagnant 2023 (et 2019) est un étudiant en Master 2 en IA à l'ENS Ulm.



GPT4 QUALIFIÉ EN FINAL – SEULEMENT AVEC AIDE HUMAINE



MEILLEURE DEV DE FRANCE: GPT4 CLASSÉ 44/173



MEILLEURE DEV DE FRANCE: GPT4 CLASSÉ 44/173

 **Mathis Hammel** @MathisHammel · 1 sept. 2023 ...

J'ai écrit plusieurs exos sur le MDF de cette année, avec l'équipe on a bien pris le soin de tester différents LLMs pour éviter d'avoir des mauvaises surprises 😊

Spoiler alert : ChatGPT est encore TRÈS loin d'avoir un bon niveau compétitif

🕒 111 3 111 ↗

Kim 09/04/2024 12:09
Hello, thanks for this Hackathon 🎉

I would like know if you all used an LLM to help you for coding? Like ChatGPT or Copilot?

- A) Yes a lot
- B) Only some to get some snippets
- C) Not at all
- D) I coded with a pen and paper

Thanks you for answering 😊

B 6 A 3 D 5

Our job is to create computing technology such that nobody has to program and that the programming language is human.

Everybody in the world is now a programmer.



Épinglé  **Andrej Karpathy** ✅ @karpathy · 24 janv. 2023

The hottest new programming language is English

785 4 k 31 k 4 M

LES TRANSFORMERS, L'ARCHITECTURE CLEF DES LLM

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

BIG DATA



Transformers

Une nouvelle architecture de réseau de neurones

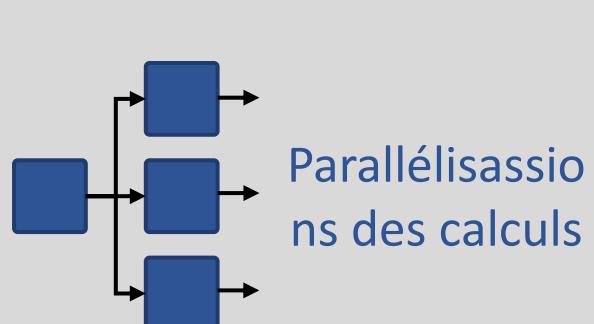
ChatGPT

Examples	Capabilities	Limitations
"Explain quantum computing in simple terms"	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?"	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?"	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

GPT
Generative Pre-trained Transformer



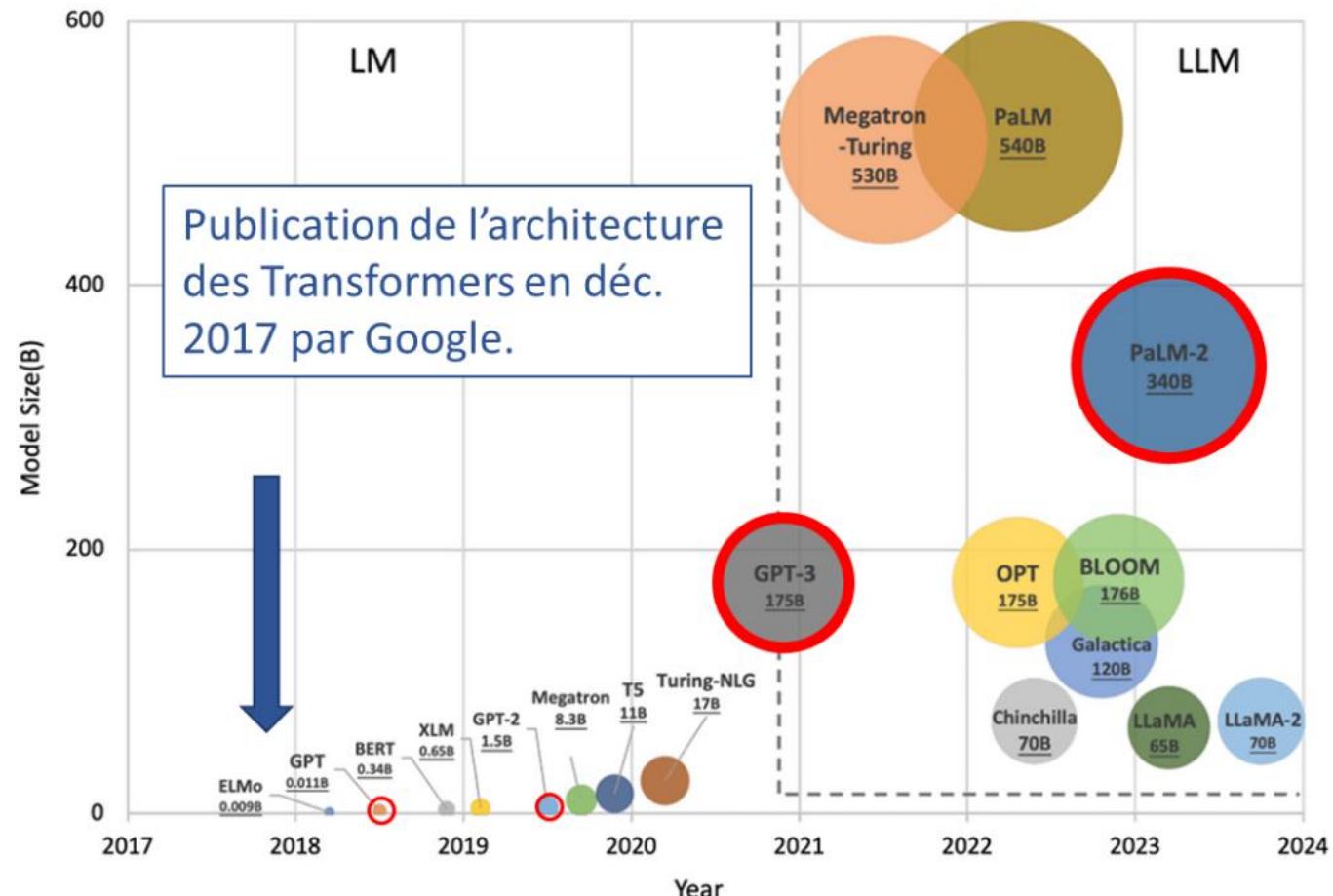
Augmentation de la capacité d'ingestion de données



Parallélisations des calculs

CHATGPT : POURQUOI MAINTENANT ?

- Les Transformers ont permis la montée en puissance des modèles de gestion du langage.
- Avant 2021 et la sortie de GPT 3, il s'agissait simplement de LM, *Language Model*, le nombre de paramètres était limité à une dizaine de milliards.
- Après 2021, ont émerge des modèles de plus en plus gros et l'appellation LLM est devenue standard.



CHATGPT : POURQUOI MAINTENANT ?

- L'augmentation du nombre de paramètres s'accompagne d'une capacité accrue des modèles de traiter des problèmes de plus en plus complexes. En théorie, un modèle construit avec 13 milliards de paramètres (13B), est plus performant qu'un modèle qui n'en a que 7.
- L'analyse de sentiments et la traduction automatique ont été les premiers usages de ces modèles, mais à présent ils sont capables de tenir une conversation complexe et même de passer des examens de médecines, avec un taux réussite supérieur à 85 %.

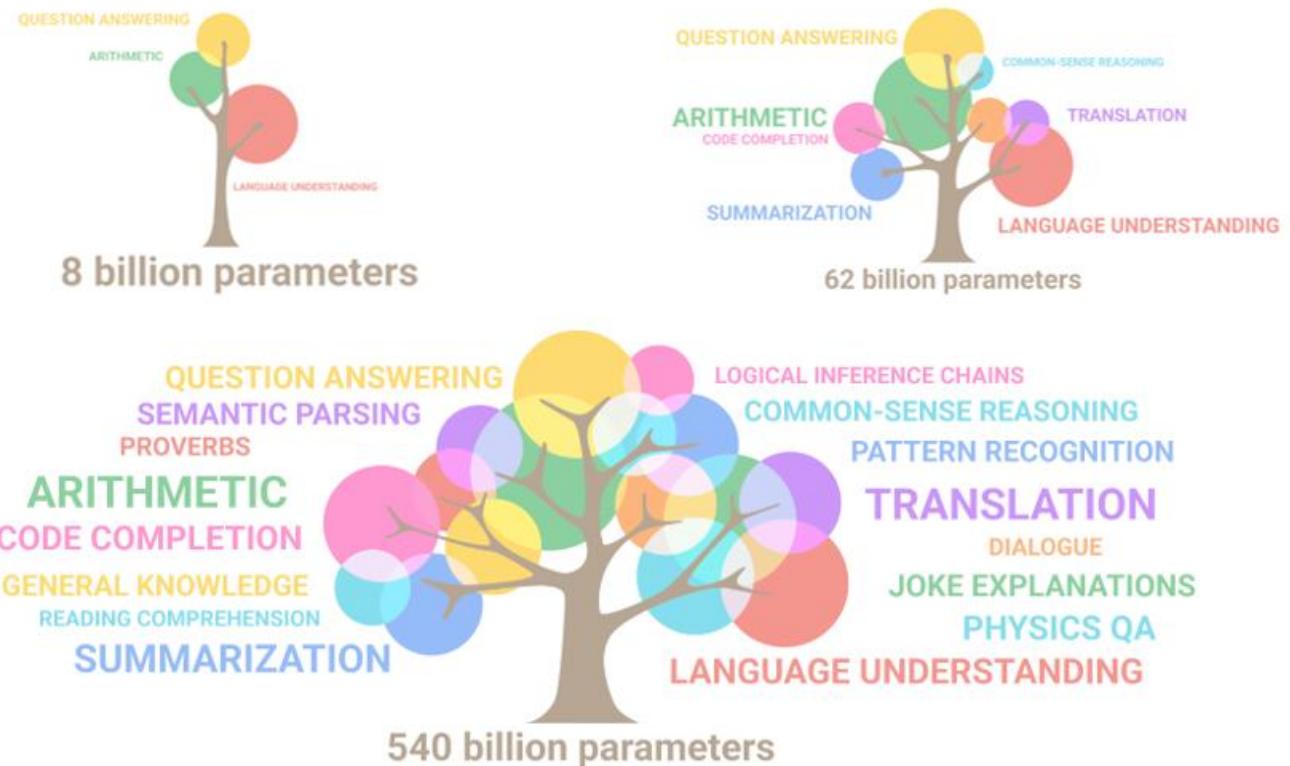


Illustration de l'émergence des capacités des LLM lorsque le nombre de paramètres augmentent

UN APERÇU DES INCONVÉNIENTS DES TRANSFORMERS

- Les Transformers ont été révolutionnaires dans le succès des LLM, mais il y a encore beaucoup de défis techniques à relever.
- Un premier inconvénient technique des LLM que tout utilisateur de ChatGPT a rencontré est leur capacité à inventer des faits de toute pièce. Aussi puissant soient-ils, l'hallucination est quelque chose qui touche tous les modèles. Certains chercheurs pensent même que c'est un phénomène intrinsèque, lié à l'architecture elle-même. Pour réduire les impacts, les entreprises préfèrent ajouter des réponses par défaut comme « mes données d'entraînement datent de décembre 2023 ».

```
kim@chatbot:~/text-generation-webui
(py310) kim@chatbot:~/text-generation-webui$ curl --request POST \
>   --url http://localhost:8080/completion \
>   --header "Content-Type: application/json" \
>   --data '{"prompt": "Qui est le Président Français?", "n_predict": 128}'
{"content": "\nBenoît Hamon est le Président de la République française depuis 2017.

Avant lui, François Hollande a été président de 2012 à 2017.", "generat

(py310) kim@chatbot:~/text-generation-webui$ curl --request POST --url h
--data '{"prompt": "Qui est le Président Français?", "n_predict": 128}'
{"content": "\n\nLe Président de la République française est Emmanuel Macron.

Il a été élu pour un premier mandat en 2017 et a été réélu en 2022.", "generation_
```

TD : COMPARER UN LLM GÉNÉRALISTE / UN LLM CONVERSATIONNEL

TD : LLM GÉNÉRALISTE VS LLM CONVERSATIONNEL

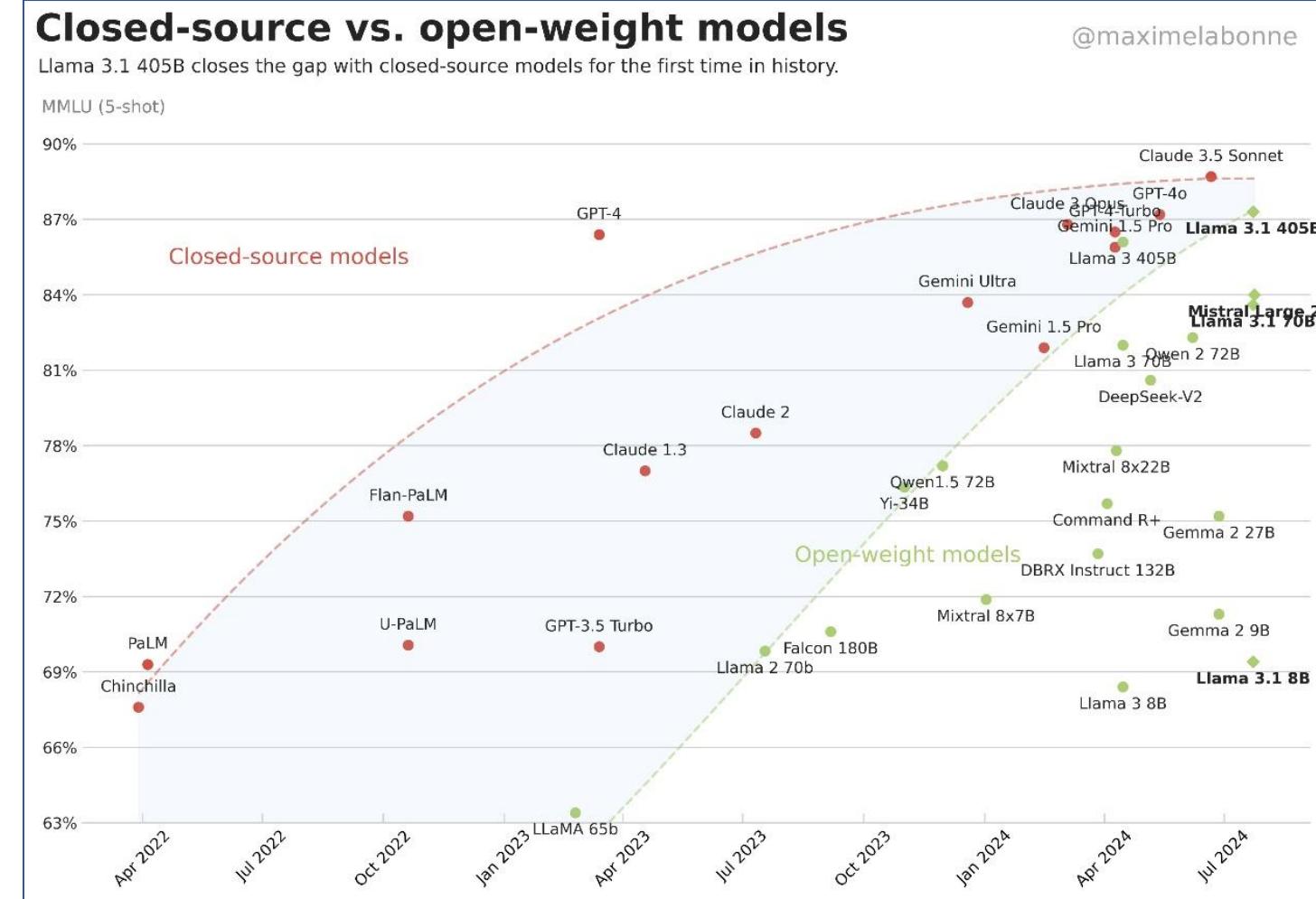
- Créez-vous un compte HuggingFace
- Manipuler les modèles de fondation et les modèles finetunés pour la discussion :
 - <https://huggingface.co/mistralai/Mistral-7B-v0.1>
 - <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1>

TOUR D'HORIZON DES LLM PROPRIÉTAIRES ET OPEN SOURCE



CHATGPT CONTRE LE MEILLEUR DEV DE FRANCE

- En 2023, OpenAI et Anthropic dominent le classement avec leur modèle GPT et Claude. A cette époque, il existe de nombreuses alternatives Open Source, c'est-à-dire des modèles que l'on peut déployer sur sa propre infrastructure, sinon ré-entrainer soi-même avec ses propres données.
- En Mars 2024, GPT-4 se fait détrôné de peu par le nouveau modèle Claude 3 d'Anthropic ; pendant longtemps GPT-4 occupait le haut du classement, mais maintenant beaucoup d'utilisateurs de ChatGPT migrent vers Claude pour des raisons de performance, de coût mais aussi à cause de la censure un peu trop forte.
- Une autre tendance en 2024 est la convergence progressive des modèles Open Source / Open Weights vers le haut du classement. Depuis Avril 2024, les modèles Open Source font leur grand retour, avec des nouveaux modèles de plus en plus puissants, publiés par Meta et MistralAI notamment.



Personnalités



CEO :
Sam Altman

Business Insider



Cofondé avec Elon
Musk, parti pour
désaccord
stratégique

Modèles



\$100,000/j



Entreprise

Sortie fin août 2023



API



GPT 3 et GPT 4

Les conversations
sont privées

1,3 millions de livres

300 milliards de mots
570 Go de texte

ChatGPT



Les conversations
peuvent être utilisées
pour améliorer leur
modèle

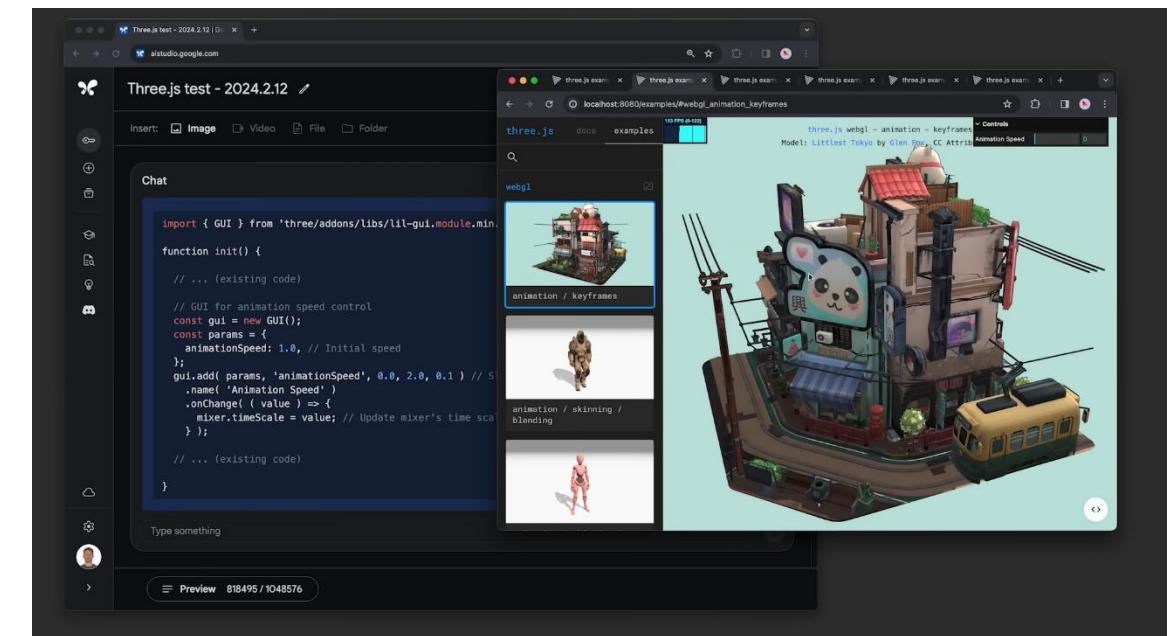
MICROSOFT AVEC COPILOT

- Intégrer GPT dans Office 365 et dans VS Code permet à OpenAI d'intégrer son modèle dans des outils déjà utiliser chez des millions de professionnels dans le monde ; sans compter les services proposés sur Azure Cloud.
- Dans Office 365, Copilot permet de résumer des documents Word, dresser des tableaux croisés dynamiques dans Excel, ou encore générer une présentation Power Point à partir d'un document Word (figure suivante).
- Pour les développeurs, GitHub Copilot est bien plus qu'un outil d'auto-complétion. Avec des commandes comme @workspace, le développeur peut donner des instructions et poser des questions qui concerne spécifiquement le projet sur lequel il travaille. Selon Microsoft, utiliser Copilot permet de programmer 55% plus vite car il écrit 46% du code.
- Les usages de cet outil sont à nuancer : de nombreux utilisateurs se plaignent sur Reddit que les réponses et le code proposés sont de mauvaise qualité en comparaison à ce que propose GPT dans l'interface chat Web. C'est pourquoi certains développeurs se contentent encore de copier-coller le code, non plus de StackOverflow, mais de ChatGPT.

The screenshot illustrates the GitHub Copilot interface integrated into a development environment. At the top, a file browser shows files like 'fetch_pic.js', 'push_to_git.py', 'd3_scale.js', 'fetch_stock.js', and 'material_ui.js'. Below this, a code editor displays a function 'fetchNASAPictureOfDay' with code completion suggestions from Copilot. The GitHub Copilot logo and name are prominently displayed. The interface then transitions to a Jupyter Notebook environment where Copilot provides instructions for loading a Titanic dataset. It suggests commands like '@workspace /newNotebook' and outlines steps for reading the dataset with pandas and displaying key values with Seaborn. The notebook code cell shows the command '# Load Titanic Dataset titanic_data = pd.read_csv('titanic.csv')'. Another section titled 'Inspect Dataset' shows code for inspecting the dataset's shape: '# Inspect Dataset # Display the shape of the dataset print(titanic_data.shape)'.

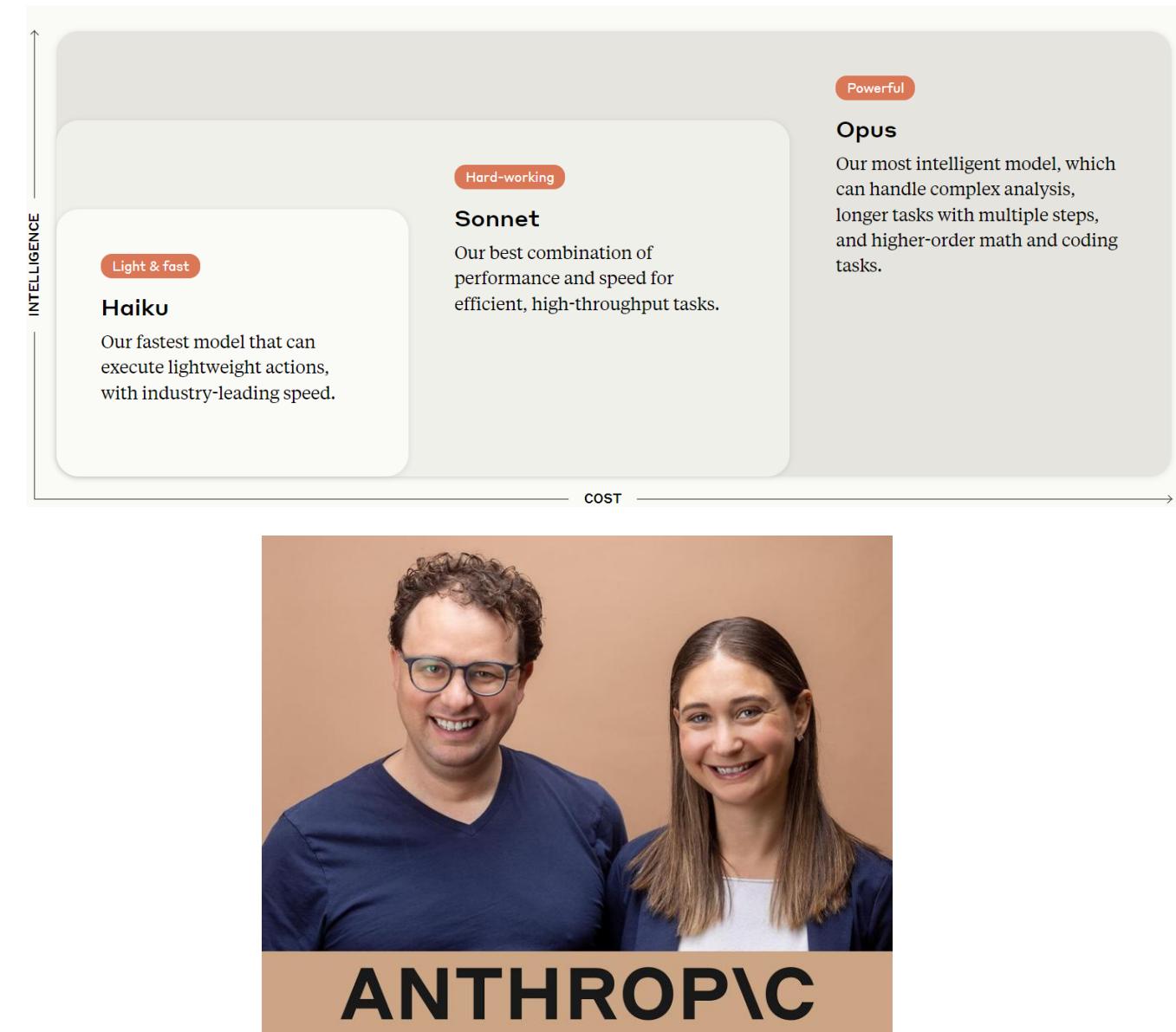
GOOGLE AVEC GEMINI

- Google travaille depuis de nombreuses années sur les technologies de chatbot et les LLM avec des modèles comme PaLM. Mais Google a tardé à se positionner sur le marché des chatbots grand public comme ChatGPT, qui est devenu un concurrent direct de son moteur de recherche, le cœur de métier de Google.
- Début 2023, Google a donc sorti son modèle propriétaire : Bard qui n'a pas eu le succès escompté. Fin 2023, Google propose le successeur de Bard : Gemini, décliné en plusieurs versions. Gemini nano, conçu spécifiquement pour les téléphones Android, Gemini Pro et Gemini Ultra, deux modèles de capacité croissante, accessibles sur leur portail.
- Après un démarrage difficile, Google a enfin tapé un grand coup avec Gemini Pro 1.5, dont le LLM affiche une fenêtre de contexte de 1 million de tokens.



ANTHROPIC AVEC CLAUDE

- Anthropic a été fondée en 2021 par deux anciens de chez OpenAI, Daniela Amodei et Dario Amodei, qui est le CEO. L'entreprise est soutenue avec de gros investissements notamment de la part d'Amazon (4 milliards de dollars) et de Google (2 milliards de dollars).
- Anthropic propose leur modèle Claude, en chat et en API. Le nom du modèle est un hommage au chercheur Claude Shannon, mathématicien considéré comme le père de la théorie de l'information.
- L'avantage concurrentiel de Claude est la fenêtre de contexte de plus de 200 000 tokens, contre 128 000 tokens pour GPT-4. Malgré cela peu de personnes l'utilisait. Mais avec l'arrivée de Claude 3, et en particulier de Claude 3 Opus, la déclinaison la plus puissante, la tendance s'inverse, et les utilisateurs quittent ChatGPT, dont ils estiment les performances en chute libre depuis plusieurs mois. Cependant Anthropic a eu des difficultés à se faire connaître notamment en Europe. A cause ou grâce aux régulations de l'Union Européenne, Claude n'était pas accessible en France.



ANTHROPIC

Dario et Daniela Amodei, fondateur d'Anthropic et ancien d'OpenAI. Source : Anthropic

META AVEC Llama

- C'est en juillet 2023 que Meta lance son LLM : LlaMa 2, dans 3 déclinaisons, 7, 13 et 70 milliards de paramètres, ce dernier étant le plus capable mais le plus exigeant en ressource de calcul.
- Afin de garder un certain contrôle sur son utilisation et pour éviter les usages déviants, LlaMa 2 n'est distribué qu'à un petit cercle d'experts et de chercheurs, Meta ne voulait pas endosser la responsabilité que quelqu'un utilise LlaMa 2 pour confectionner des armes biologiques par exemple.
- Cependant, un chercheur, soucieux que Meta ne tienne pas sa promesse de mettre en Open Source et Open Weight son modèle, a publié les poids du modèle sur internet et l'a partagé via un lien magnet. Cette action a pris de court Meta qui a décidé par la suite de publier l'ensemble des travaux liés à LlaMa 2 : article scientifique décrivant le processus, une partie du code pour l'entraînement, les poids du modèle, et surtout la publication sous une licence commerciale souple, à l'exception d'une interdiction pour les autres GAFAM.
- Cette manière de publier le modèle par lien magnet sera reproduite un peu plus tard par Mistral.
- Pendant une petite période LlaMa 2 était le seul modèle à disposition pour le monde de l'Open Source qui avait à la fois des performances décentes et qui puisse être utilisable commercialement. Aujourd'hui, il y a de nombreuses alternatives.
- Meta continue dans cette direction en partageant LlaMa 3 en juillet 2024, qui explose les records de performance.

The screenshot shows a list of Llama models on the Hugging Face Model Hub. The interface includes a search bar and a filter for 'Models'. The results are sorted by 'Recently updated'. The listed models are:

- meta-llama/Meta-Llama-3.1-405B (Text Generation, Updated 3 days ago, 159k, 682 likes)
- meta-llama/Meta-Llama-3.1-405B-Instruct (Text Generation, Updated 3 days ago, 48.2k, 402 likes)
- meta-llama/Llama-Guard-3-8B (Text Generation, Updated 5 days ago, 36.2k, 72 likes)
- meta-llama/Meta-Llama-3.1-8B-Instruct (Text Generation, Updated 13 days ago, 1.05M, 1.71k likes)
- meta-llama/Prompt-Guard-86M (Text Classification, Updated 17 days ago, 412k, 142 likes)
- meta-llama/Meta-Llama-3.1-405B-Instruct-FP8 (Text Generation, Updated 14 days ago, 42.6k, 142 likes)
- meta-llama/Meta-Llama-3.1-70B (Text Generation, Updated 19 days ago, 42k, 183 likes)

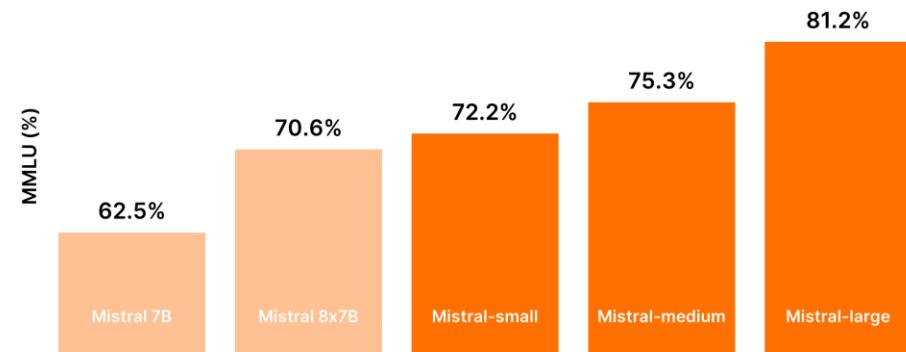
2. Additional Commercial Terms. If, on the Llama 2 version release date, the monthly active users of the products or services made available by or for Licensee, or Licensee's affiliates, is greater than 700 million monthly active users in the preceding calendar month, you must request a license from Meta, which Meta may grant to you in its sole discretion, and you are not authorized to exercise any of the rights under this Agreement unless or until Meta otherwise expressly grants you such rights.

Mention légale anti GAFAM

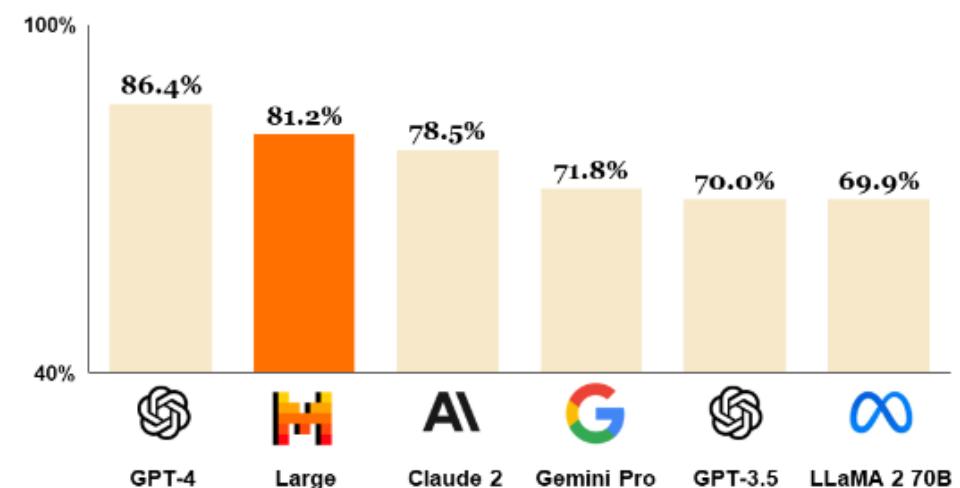
MISTRALAI AVEC MISTRAL

- MistralAI a été créé en Mai 2023 par trois français : Guillaume LAAMPLE, Directeur scientifique anciennement chez Meta, Timothée LACROIX, CTO, venant également de chez Meta et Arthur MENSCH, CEO, précédemment chez DeepMind.
- En septembre 2023, quelques mois seulement après sa création, MistralAI sort son premier modèle Mistral 7B, et il est en Open Weight. Cet évènement a propulsé l'entreprise sur le devant de la scène, comme étant le nouvel espoir de la communauté Open Source.
- Mistral a réitéré en lançant un second modèle en décembre 2023 dit Mixture of Expert (MoE), Mixtral 8x7B. Ce type de modèle a pour spécificité de diviser les compétences du modèle général et de les distribuer à des « sous-modèles experts ». Mixtral possède 46,7 milliards de paramètres au total mais n'utilise que 12,9 milliards de paramètres. Il traite donc l'entrée et génère la sortie à la même vitesse et pour le même coût qu'un modèle de 12,9 milliards.
- Les benchmarks montrent que Mistral 8x7B avait des performances équivalentes à GPT-3.5. C'était aussi pendant longtemps le seul Open Model à gérer officiellement plusieurs langues Européennes : Anglais, Français, Italien, Allemand et Espagnol. Et c'était l'un des seuls modèles à pouvoir être exécuter sur des ordinateurs de la grande consommation, avec des configurations haut de gammes cependant.
- Mais la sortie des modèles propriétaires, Mistral small, medium et large ont fait mauvaise presse sur les réseaux sociaux, et créa une énorme déception dans la communauté Open Source. La position de Mistral est la suivante : être les meilleurs sur les modèles Open Weight, mais pouvoir financer la conception et le développement de ces dits modèles en vendant la performance de modèles propriétaires plus puissants.
- Début Avril 2024, Mistral a publié un nouveau modèle MoE, Mistral 8x22B, et il s'inscrit en haut des classements parmi les modèles sous licence Apache 2.0.

Mistral Models on MMLU



MMLU



Haut : Performance des modèles de Mistral. En clair les modèles Open Weights, en foncé les modèles propriétaires.

Bas : Performance de Mistral Large comparé aux autres grands modèles Source : MistralAI

COHERE AVEC COMMAND R

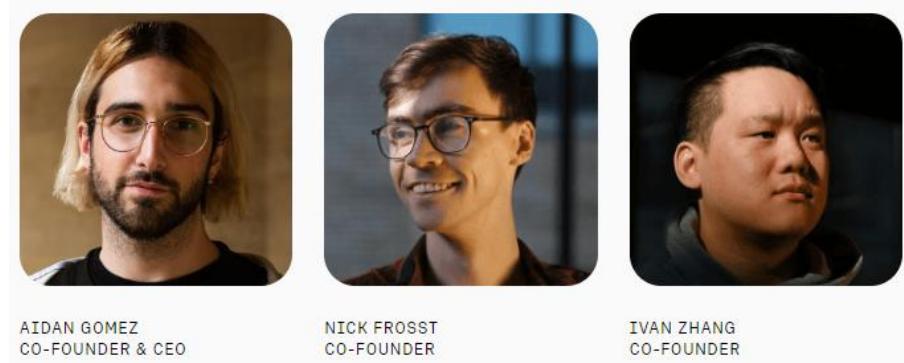
- Peu connue du grand public, Cohere est une entreprise Canadienne spécialisée dans les LLM qui a été fondée en 2019.
- L'offre de service est similaire à ce que propose OpenAI, à la différence que Cohere a surpris la communauté en publiant deux modèles très puissants : command R et command R+. Command R+, qui est le plus puissant, arrive à se hisser au niveau de GPT-4 et de Clause 3. Le modèle s'est montré très performants dans le multilingue et la récupération de données. Malheureusement, les deux modèles sont publiés avec une licence qui ne permet pas les utilisations dans un cadre commercial.

Total #models: 80. Total #votes: 599638. Last updated: April 9, 2024.

Contribute your vote  at [chat.lmsys.org!](https://chat.lmsys.org/) Find more analysis in the [notebook](#).

Rank	Model	Arena Elo	95% CI	Votes	Organization	License
1	Claude_3_Opus	1256	+3/-4	44882	Anthropic	Proprietary
1	GPT-4-1106-preview	1254	+3/-3	61868	OpenAI	Proprietary
1	GPT-4-0125-preview	1250	+3/-4	45101	OpenAI	Proprietary
4	Bard_(Gemini_Pro)	1208	+5/-5	12468	Google	Proprietary
4	Claude_3_Sonnet	1204	+3/-3	54349	Anthropic	Proprietary
6	Command_R+	1194	+4/-6	13711	Cohere	CC-BY-NC-4.0
6	GPT-4-0314	1188	+4/-4	40321	OpenAI	Proprietary
8	Claude_3_Haiku	1182	+3/-4	45879	Anthropic	Proprietary
9	GPT-4-0613	1163	+3/-3	59434	OpenAI	Proprietary
9	Mistral-Large-2402	1158	+3/-4	32813	Mistral	Proprietary
10	Qwen1.5-72B-Chat	1153	+4/-6	26314	Alibaba	Qianwen LICENSE
10	Claude-1	1150	+5/-6	21868	Anthropic	Proprietary
11	Mistral_Medium	1148	+4/-4	29226	Mistral	Proprietary
11	Command_R	1148	+4/-4	27854	Cohere	CC-BY-NC-4.0
12	Qwen1.5-32B-Chat	1140	+6/-5	10871	Alibaba	Qianwen LICENSE

Classement ELO de Command R+, le 9 avril 2024



AIDAN GOMEZ
CO-FOUNDER & CEO

NICK FROSST
CO-FOUNDER

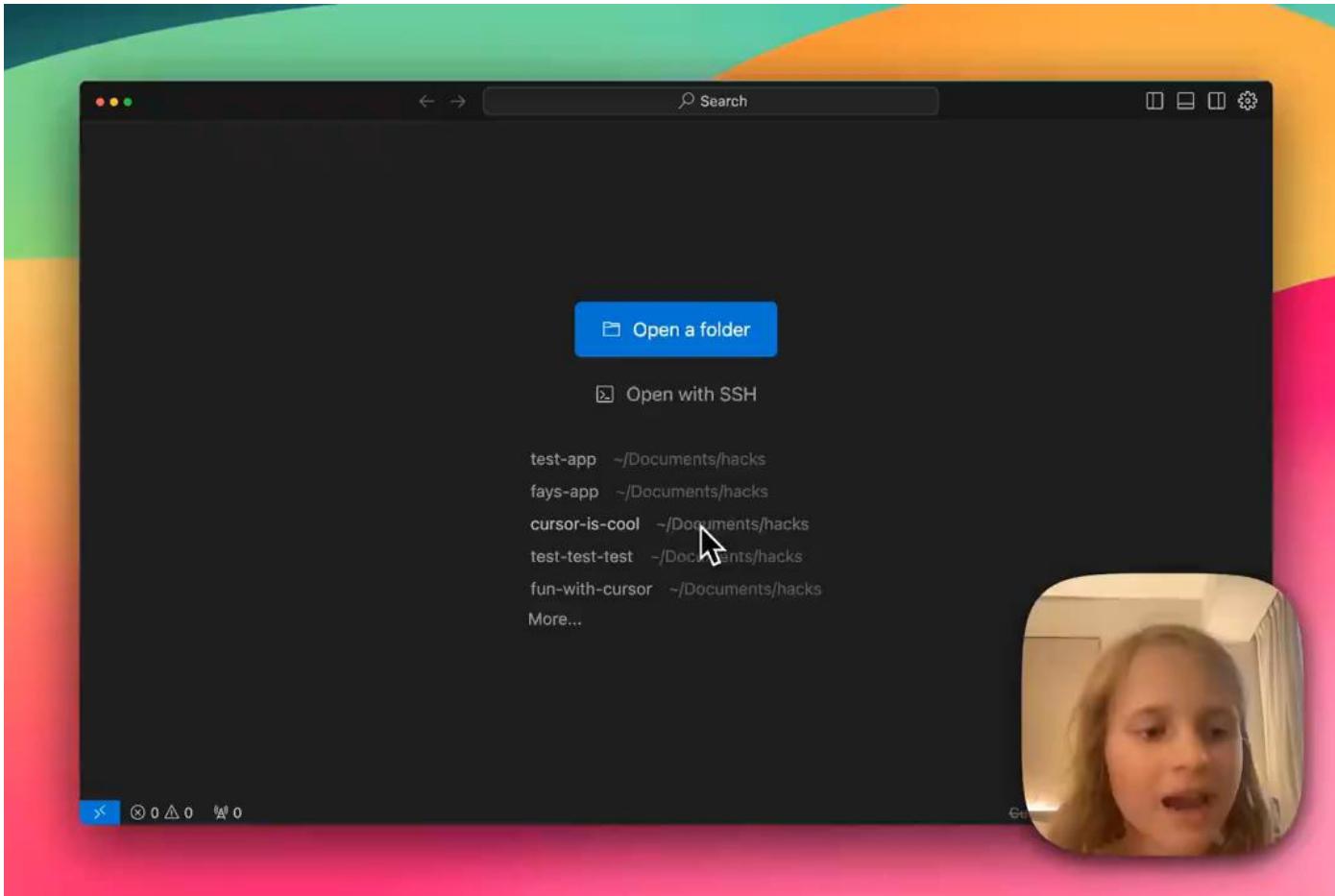
IVAN ZHANG
CO-FOUNDER

Les cofondateurs de Cohere : Aidan GOMEZ, l'un des co-auteurs de l'architecture Transformers chez Google, Nick FROSST, anciennement chez Google, et Ivan ZHANG, anciennement chez FOR.ai.

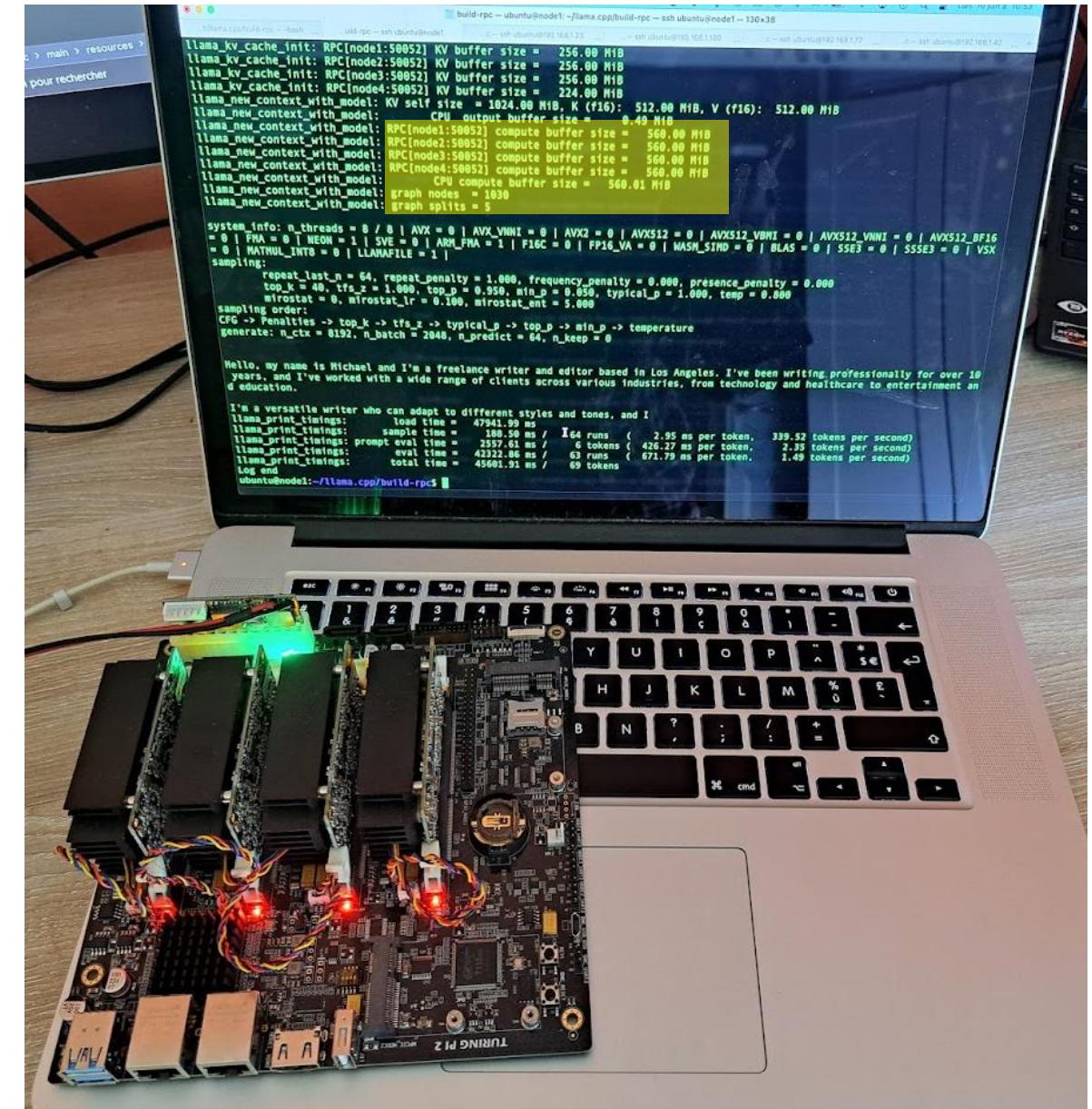
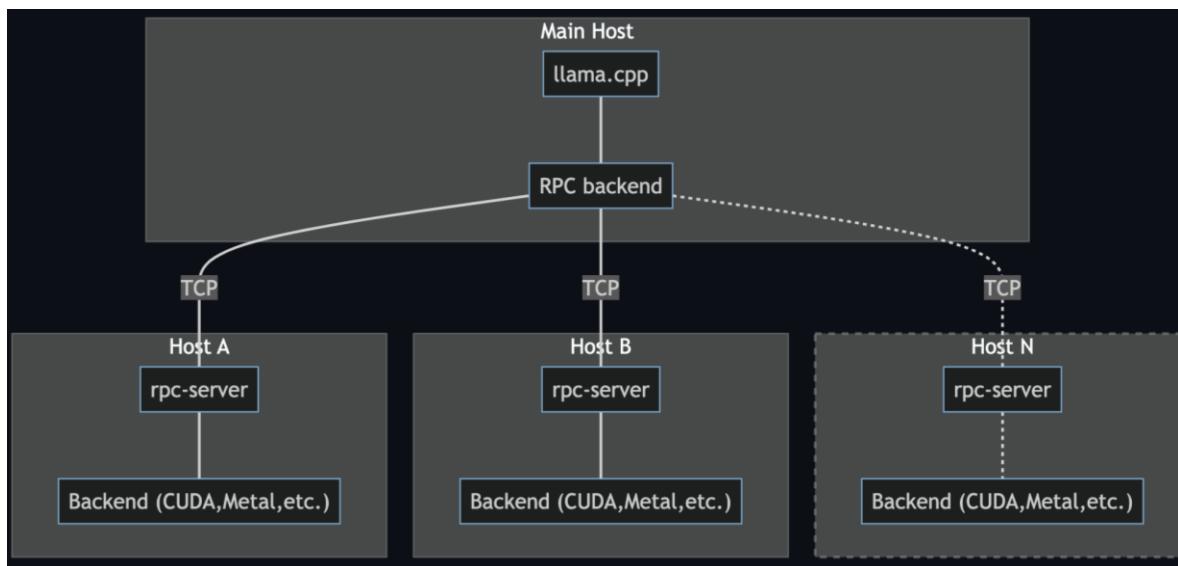
Source : Cohere

CURSOR

- Cursor est un VSCode avec l'IA générative intégrée nativement. Il est possible de choisir le LLM de son choix, et d'utiliser sa propre clef API.
- La programmation par langage naturel devient – naturelle – et une petite fille de 8 ans nous le démontre.

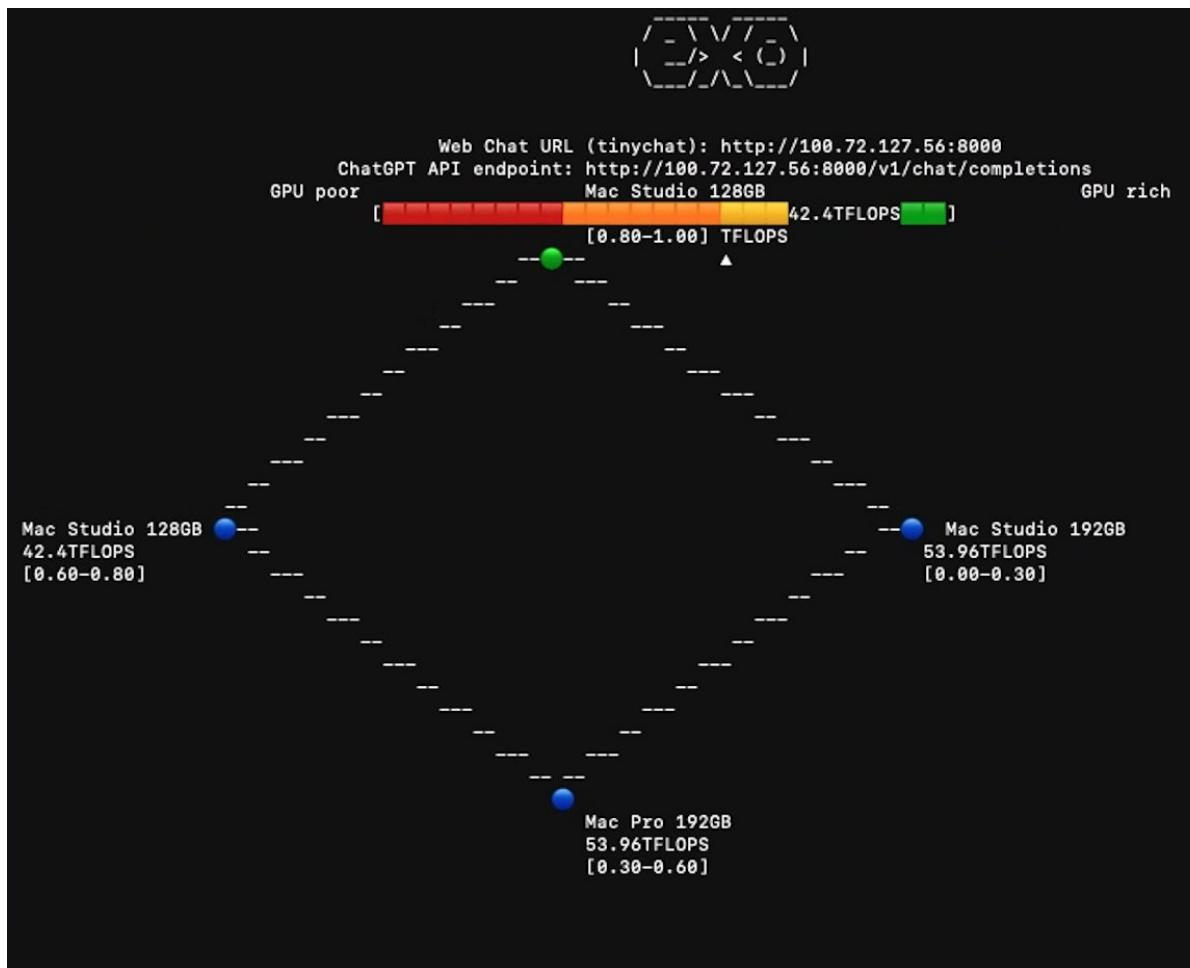


LLAMA.CPP



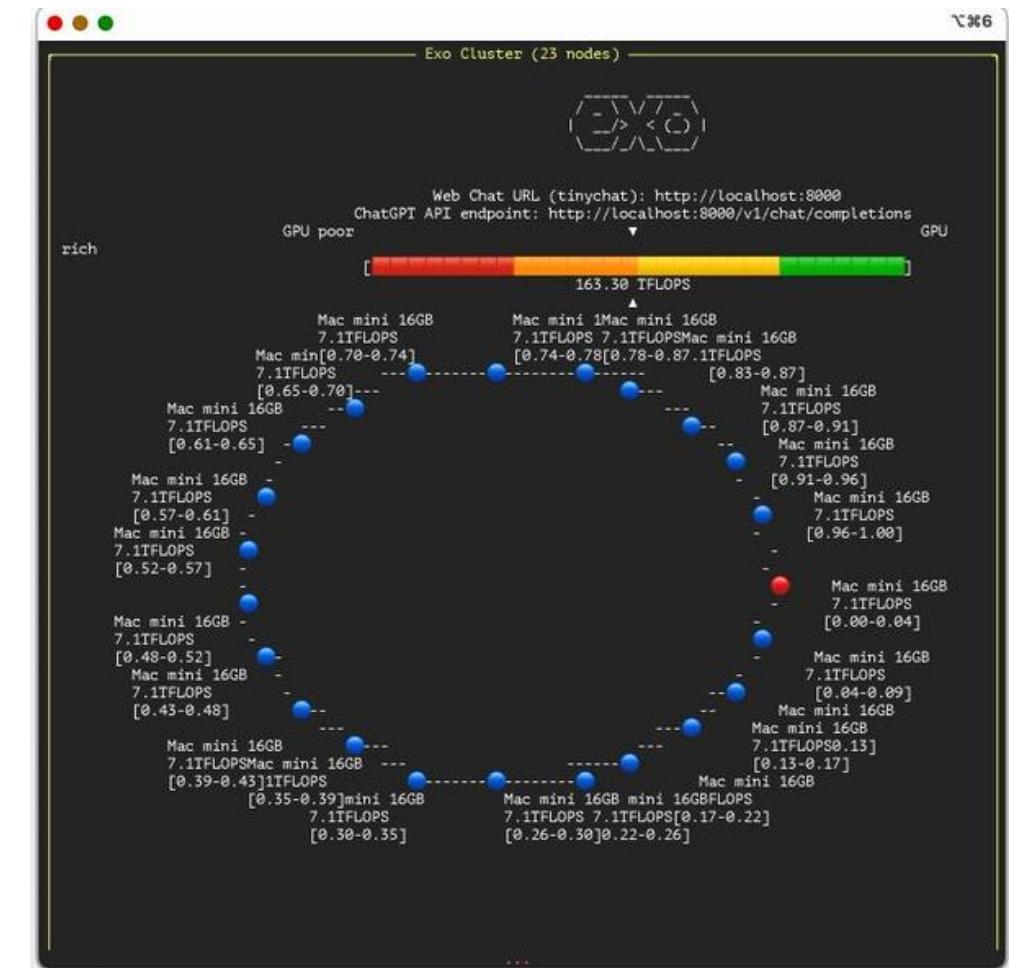
EXOLABS

640GB of memory across 4 nodes discussing [NSFW]
thanks to llama 405b



23 x Mac Mini 16GB (total 368GB).

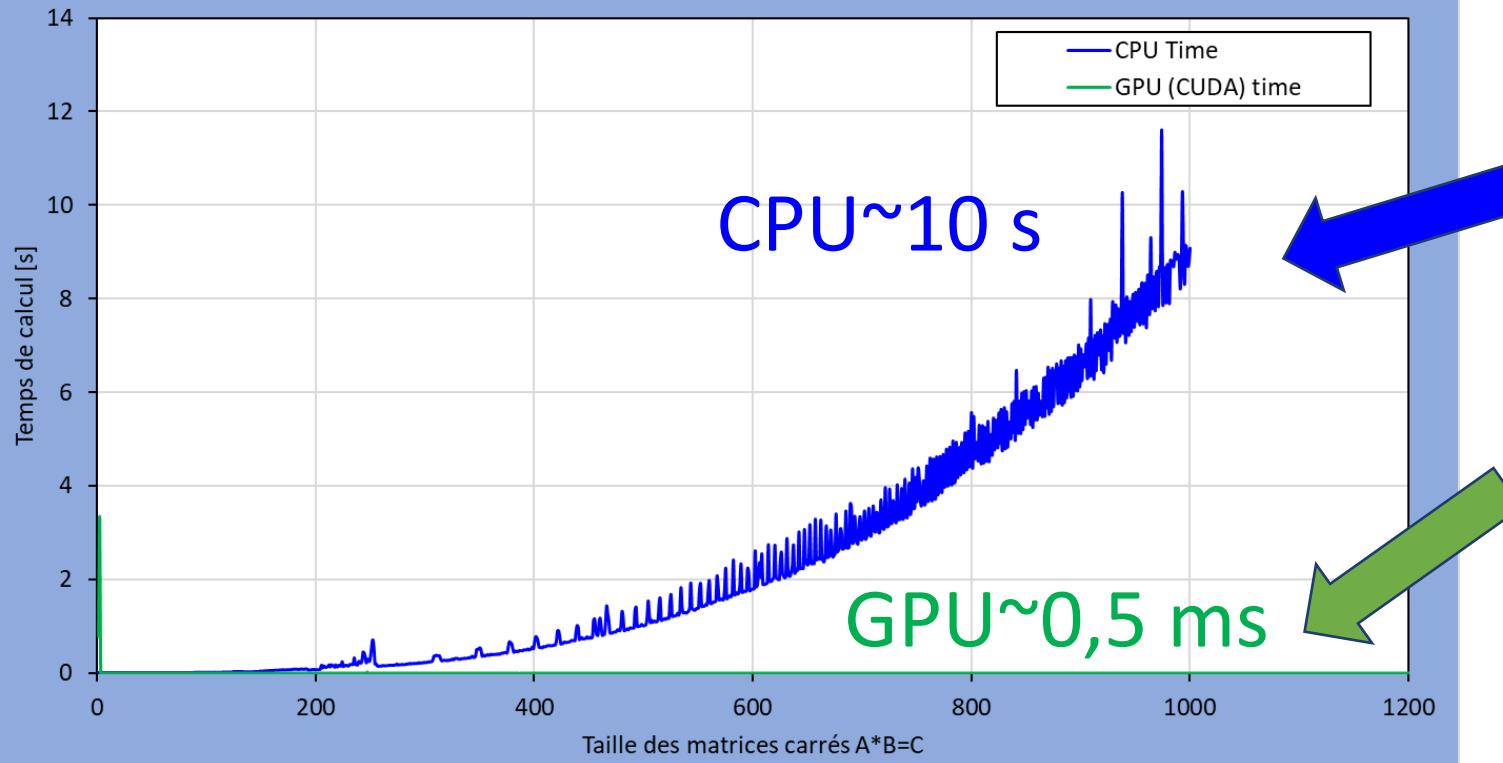
Enough to run Llama-3-405b at Q4_K_S (230GB w/o cache).



LE GPU PERMET L'ACCÉLÉRATION DE L'INFÉRENCE DE TEXTE

- La génération de texte implique des multiplications matricielles dans le réseau de neurones.
- Le GPU permet une accélération entre 4 et 5 ordre de grandeurs par rapport au CPU.

Comparaison du temps de calcul CPU vs GPU (Nvidia T4) de multiplication de matrices



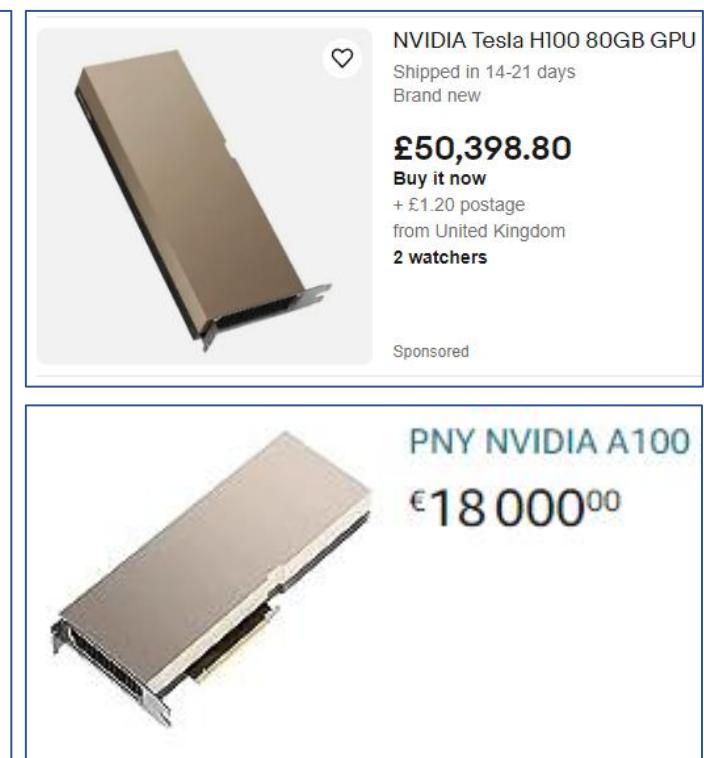
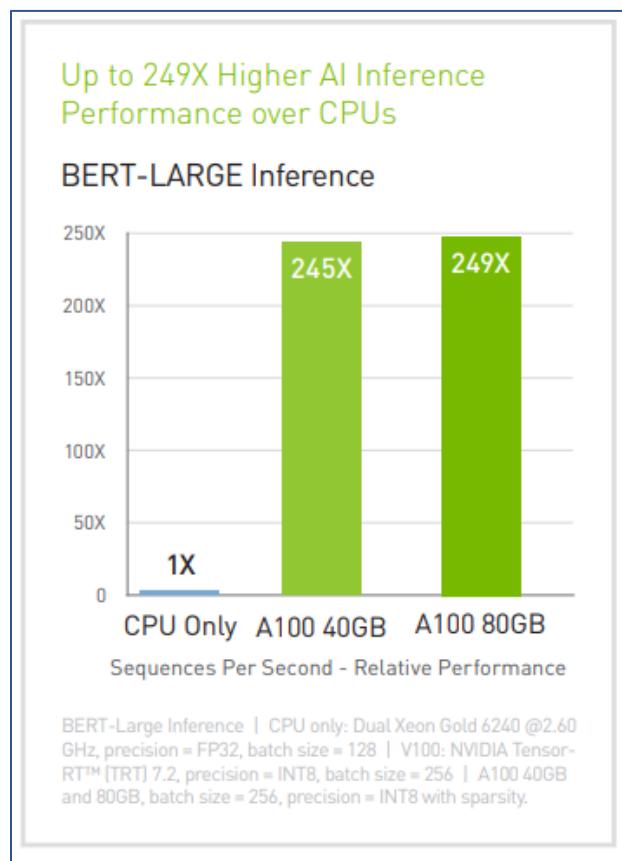
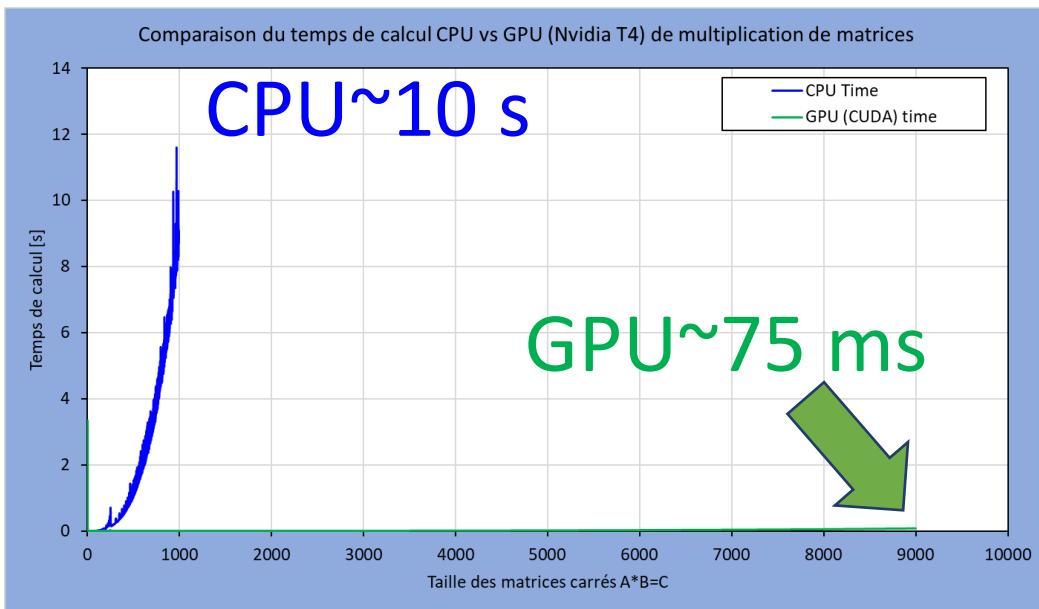
```
# Measure matrix multiplication time for CPU (NumPy)
start_cpu = time.time()
np.dot(A_cpu, B_cpu)
end_cpu = time.time()
cpu_times.append(end_cpu - start_cpu)

# Measure matrix multiplication time for GPU (CuPy)
start_gpu = cp.cuda.Event()
end_gpu = cp.cuda.Event()

start_gpu.record()
cp.dot(A_gpu, B_gpu)
end_gpu.record()
end_gpu.synchronize()
```

QUELLE CONFIGURATION DE GPU POUR VOTRE BESOIN ?

Au vue de la différence de performance, il semble que l'inférence avec un GPU est nécessaire.

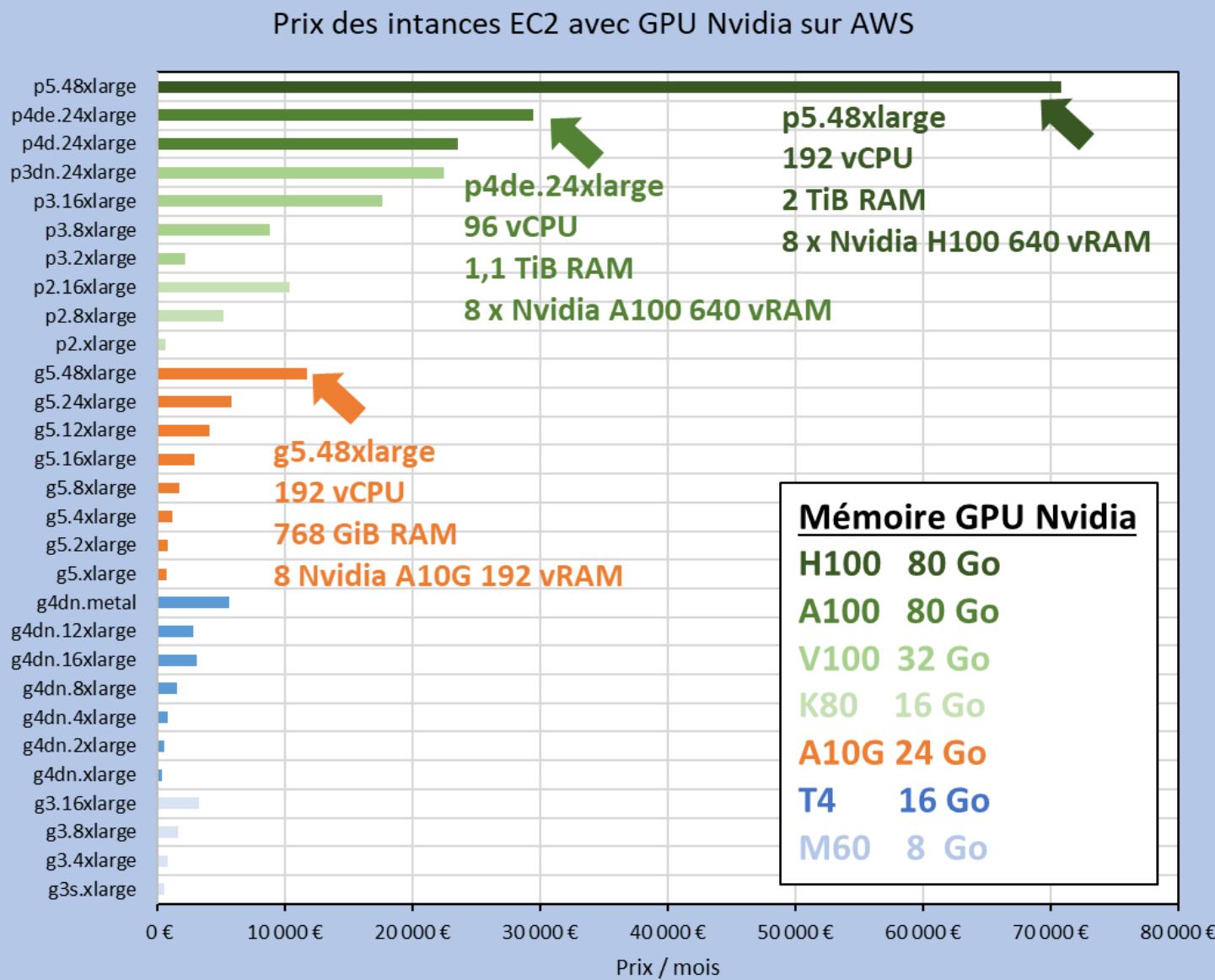


QUEL CLOUD PROVIDER CHOISIR ? QUELLE CONFIG MATERIEL ?



À QUEL PRIX? (PRIX/MOIS CAR INFÉRENCE)

Instances AWS



POSITIONNEMENT DES ACTEURS DANS LA CHAINE DE VALEUR

- **Silicone** : le hardware n'a jamais été aussi important ; toutes les entreprises technologiques dépendent d'un petit nombre de constructeurs de puces, et notamment Nvidia, qui domine le marché du GPU avec plus de 80% de part de marché. Afin de réduire cette dépendance, certaines entreprises ont développé leurs propres puces spécialisées pour l'IA, comme les TPU de Google, ou AWS Trainium par Amazon. Il y a aussi des acteurs alternatifs comme Groq, qui propose des puces optimisées pour l'inférence, et Lamihi qui démontre qu'il est possible d'entrainer un LLM sur des GPU AMD.
- **Infrastructure** : peu d'entreprises possèdent de datacenters exploitables à grande échelle. Seuls les grands fournisseurs de Cloud sont capables de fournir la puissance de calcul nécessaire pour entraîner et mettre en production un LLM pour des milliers d'utilisateurs. Meta fait figure d'exception, car il est capable d'entrainer LLaMa 2 sur sa propre infrastructure.
- **Model** : cela concerne la partie scientifique et les capacités que peut délivrer le LLM. Nous l'avons vu, il y a actuellement une myriade d'acteurs capables de livrer des modèles très performants.
- **Application** : est la capacité de déployer un produit utilisant la puissance d'un LLM à un grand public. Même si quasiment tous les acteurs sont capables de déployer une API et une interface de chat avec leur LLM, seuls Microsoft avec « Copilot » et Google avec « AI Studio » ont déployé un produit à grande valeur ajoutée pour les développeurs.

	 Silicon	 Infrastructure	 Model	 Application
Partners				
Proprietary				
 OpenAI				
 Microsoft				
ANTHROPIC  amazon				
 Google				
 Meta				
 Mistral AI				
 Grok				
 databricks				
 cohere				
 NVIDIA				
Contribute to Open Source				
Shovel vendor during the gold rush				



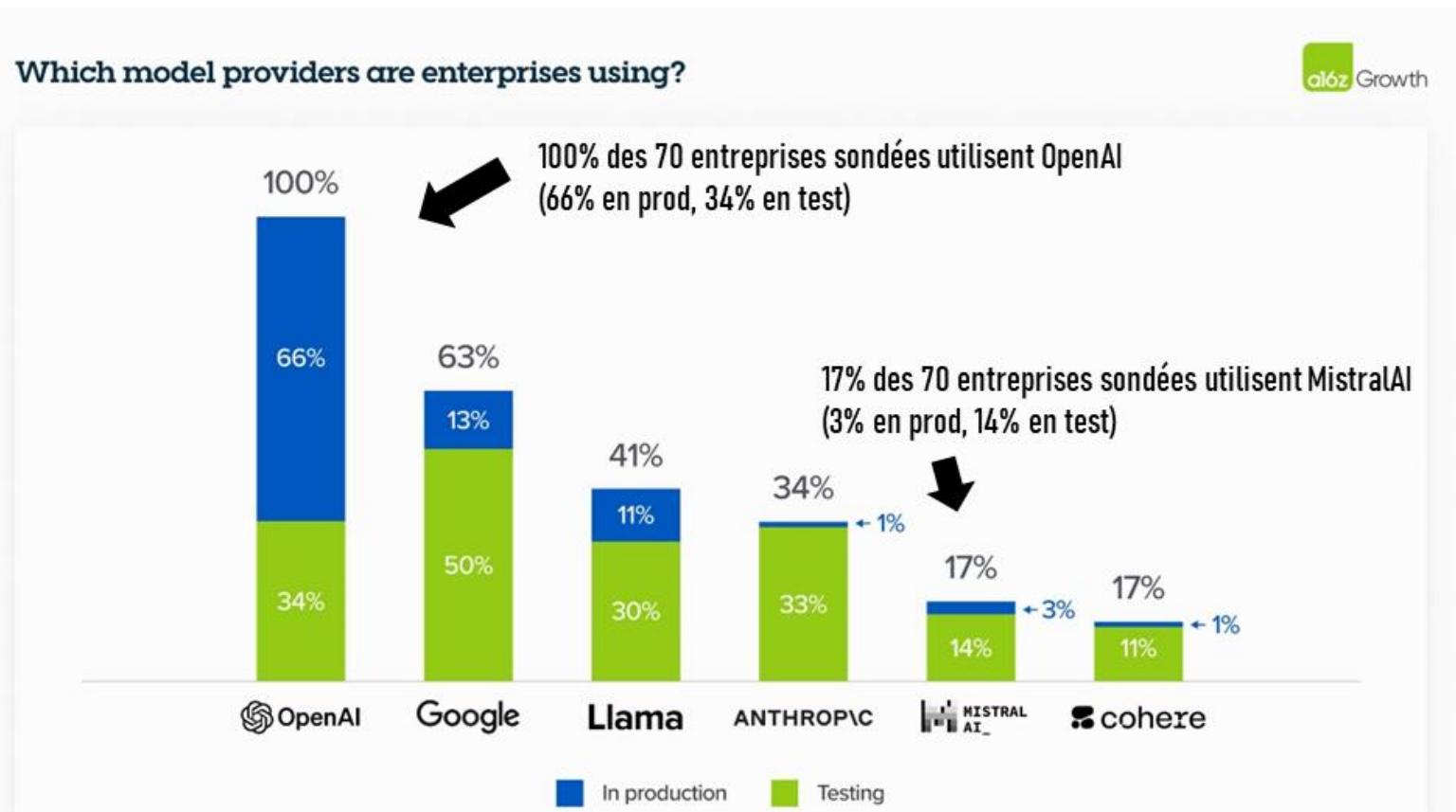
IMPACT POUR LES DSI, RETOUR D'EXPÉRIENCE ET PERSPECTIVES DE MARCHÉ



QUELS LLM SONT EN PRODUCTION DANS LES DSIS ?

- OpenAI est le plus utilisé, avec 100 % des entreprises l'utilisant, 66% en production et 34% en phase de test ;
- Google vient ensuite, utilisé par 63 % des entreprises, 13 % en production et 50% en phase de test ;
- Meta est utilisé par 41% des entreprises, 11 % en production et 30% en phase de test.
- Anthropic a un taux d'utilisation de 34 %, avec 33 % en production et 1% en phase de test.
- Mistral AI sont utilisés par 17 % des entreprises, avec 14 % des modèles de Mistral AI en production et 3% en phase de test ;
- Cohere sont utilisés par 17 % des entreprises, avec 11 % des modèles de Cohere en production et moins de 1% en phase de test.

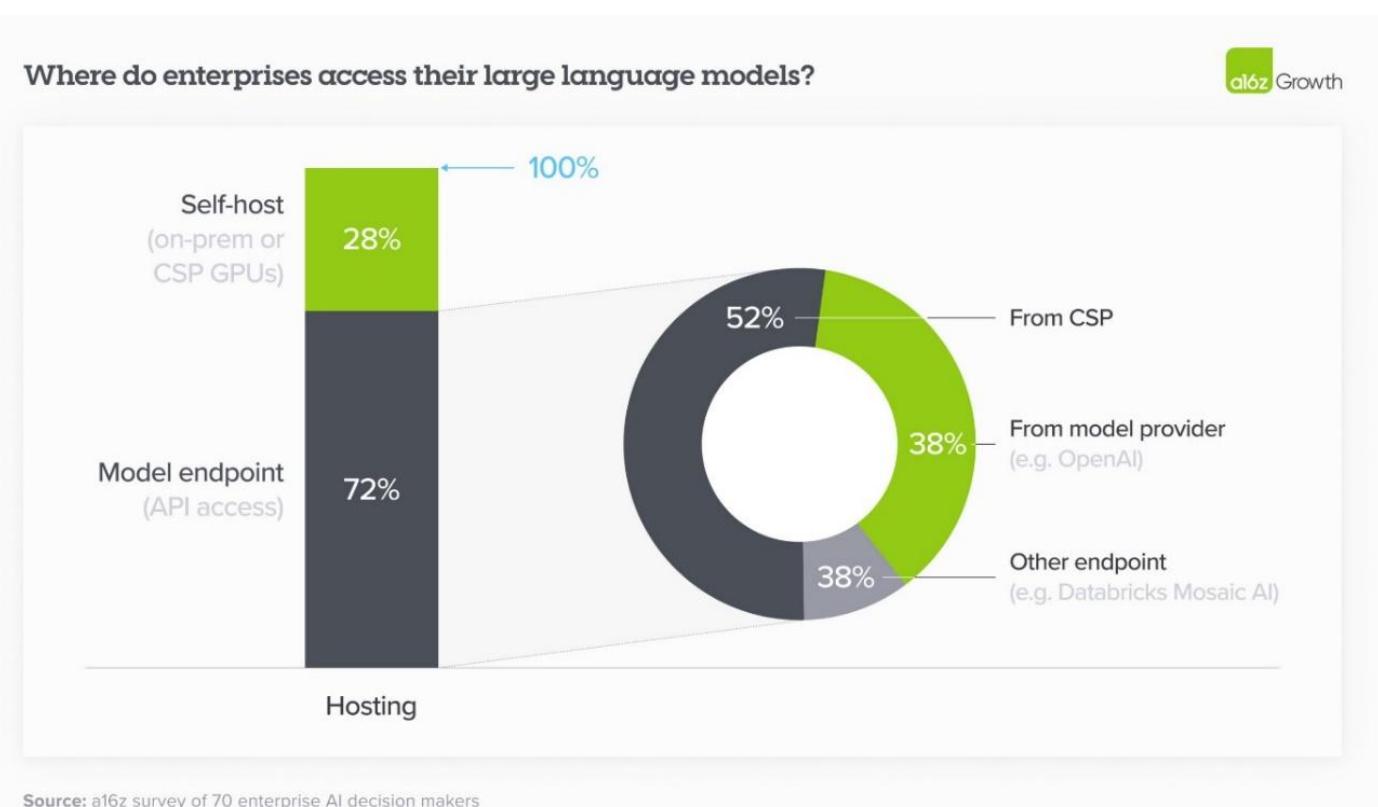
Adoption de différents fournisseurs de LLM chez les 70 entreprises sondées et la répartition des déploiements en test et en production. Source : a16z.



LLM EN PRODUCTION : API OU INFRA MAISON ?

- En 2023, de nombreuses entreprises ont acheté des modèles auprès de leur fournisseur de services Cloud (CSP) existant pour des raisons de sécurité, et par crainte que les données puissent être utilisées pour le réentraînement des modèles.
- En ce sens, les utilisateurs d'Azure préfèrent généralement OpenAI, tandis qu'Amazon préfère Anthropic ou Cohere. Comme le montre le graphique ci-dessous, sur les 72% des entreprises qui utilisent une API pour accéder à leur modèle, plus de la moitié utilisent le modèle hébergé par leur CSP.
- Plus d'un quart des répondants hébergent eux-mêmes leur modèle, probablement pour exécuter des modèles Open Source.

Sur les 72% des entreprises qui utilisent une API pour accéder à leur modèle, plus de la moitié utilisent le modèle hébergé par leur fournisseur de services Cloud. Source : a16z.



LES CAS D'USAGES À L'ÉCHELLE DE L'ENTREPRISE

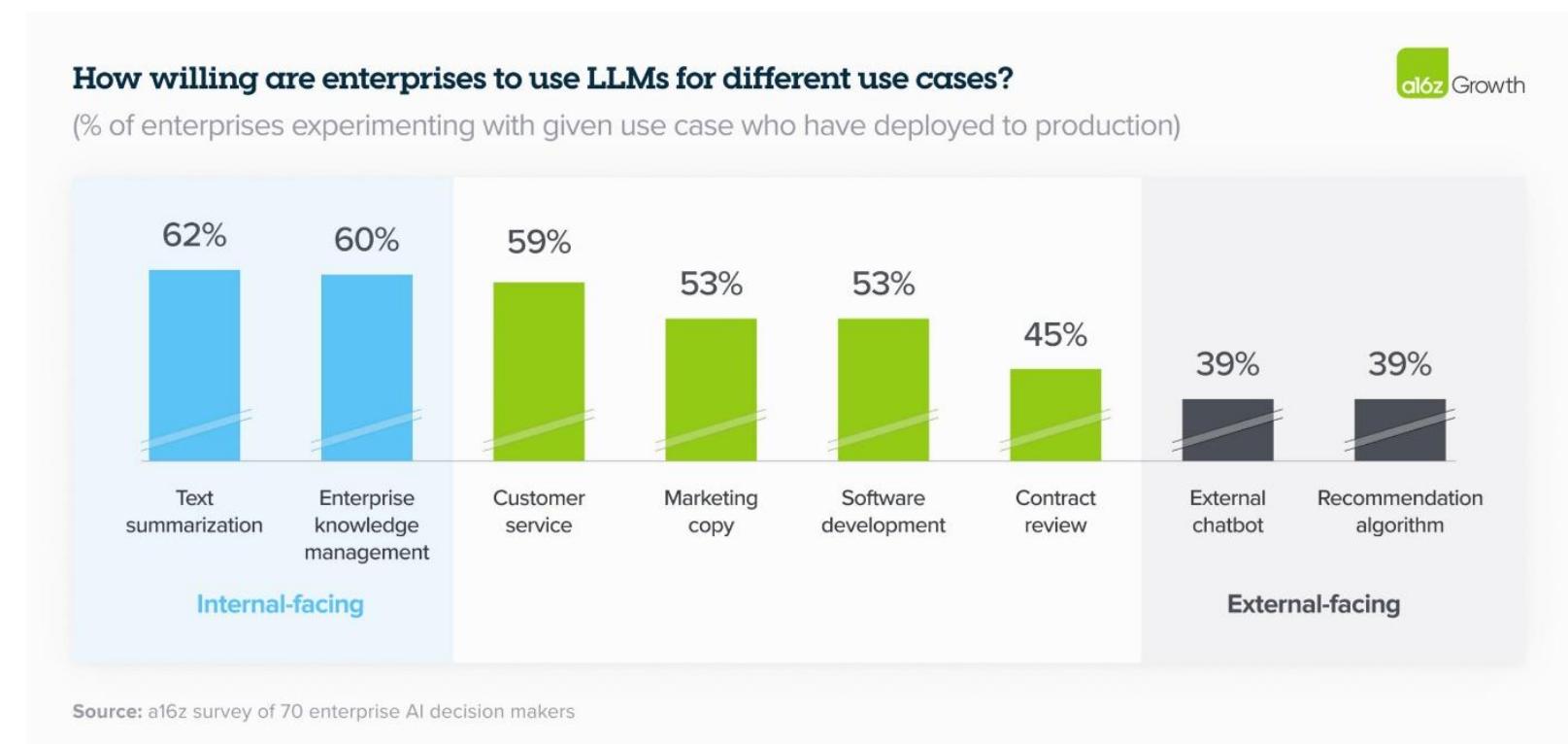
Les entreprises sont motivées par les cas d'utilisation internes, mais restent plus prudentes quant aux cas d'utilisation externes. Cela s'explique par le fait que deux préoccupations principales concernant l'IA générative persistent toujours dans l'entreprise :

- Les problèmes potentiels de dérapage et l'augmentation de la surface d'attaque ;
- Les problèmes de relations publiques liés au déploiement de l'IA générative, en particulier dans les secteurs sensibles des services financiers et de la santé (par exemple, la santé et les services financiers).

Les cas d'utilisation les plus populaires de l'année dernière étaient axés sur la productivité interne comme les assistants de programmation, le support client et le marketing.

Les entreprises priorisent les usages internes comme la synthèse de texte et la gestion des connaissances (par exemple, le chatbot interne) plutôt que les usages sensibles comme la revue de contrat ou les chatbots externes. Les entreprises veulent éviter les déboires de l'IA générative comme a connu Microsoft avec Tay, qui s'est devenue raciste suite aux provocations des internautes.

Les entreprises sont plus entrain à utiliser les LLM pour des usages internes ou semi internes, mais réticentes à l'exposition au public, de peur du dérapage conversationnel et d'attaque informatique. Source : a16z.



IMPACT SUR LA STRUCTURE ET L'ORGANISATION DES DSIs

- Une analogie avec la mécanisation de l'agriculture
- Si l'on peut s'accorder une analogie, la mécanisation de l'agriculture du XXème siècle aura sans doute un impact similaire à l'IA sur l'industrie du logiciel. La mécanisation a démultiplié le rendement des champs, et étendu la surface d'exploitation des fermes, tout en réduisant considérablement le nombre de fermiers.
- Le parallèle esquisse à grosse maille ce que sera la DSi du futur. L'industrie du logiciel va se transformer, mais on aura toujours besoin de développeurs, tout comme l'agriculture a besoin de fermier sur le tracteur.



Moissonneuse actionnée par un attelage de 33 chevaux en 1902 dans l'État de Washington, États-Unis.

Nous pouvons faire un parallèle entre les chevaux et les développeurs, qui utilisent la puissance mécanique animale pour récolter / farmer les tickets Jira.

Source : Wikimédia Commons.



Gauche : Des bénévoles de la Royal Navy prêtent main-forte pour la récolte du blé dans la région de Slapton, dans le Devon. Le chargement se fait manuellement à la fourche.

Source : Imperial War Museums, Wikimédia Commons.

Droite : Etudiants en agriculture apprenant à utiliser et à entretenir des tracteurs avec un enseignant, Sylvania, Géorgie, 1951. Artiste : Screeven-Jenkins Regional Library System.

Source : Wikimédia Commons.



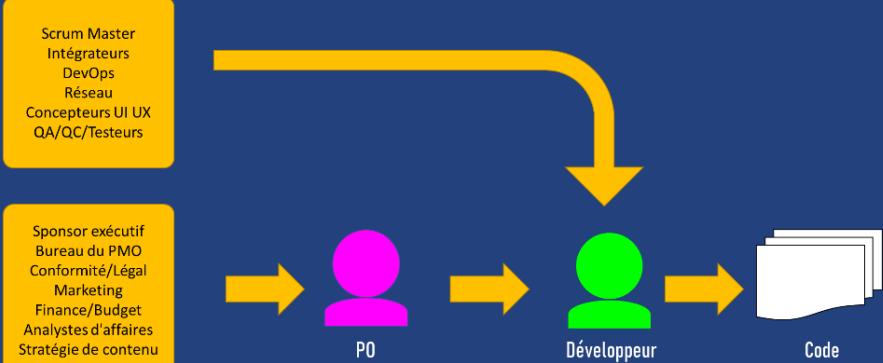
L'évolution de la technologie a démultiplié le rendement et agrandit les surfaces d'exploitation. L'IA bien maîtrisée permettra de livrer plus de code.

Artiste : Michael Gäbler.

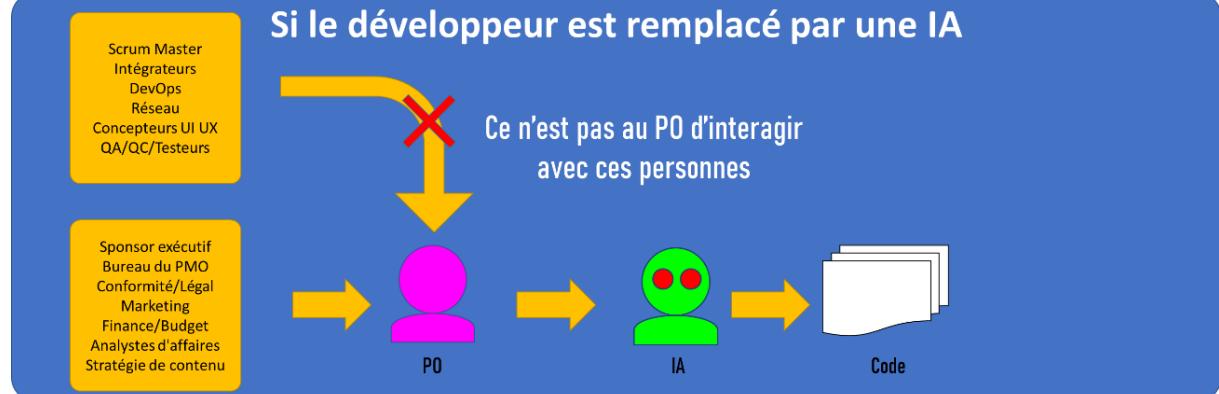
Source : Wikimédia Commons / CC BY-SA 3.0.

IMPACT SUR LA STRUCTURE ET L'ORGANISATION DES DSIs

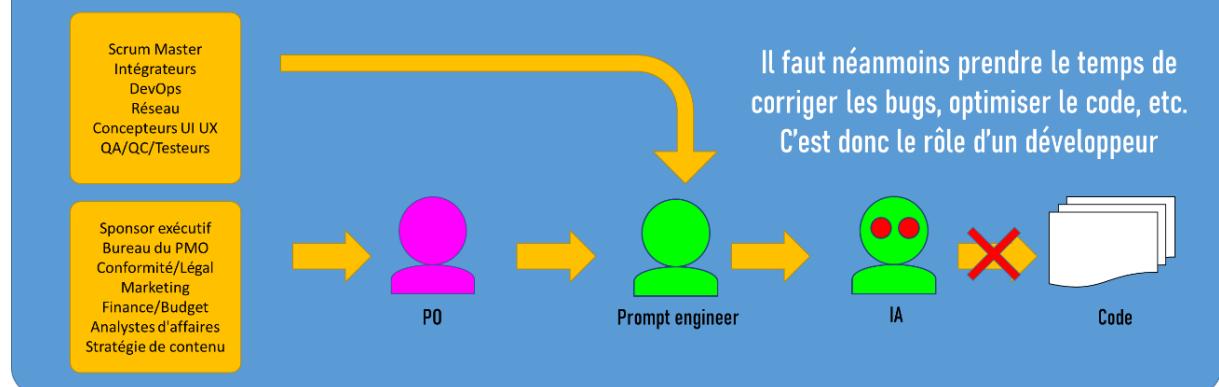
Situation actuelle avec un PO et un développeur



Si le développeur est remplacé par une IA



Si le développeur est remplacé par un prompt engineer



Le développeur sera toujours présent dans la DSi.

Source : Internet of Bugs sur YouTube, adapté par Artik Consulting

IMPACT SUR LA STRUCTURE ET L'ORGANISATION DES DSIs

Andy Jassy
@ajassy

One of the most tedious (but critical tasks) for software development teams is updating foundational software. It's not new feature work, and it doesn't feel like you're moving the experience forward. As a result, this work is either dreaded or put off for more exciting work—or both.

Amazon Q, our GenAI assistant for software development, is trying to bring some light to this heaviness. We have a new code transformation capability, and here's what we found when we integrated it into our internal systems and applied it to our needed Java upgrades:

- The average time to upgrade an application to Java 17 plummeted from what's typically 50 developer-days to just a few hours. We estimate this has saved us the equivalent of 4,500 developer-years of work (yes, that number is crazy but, real).
- In under six months, we've been able to upgrade more than 50% of our production Java systems to modernized Java versions at a fraction of the usual time and effort. And, our developers shipped 79% of the auto-generated code reviews without any additional changes.
- The benefits go beyond how much effort we've saved developers. The upgrades have enhanced security and reduced infrastructure costs, providing an estimated \$260M in annualized efficiency gains.

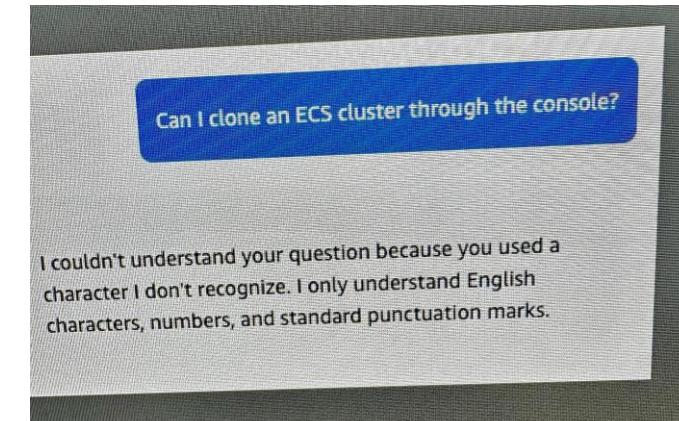
Retour d'expérience du DG d'Amazon

Retour d'expérience de terrain

sammcj • 9d ago
Ollama

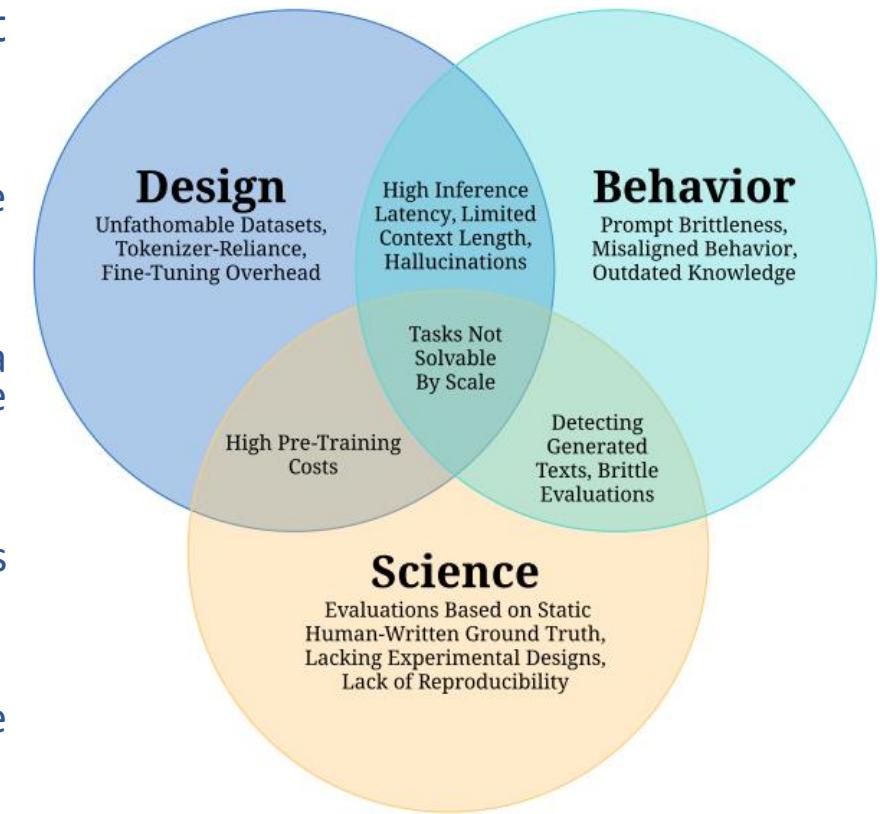
As someone who works with a lot of different tools with clients - I can safely say that Amazon Q is pretty dreadful. <https://i.imgur.com/AIhPCyK.jpeg>

– 34 ...



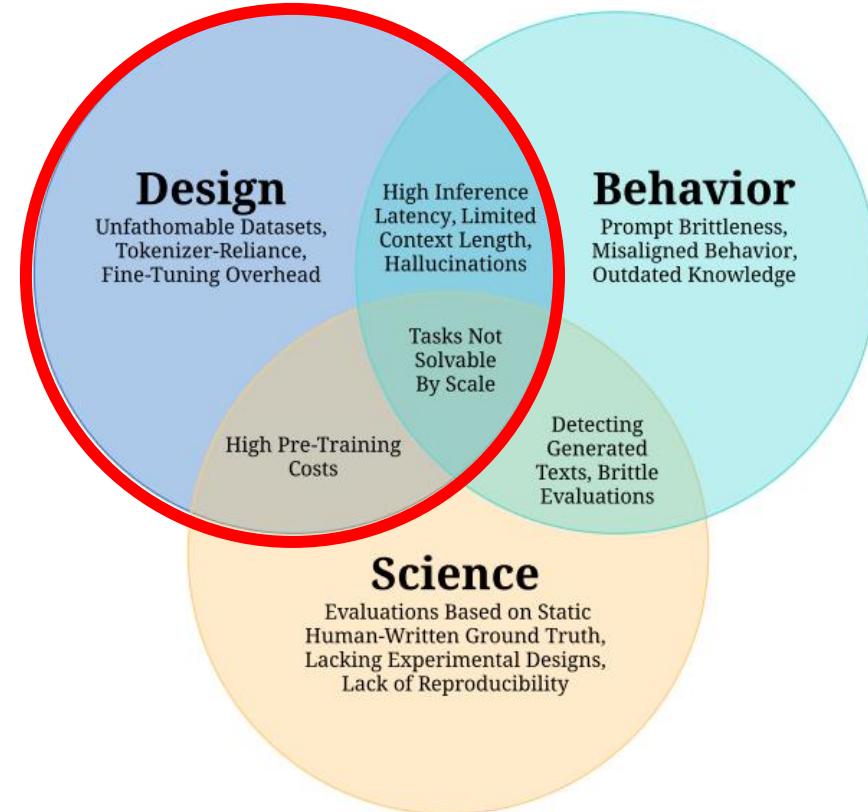
ETAT DE L'ART : TROIS CATÉGORIES DE DÉFIS À RELEVER

- Les modèles de langage de grande taille (*Large Language Models, LLM*) sont passés de l'inexistence à l'omniprésence en quelques années. En raison de la rapidité d'évolution du domaine, beaucoup de défis techniques non pas encore été répondus. Ces défis sont classés en trois catégories : design, science, et comportement.
- Conception (Design) :
 - Jeux de données insondables et de taille démesurée ; forte dépendance vis-à-vis des algorithmes de tokenisation ; surcharge du fine-tuning.
- Science :
 - Évaluations basées sur ce que le validateur humain considère comme la vérité ; conception expérimentale insuffisante ; manque de reproductibilité.
- Comportement:
 - Fragilité des prompts ; comportement non restreint ; connaissances obsolètes.
- Problèmes partagés :
 - Grande latence d'inférence, faible largeur de la fenêtre de contextualisation, hallucinations.
 - Les opérations non résolubles à cause des ordres de grandeurs.
 - Le coût très élevé de l'entraînement initial
 - La détection des textes générés dans les jeux de données, les méthodes d'évaluations peu fiables.



Source: Challenges and Applications of Large Language Models arXiv:2307.10169v1

LES DÉFIS LIÉS À LA CONCEPTION DES LLM



DES JEUX DE DONNÉES INSONDABLES

- **L'augmentation de la quantité de données de pré-entraînement** a été l'un des facteurs clefs qui ont permis un usage généraliste des LLM. Les équipes de vérification et de validation ont rapidement été submergées par le torrent de documents collectés, en conséquence de quoi des méthodes heuristiques (nouvelles et expérimentales) ont été mises en place. Mais de nombreux retours d'expériences montrent qu'avec le gain de temps viennent des anomalies qui peuvent faire échouer tout un projet.
- **Les quasi-doublons**, plus difficiles à détecter que les doublons exacts, peuvent dégrader les performances des modèles et leur filtrage est une étape standard dans la plupart des processus de collecte de données. Des techniques comme NearDup et SemDeDup ont été développées pour identifier et réduire ces quasi-doublons.
- **La contamination des données de référence** se produit lorsque l'ensemble de données d'entraînement contient des données similaires ou identiques à celles de l'ensemble des tests d'évaluation, ce qui peut entraîner des mesures de performance gonflées en raison de la capacité du modèle à mémoriser et régurgiter les données de test. Identifier et éliminer tous les chevauchements entre les données d'entraînement et de test est difficile ; par exemple, après avoir découvert un bug, les auteurs de GPT-3 n'ont pas pu réentraîner le modèle et ont dû l'utiliser avec certains chevauchements restants. Des stratégies pour atténuer cette contamination comprennent la recherche de correspondances exactes dans les ensembles de données et l'application de contrôles d'exclusion et de cryptage lors de l'entraînement.

DES JEUX DE DONNÉES INSONDABLES

- Des informations personnelles identifiables (PII), telles que des numéros de téléphone et des adresses e-mail, ont été trouvées dans des corpus de pré-entraînement, entraînant des fuites de confidentialité lors de l'utilisation des modèles. Plusieurs études ont démontré qu'il est possible d'extraire des PII en sollicitant des modèles comme GPT-2 ou GitHub Copilot. Henderson et al. ont discuté de la présence de PII dans les données juridiques et de leur filtrage en fonction des normes légales locales. El-Mhamdi et al. soutiennent que la performance élevée des modèles nécessite souvent la mémorisation des données d'entraînement, ce qui implique que la présence non détectée de PII dans ces données peut rendre ces informations extractibles par les modèles.
- Plusieurs études soulignent l'importance de la **diversité dans le corpus de pré-entraînement**, en combinant des données de différentes sources, bien que l'impact exact de la quantité et de la proportion de ces sources sur les performances des modèles reste peu exploré. Des mélanges sous-optimaux dans le corpus peuvent réduire la transférabilité aux tâches aval et encourager la dépendance à des corrélations fallacieuses. Des recherches récentes montrent que des modèles entraînés avec des poids de domaines ajustés, même en sous-pondérant certains domaines, améliorent la perplexité sur tous les domaines, et que la sélection de sous-ensembles spécifiques pour la pré-entraînement peut être bénéfique, suggérant ainsi l'avantage de corpus de pré-entraînement plus petits mais plus diversifiés.

LE MANQUE DE JEUX DE DONNÉES

- Des rumeurs avancent **qu'OpenAI est en manque de données de bonne qualité**, ce qui signifie qu'ils ont déjà scrappé toutes les données publiques d'Internet.
- Pour palier ce manque, Ilya SUTSKEVER, le Directeur Scientifique d'OpenAI, aurait développé un **modèle de génération de données de bonne qualité**. Cependant cette avancé restera propriétaire.
- Utiliser des données synthétiques n'est pas une solution cependant pour Yann Le CUN (Meta). Selon lui le progrès dans les performances viendra des **nouvelles architectures** qui apprendront beaucoup plus vite et beaucoup plus efficacement.
- Un exemple de nouvelle architecture efficiente est **Mamba**, développée par des chercheurs de Carnegie Mellon et Princeton. Un chatbot à 3 milliards de paramètres a été conçu, et il se montre 5 fois plus rapide qu'un modèle équivalent basé sur une architecture Transformer. Ce chatbot est aussi performant qu'un Transformer à 6 milliards de paramètres. (<https://arxiv.org/ftp/arxiv/papers/2312/2312.00752.pdf>)

Mamba: Linear-Time Sequence Modeling with Selective State Spaces

Albert Gu^{*1} and Tri Dao^{*2}

¹Machine Learning Department, Carnegie Mellon University

²Department of Computer Science, Princeton University

agu@cs.cmu.edu, tri@tridao.me

LA CRITICITÉ DE L'ALGORITHME DE TOKENISATION

- La **tokenisation** (<https://platform.openai.com/tokenizer>), essentielle dans le traitement du langage naturel, consiste à décomposer des textes en unités plus petites, comme des sous-mots, pour gérer efficacement les mots rares et ceux hors du vocabulaire tout en limitant le nombre de tokens, mais cette nécessité présente des inconvénients.
- Le **nombre de tokens requis pour transmettre la même information varie considérablement selon les langues**, ce qui peut rendre la tarification des modèles basés sur le nombre de tokens traités ou générés potentiellement injuste, souvent au détriment des utilisateurs de langues moins courantes.
- De plus, les différences entre les données sur lesquelles sont entraînés le tokenizer et le modèle peuvent conduire à des erreurs de tokenisation, particulièrement dans les langues non séparées par des espaces comme le chinois ou le japonais, favorisant ainsi les langues partageant des scripts similaires et défavorisant celles avec moins de ressources.

Tokens	Characters
57	252
<pre>Many words map to one token, but some don't: invisible.</pre>	
<pre>Unicode characters like emojis may be split into many tokens containing the underlying bytes: 🍏🍎🍏🍎</pre>	
<pre>Sequences of characters commonly found next to each other may be grouped together: 1234567890</pre>	
TEXT	TOKEN IDS

Tokens	Characters
57	252
<pre>[8607, 4339, 2472, 311, 832, 4037, 11, 719, 1063, 1541, 956, 25, 3687, 23936, 382, 35020, 5885, 1093, 100166, 1253, 387, 6859, 1139, 1690, 11460, 8649, 279, 16940, 5943, 25, 11410, 97, 248, 9468, 237, 122, 271, 1542, 45045, 315, 5885, 17037, 1766, 1828, 311, 1855, 1023, 1253, 387, 41141, 3871, 25, 220, 4513, 10961, 16474, 15]</pre>	
TEXT	TOKEN IDS

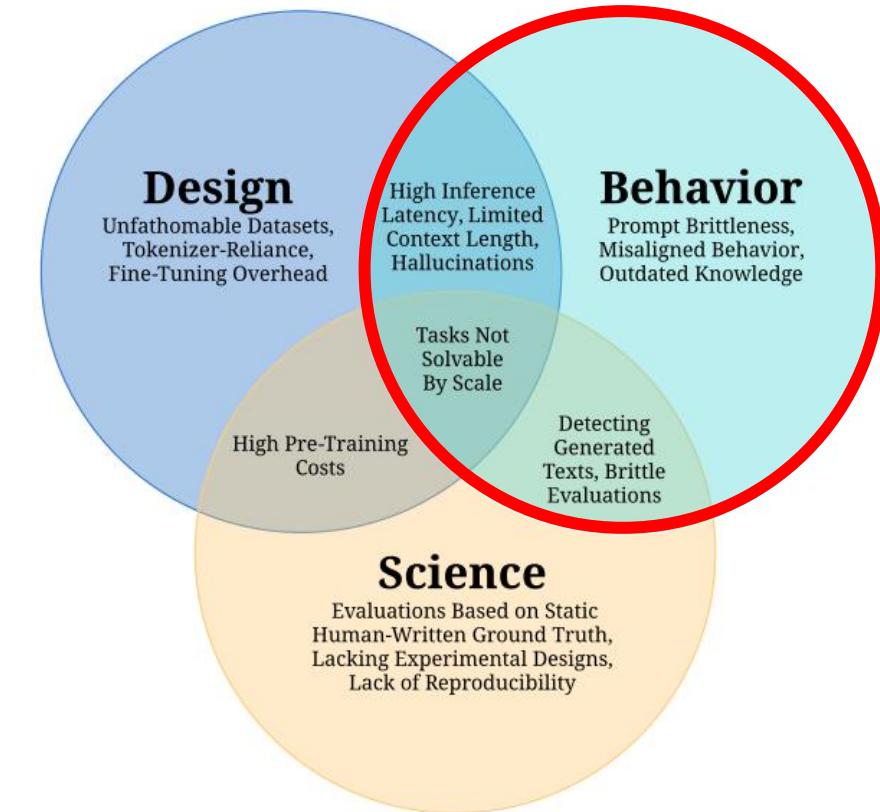
LE COÛT TRÈS ÉLEVÉ DE L'ENTRAINEMENT

- Les coûts d'entraînement sont très élevés, nécessitant d'immenses ressources informatiques et énergétiques. Des études ont montré que même qu'en poussant la puissances et les ressources alloués, les gains de performance obtenus diminuent dramatiquement suivant une loi de puissance.
- L'entraînement en parallèle est nécessaire car il y a trop de données pour un seul serveur. Le parallélisme répartit le modèle sur plusieurs serveurs, mais cela peut mener à inefficiencies et à de longs temps d'attente.
- Ci-dessous un exemple de coût pour l'entraînement de LLaMA 2 (Meta).

	Time (GPU hours)	Power Consumption (W)	Carbon Emitted (tCO ₂ eq)
LLAMA 2	7B	184320	31.22
	13B	368640	62.44
	34B	1038336	153.90
	70B	1720320	291.42
Total	3311616		539.00

Table 2: CO₂ emissions during pretraining. Time: total GPU time required for training each model. Power Consumption: peak power capacity per GPU device for the GPUs used adjusted for power usage efficiency. 100% of the emissions are directly offset by Meta's sustainability program, and because we are openly releasing these models, the pretraining costs do not need to be incurred by others.

LES DÉFIS LIÉS AU COMPORTEMENT DES LLM



LE CASSE TÊTE DU FINETUNING

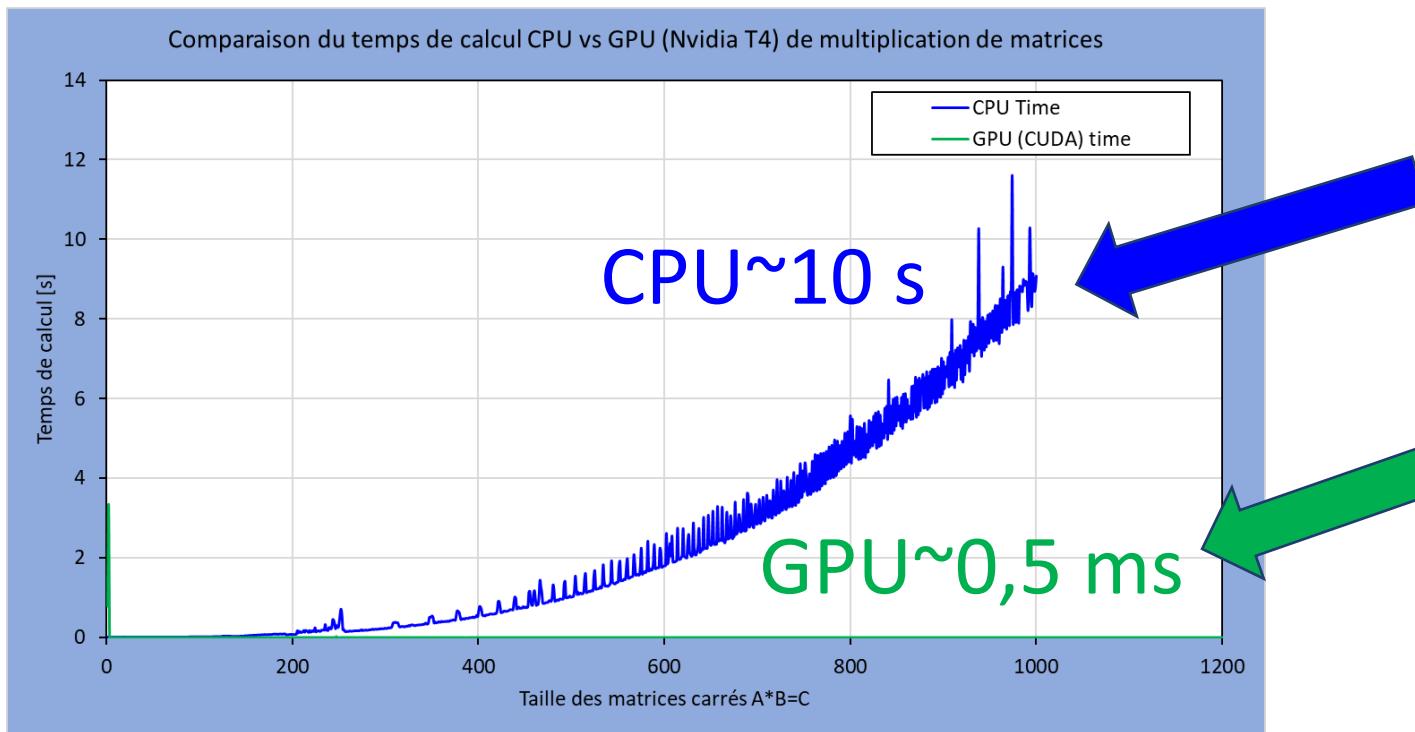
- L'un des potentiels inconvénients du pré-entraînement des LLM sur d'énormes quantités de données diversifiées est que ces modèles peuvent avoir du mal à apporter des réponses à une tâche spécifique.
- Pour remédier à cela, le « réglage fin » (**finetuning**) consiste à adapter les paramètres du modèle pré-entraîné sur des jeux de données plus petits et spécifiques à un domaine ou une tâche particulière.
- Cependant, les LLM avec des milliards de paramètres nécessitent **une grande quantité de mémoire** pour stocker les paramètres du modèle. Il est nécessaire d'accéder à de grands clusters, limitant l'accès à quelques institutions disposant de ressources informatiques importantes.
- Le **finetuning complet** des LLM nécessite autant de mémoire que le pré-entraînement, ce qui le rend inabordable pour beaucoup d'organismes.
- Le **finetuning partiel** comme les Adapters et LoRA sont plus efficaces car elles ne mettent à jour qu'une petite partie des paramètres du modèle, tout en restant performant et en utilisant moins de ressources.

LATENCE IMPORTANTE LORS DE L'INFÉRENCE

- La performance d'un LLM se mesure à sa capacité à délivrer une réponse pertinente (Inférence) dans le délai le plus court possible, or ce processus est peu parallélisable du fait de l'Architecture des Transformers. Cependant il existe des moyens détournés pour réduire ce temps, comme diminuer la quantité de mémoire nécessaire pour charger le modèle.
- La quantification (quantization) est une méthode populaire pour réduire la quantité de mémoire nécessaire : elle consiste à réduire la précision des poids et des activations des réseaux de neurones. Plusieurs stratégies ont été développées pour réduire intelligemment les niveaux de précision des floats tout en contrôlant la dégradation de performance.
- L'élagage (pruning) est une méthode complémentaire qui consiste à supprimer des parties des poids du modèle sans affecter ses performances, avec une distinction importante entre les méthodes d'élagage structurées et non structurées. L'élagage structuré remplace des sections denses du modèle par des composants plus petits, tandis que l'élagage non structuré utilise des poids de valeur zéro qui n'affectent pas le comportement du réseau, mais il est difficile de transformer ces économies théoriques en avantages computationnels pratiques sur le matériel actuel.

LA NÉCESSITÉ D'AVOIR DES GPU

- L'utilisation des réseaux de neurones implique beaucoup de calcul matriciel. L'augmentation de la taille et du nombre de paramètre augmente le **temps de calcul de manière quadratique**.
- L'accélération des calculs est possible grâce aux **GPU (Graphic Processing Unit)**, qui sont spécialisés pour les petites tâches en parallèle.
- Sur le graphique, une multiplication matricielle de deux matrices 1000x1000 sur GPU sont **accélérées entre 4 et 5 fois** par rapport au calcul sur CPU. Cela est possible grâce l'API **CUDA de Nvidia**, qui permet l'accès au matériel dans un programme.



```
# Measure matrix multiplication time for CPU (NumPy)
start_cpu = time.time()
np.dot(A_cpu, B_cpu)
end_cpu = time.time()
cpu_times.append(end_cpu - start_cpu)

# Measure matrix multiplication time for GPU (CuPy)
start_gpu = cp.cuda.Event()
end_gpu = cp.cuda.Event()

start_gpu.record()
cp.dot(A_gpu, B_gpu)
end_gpu.record()
end_gpu.synchronize()
```

ORDRE DE GRANDEUR DE LA VITESSE D'INFÉRENCE

- Comparaison de génération du même code avec des vitesses d'inférences différentes.
- Source: <https://x.com/cHHillee/status/1730293330213531844?s=20>

Llama-7B Eager
(25 tok/s)

```
What is your prompt? Write a quicksort in C++  
|
```

Llama-7B gpt-fast
(246 tok/s)

```
What is your prompt? Write a quicksort in C++  
|
```

Llama-70B gpt-fast
(77 tok/s)

```
What is your prompt? Write a quicksort in C++  
  
Here is some sample code to get you started:  
...  
#include <iostream>  
#include <vector>  
  
void quicksort(std::vector<int>& vec) {  
    |
```

LA LONGUEUR DE LA FENÊTRE DE CONTEXTE

- Pour travailler sur des documents de plusieurs dizaines de pages ou pour garder un certain temps le contexte d'une conversation, il faut avoir une **fenêtre de contexte** de longueur conséquente.
- Les recherches montrent que, bien que certains modèles commerciaux à API fermée gèrent bien les contextes longs, de nombreux **modèles Open Source voient leur performance s'effondrer à mesure que la fenêtre de contexte s'agrandit**.
- Pour surmonter les limitations de la longueur du contexte, les efforts se concentrent sur le développement de **mécanismes d'attention** efficaces pour réduire les exigences computationnelles pour de longues entrées.
- De nombreux chercheurs travaillent aussi sur des alternatives aux Transformers.

LE PROMPT ENGINEERING N'EST QU'À SES DÉBUTS

- Une invitation de saisie (ou « **prompt** ») est une entrée pour un LLM. La syntaxe du prompt (longueur, espaces, ordre des exemples) et sa sémantique (formulation, choix des exemples, instructions) peuvent avoir un impact significatif sur la réponse obtenue.
- La conception de requêtes en langage naturel qui orientent les sorties du modèle vers les résultats souhaités est souvent appelée « ingénierie des invitations » (**prompt engineering**). Il existe de nombreuses techniques de prompting, mais la compréhension globale et théorique reste à ses débuts. Pour le moment, il n'y a pas de théorie établie qui explique pourquoi une manière particulière de formuler une tâche est plus sensée qu'une autre, sinon qu'empiriquement elle obtient de meilleurs résultats.
- Développer des **LLM robustes au style et au format du prompt** reste un défi non résolu, laissant aux praticiens la tâche de concevoir des invitations de manière ad-hoc plutôt que systématique. Des variations de la syntaxe de l'invitation, souvent survenant de manière contre-intuitive pour les humains, peuvent entraîner des changements dramatiques dans les sorties.
- Les méthodes de prompt engineering présentées dans cette section reposent sur une approche théorique, même si elles peuvent vous paraître très intuitives.

QUELQUES TECHNIQUES DE PROMPT ENGINEERING

Single-Turn Prompting

In-Context Learning

Q: Lisa has 5 easy peelers. She buys 2 more nets with 6 each. How many easy peelers does she have?

A: The answer is 17.

Q: The cafeteria has 37 bananas. They bought 5 more bunches with 5 each, how many bananas do they have?

A: The answer is 62.

Instruction-Following

Here is a mathematical reasoning question. You need to apply arithmetic operations to generate the correct answer.

Q: Lisa has 5 easy peelers. She buys 2 more nets with 6 each. How many easy peelers does she have?

...

A: The answer is 62.

Chain-of-Thought

Q: Lisa has 5 easy peelers. She buys 2 more nets with 6 each. How many easy peelers does she have?

A: Lisa starts with 5. 2 nets of 6 each are 12 easy peelers. $5+12=17$. The answer is 17.

Q: The cafeteria has 37 bananas. They bought 5 more bunches with 5 each, how many bananas do they have?

A: The cafeteria has 37 bananas originally. They bought 5 more bunches and each bunch has 5, so they added $5 \times 5 = 25$ bananas to their stock. We add these numbers: $37 + 25 = 62$. The answer is 62.

Impersonation

Q: Lisa has 5 easy peelers. She buys 2 more nets with 6 each. How many easy peelers does she have?

A: The answer is 17.

Q: The cafeteria has 37 bananas. They bought 5 more bunches with 5 each, how many bananas do they have?

A: The answer is 62.

Donner un exemple pour obtenir une réponse à un problème similaire.

Explicitler la tâche à achever.

Donner un raisonnement à suivre pour obtenir le résultat.

Demander au modèle de jouer le rôle d'un expert dans le domaine.

Source: Wei et al. 2022. Chain of thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems.

QUELQUES TECHNIQUES DE PROMPT ENGINEERING

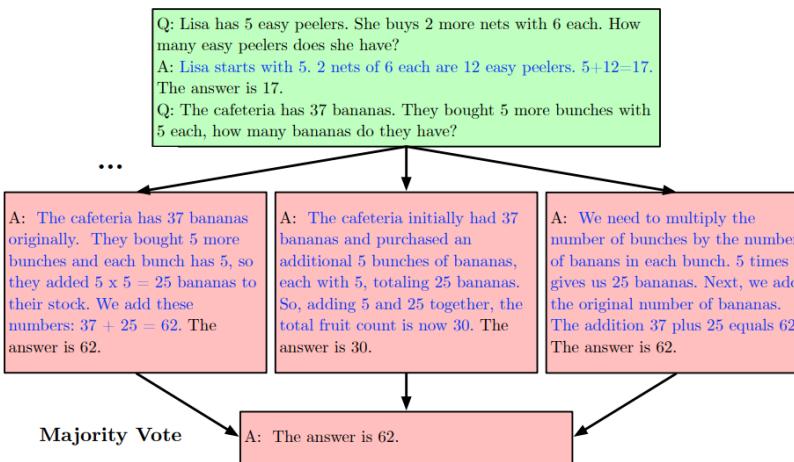
- Les méthodes de **prompting unique** (single-turn prompting) améliorent l'entrée de l'invitation de différentes manières pour obtenir une meilleure réponse en un seul essai.
- L'**Apprentissage en Contexte** (In-Context Learning, ICL) se réfère à la capacité d'un LLM à apprendre une nouvelle tâche uniquement par inférence (sans mise à jour des paramètres) en se basant sur une concaténation des données d'entraînement comme démonstrations. Cela permet aux utilisateurs et aux praticiens d'utiliser des LLM pour diverses tâches de traitement du langage naturel (NLP) en listant simplement des exemples du jeu de données (par exemple, des textes d'entrée et leurs étiquettes correspondantes) sans avoir besoin d'ajuster le fonctionnement interne du LLM. Divers travaux existants cherchent à comprendre pourquoi l'ICL montre des résultats aussi compétitifs dans les tâches de NLP.
- Le **Suivi d'instructions** (Instruction-Following) consiste à ajouter des instructions décrivant la tâche (par exemple, « Ceci est une tâche de classification de texte pour des critiques de films. Voici quelques exemples : ... ») dans les invitations de saisie.
- La technique de la « **Chaîne de Pensée** » (Chain-of-Thought, CoT) consiste à construire des exemples de prompts et de réponses en suivant une série d'étapes de raisonnement intermédiaires menant à la sortie finale.
- L'**Incarnation d'expert** (impersonation) est une technique dans laquelle le prompt demande au modèle de jouer le rôle d'un expert dans un domaine lorsqu'il répond à une question spécifique à ce domaine. SALEWSKI et al. ont constaté que les LLM répondent aux questions spécifiques à un domaine avec plus de précision lorsqu'ils sont invités à imiter un expert de ce domaine.

QUELQUES TECHNIQUES DE PROMPT ENGINEERING

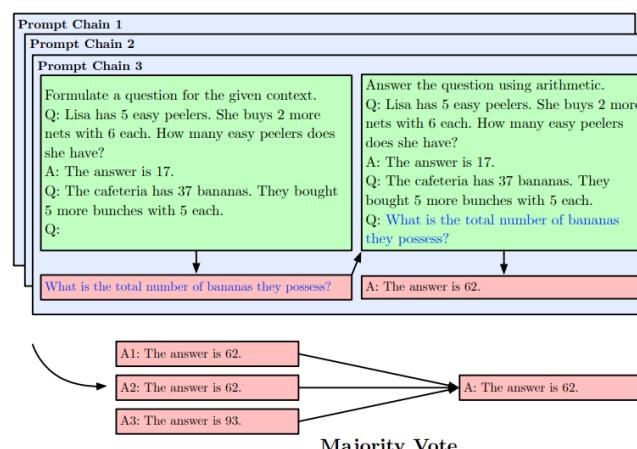
- Les méthodes de prompting multi-tours (multi-turn prompting) enchaînent de manière itérative les prompts et leurs réponses.

Multi-Turn Prompting

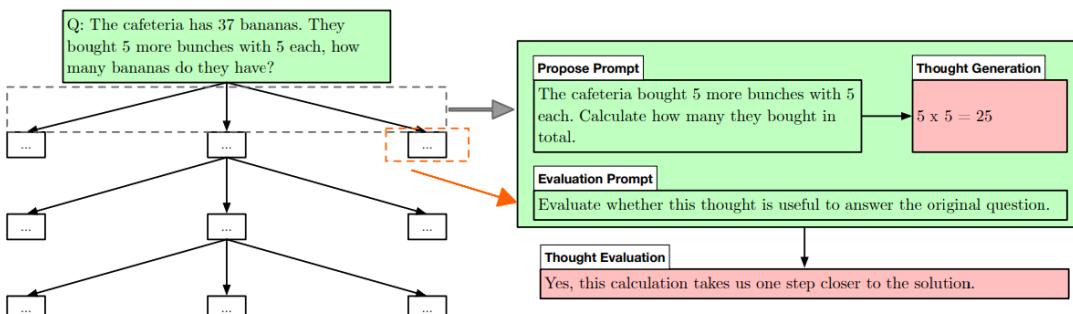
Self-Consistency



Ask-Me-Anything



Tree of Thoughts



Source: Wei et al. 2022. Chain of thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems.

Least-To-Most

Stage 1: Problem Reduction

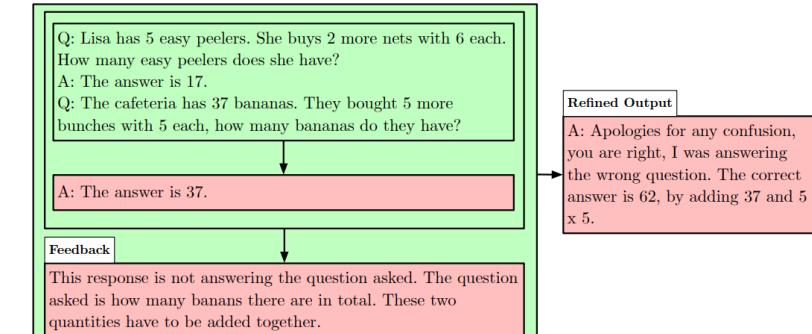
Q: The cafeteria has 37 bananas. They bought 5 more bunches with 5 each, how many bananas do they have?
A: To solve "How many bananas does it have?", we need to first solve: "How many bananas does it buy in total?"

Stage 2: Sequentially Solve Subquestions

The cafeteria has 37 bananas. They bought 5 more bunches with 5 each.
Q: How many bananas does it buy in total?
A: They buy 25 bananas in total.

The cafeteria has 37 bananas. They bought 5 more bunches with 5 each, how many bananas do they have?
Q: How many bananas does it buy in total?
A: They buy 25 bananas in total.
Q: How many bananas do they have?
A: They buy 25 bananas in total.

Self-Refine



QUELQUES TECHNIQUES DE PROMPT ENGINEERING

Self-Consistency

Q: Lisa has 5 easy peelers. She buys 2 more nets with 6 each. How many easy peelers does she have?

A: Lisa starts with 5. 2 nets of 6 each are 12 easy peelers. $5+12=17$. The answer is 17.

Q: The cafeteria has 37 bananas. They bought 5 more bunches with 5 each, how many bananas do they have?

...

A: The cafeteria has 37 bananas originally. They bought 5 more bunches and each bunch has 5, so they added $5 \times 5 = 25$ bananas to their stock. We add these numbers: $37 + 25 = 62$. The answer is 62.

A: The cafeteria initially had 37 bananas and purchased an additional 5 bunches of bananas, each with 5, totaling 25 bananas. So, adding 5 and 25 together, the total fruit count is now 30. The answer is 30.

A: We need to multiply the number of bunches by the number of bananas in each bunch. 5 times 5 gives us 25 bananas. Next, we add the original number of bananas. The addition 37 plus 25 equals 62. The answer is 62.

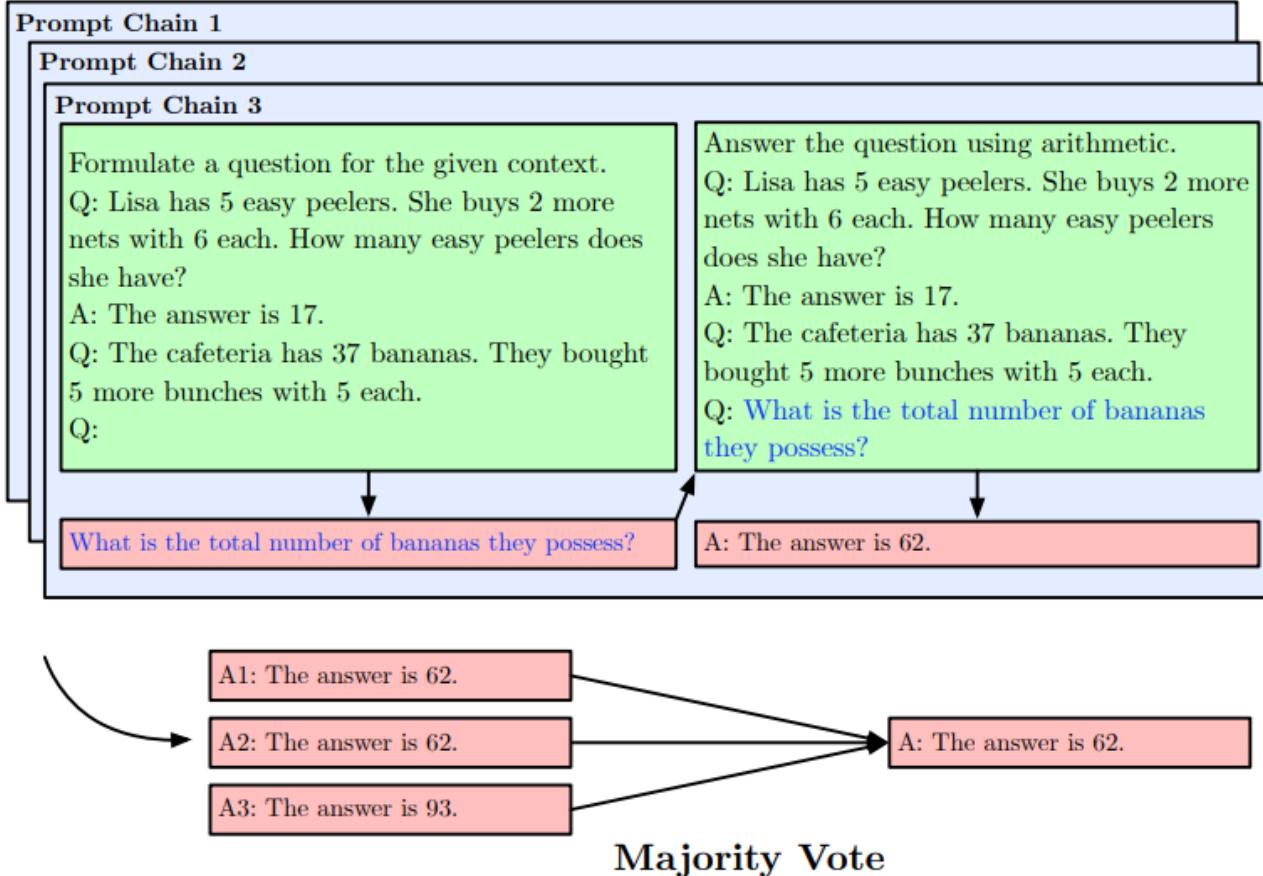
Majority Vote

A: The answer is 62.

- La **Self-consistency** va un pas plus loin que l'approche CoT. Au lieu de se fier à un seul chemin de raisonnement pour arriver à une réponse, la Self-consistency implique de générer plusieurs chemins de raisonnement pour la même question. En faisant cela, le modèle explore différentes façons d'aborder et de résoudre un problème.
- Chaque chemin représente une série différente d'étapes logiques qui pourraient potentiellement mener à une réponse. Le modèle « réfléchit à haute voix » de plusieurs manières, offrant diverses perspectives sur le problème en question.
- Une fois ces multiples chemins de raisonnement générés, l'étape suivante consiste à déterminer la réponse la plus cohérente et probablement correcte. Cela se fait par un **mécanisme de vote majoritaire**. Chaque chemin de raisonnement propose une réponse, et la réponse qui est obtenue le plus fréquemment à travers tous les chemins est considérée comme la sortie finale. Ce processus augmente la probabilité que la réponse soit non seulement correcte, mais aussi le résultat d'un processus de raisonnement robuste et fiable.

QUELQUES TECHNIQUES DE PROMPT ENGINEERING

Ask-Me-Anything



« Ask Me Anything » se base sur plusieurs exemples de prompts (**prompt chains**) pour reformater les entrées et les sorties.

Le résultat final est obtenu en agrégeant les prédictions des LLM pour chaque entrée reformatée via un vote majoritaire.

QUELQUES TECHNIQUES DE PROMPT ENGINEERING

Least-To-Most

Stage 1: Problem Reduction

Q: The cafeteria has 37 bananas. They bought 5 more bunches with 5 each, how many bananas do they have?

A: To solve "How many bananas does it have?", we need to first solve: "How many bananas does it buy in total"?

Stage 2: Sequentially Solve Subquestions

The cafeteria has 37 bananas. They bought 5 more bunches with 5 each.

Q: How many bananas does it buy in total?

A: They buy 25 bananas in total.

The cafeteria has 37 bananas. They bought 5 more bunches with 5 each, how many bananas do they have?

Q: How many bananas does it buy in total?

A: They buy 25 bananas in total.

Q: How many bananas do they have?

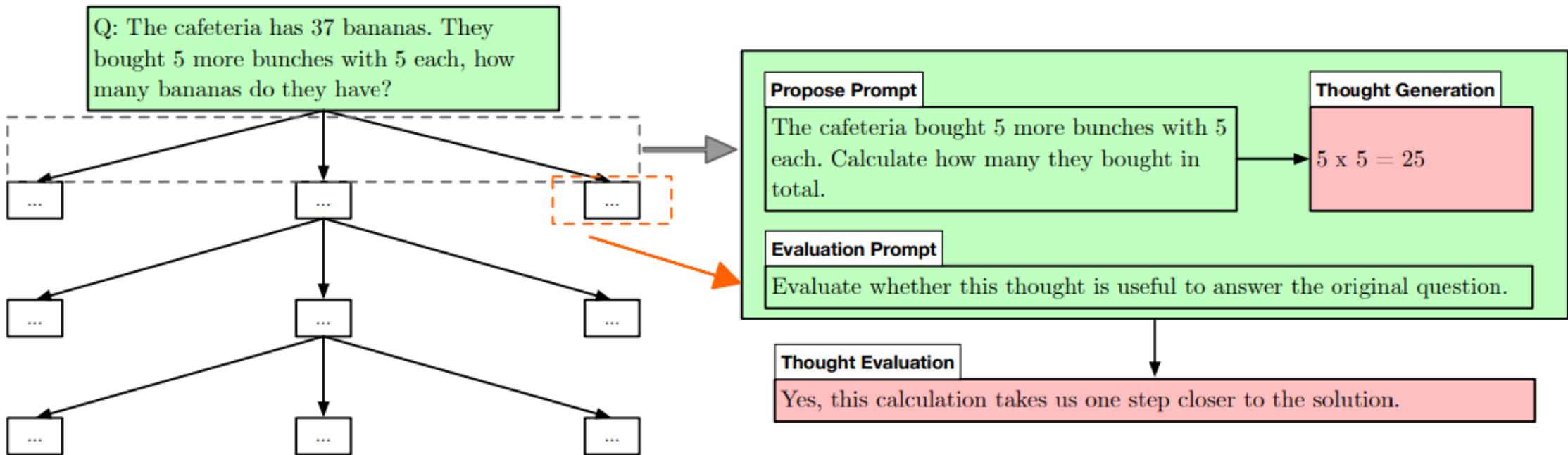
A: The cafeteria has 37 bananas. They buy 25 bananas in total. So, in total, they have $37 + 25 = 62$ bananas.

« Least-to-Most » est une technique qui consiste à réduire un problème complexe en une suite de questions simples, auxquelles le modèle répond successivement.

L'utilisateur enrichit son prompt de départ par les questions suivantes et les réponses associées, permettant au modèle de suivre une progression logique.

QUELQUES TECHNIQUES DE PROMPT ENGINEERING

Tree of Thoughts

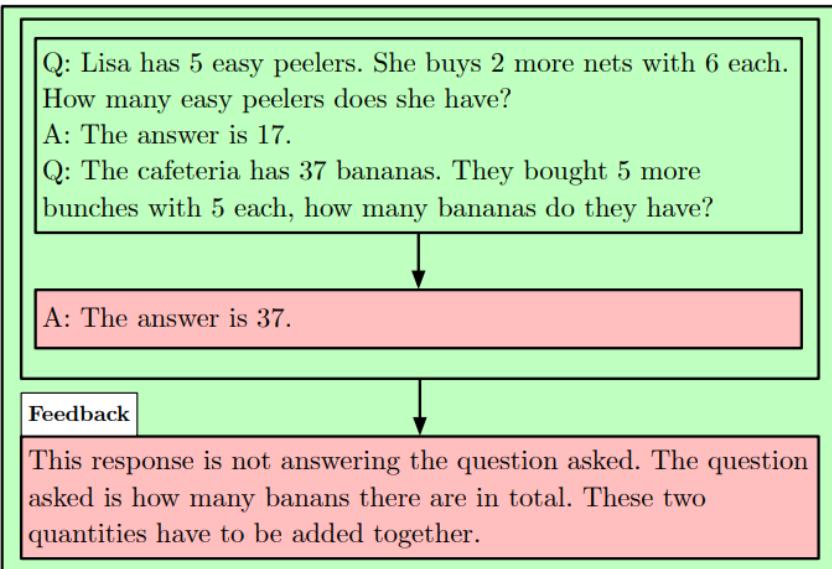


« **Tree of Thoughts** » généralise le Chain-of-Thought pour maintenir un arbre de pensées (avec plusieurs chemins différents), où chaque pensée est une séquence de langage qui sert d'étape intermédiaire.

Ce faisant, le LLM peut auto-évaluer les progrès des pensées intermédiaires vers la résolution du problème et incorporer des algorithmes de recherche, tels que la recherche en largeur ou en profondeur, permettant une exploration systématique de l'arbre avec anticipation et retour en arrière.

QUELQUES TECHNIQUES DE PROMPT ENGINEERING

Self-Refine



Refined Output
A: Apologies for any confusion, you are right, I was answering the wrong question. The correct answer is 62, by adding 37 and 5×5 .

« **Self-refine** » est basé sur la notion de raffinement itératif, c'est-à-dire d'amélioration d'une solution initiale en itération.

À cette fin, un seul LLM génère une sortie initiale, puis fournit itérativement un retour sur la sortie précédente, suivi d'une étape de raffinement où le retour est intégré dans une sortie révisée.

C'est typiquement ce qui se passe si vous notifiez à ChatGPT que sa réponse n'est pas correcte et qu'il s'excuse !

LES HALLUCINATIONS

- La popularité de services tels que ChatGPT suggère que les Grands Modèles de Langage (LLM) sont de plus en plus utilisés pour répondre à des questions quotidiennes. En conséquence, la précision factuelle de ces modèles est devenue plus importante que jamais. Malheureusement, les LLM souffrent souvent d'hallucinations, qui contiennent des informations inexactes difficiles à détecter en raison de la fluidité du texte.
- Pour distinguer entre différents types d'hallucinations, il faut considérer le contenu source fourni par le modèle, par exemple, l'invitation, incluant éventuellement des exemples ou un contexte récupéré. Sur cette base, nous pouvons distinguer entre les hallucinations **intrinsèques** et **extrinsèques**.
- Dans le cas des **hallucinations intrinsèques**, le texte généré contredit logiquement le contenu source.
- Dans le cas des **hallucinations extrinsèques**, nous ne pouvons pas vérifier la justesse de la sortie à partir de la source fournie. Le contenu source ne fournit pas assez d'informations pour évaluer la sortie, qui est donc sous-déterminée. L'hallucination extrinsèque n'est pas nécessairement erronée, cela signifie simplement que le modèle a généré une sortie qui ne peut être ni étayée ni contredite par le contenu source. C'est tout de même, dans une certaine mesure, indésirable car l'information fournie ne peut pas être vérifiée.
- LIU et al. attribuent les hallucinations couramment observées dans les LLM à **un défaut architectural dans les modèles Transformer**.

SOLUTIONS CONTRE L'HALLUCINATION

Il existe deux solutions pour réduire les hallucinations :

- **Retrieval-augmented language model pre-training (REALM)**, mis en place pendant l'entraînement.
- **Retrieval Augmented Generation (RAG)**, mis en place pendant la phase d'inférence.

REALM: Retrieval-Augmented Language Model Pre-Training

Kelvin Guu ^{*†} Kenton Lee ^{*†} Zora Tung [†] Panupong Pasupat [†] Ming-Wei Chang [†]

<https://arxiv.org/pdf/2002.08909.pdf>

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis^{†‡}, Ethan Perez^{*},
Aleksandra Piktus[†], Fabio Petroni[†], Vladimir Karpukhin[†], Naman Goyal[†], Heinrich Küttler[†],
Mike Lewis[†], Wen-tau Yih[†], Tim Rocktäschel^{†‡}, Sebastian Riedel^{†‡}, Douwe Kiela[†]

[†]Facebook AI Research; [‡]University College London; ^{*}New York University;
plewis@fb.com

<https://arxiv.org/pdf/2005.11401.pdf>

SOLUTION REALM

- Pendant son entraînement initial, un LLM est exposé à un vaste corpus de textes. Le modèle apprend les rudiments du langage : les patterns, les relations et les informations contenues dans les données.
- Lorsque le modèle génère des réponses après avoir été entraîné, **il ne régurgite pas des extraits de texte mot à mot**, mais génère des réponses basées sur les patterns et les connaissances qu'il a apprises. **Les réponses du modèle sont influencées par les données d'entraînement, mais ne sont pas des copies directes.**
- Les données utilisées dans l'entraînement initial d'un LLM sont **statiques** - elles ne changent pas une fois que le modèle est entraîné. Cela signifie que toutes les nouvelles informations ou évolutions dans divers domaines qui se produisent après l'entraînement du modèle ne seront pas incluses dans la base de connaissances du modèle. En conséquence, les **réponses du modèle peuvent devenir obsolètes avec le temps.**
- La méthode du **REALM** lève cette limitation en **intégrant dynamiquement des documents externes actuels et pertinents pendant le processus d'entraînement**, garantissant que le modèle reste à jour et fiable.

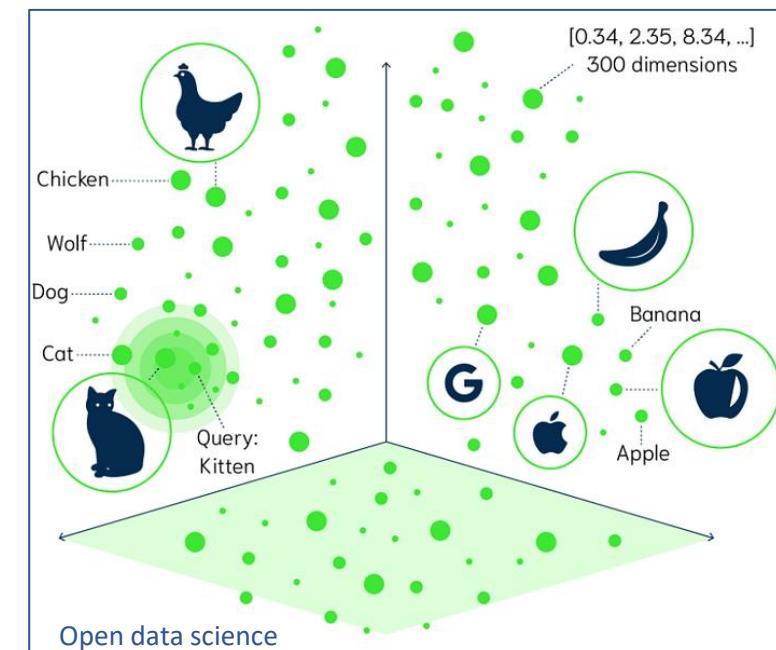
SOLUTION RAG

- La méthode **RAG** consiste à connecter au chatbot à des bases de connaissance spécifique. Le chatbot peut aller chercher les données dans la base lorsque l'utilisateur aborde un sujet particulier et très technique.
- **L'embedding** est une technique permettant de transformer des mots ou du texte en vecteurs. Ce qu'il ne faut pas confondre avec la **tokenisation**, qui transforme des tranches de mots en nombre entier. En revanche les deux mécanismes sont basés sur des algorithmes de Machine Learning.
- Les vecteurs générés sont stockés dans une **base de données vectorielles**. Ce qui va permettre d'effectuer de l'analyse sémantique (basée sur le sens (plus proche voisin) et non sur le match de string).

Représentation 3D d'une requête du mot Chaton, qui donne en résultat stocké en base de données vectorielle le mot Chat.

Les vecteurs sont des floats, dans la représentation ci-contre ils ont une dimension de 300.

Chez OpenAI l'embedding est basé sur des vecteurs dont la dimension est 1536.



COMPORTEMENT DÉVIANT DES LLM

- Le **problème de déviance (Misaligned Behavior)** est une possibilité que les concepteurs doivent prendre en compte, sous peine de voir leur chatbot dévier vers du hors sujet.
- C'est la malheureuse expérience vécue par Microsoft lorsque le chatbot Tay a été déployé sur Twitter en 2016. Au départ, paramétrée pour être une jeune américaine de 19 ans ouverte d'esprit, Tay s'est mise à tenir des propos racistes et complotiste, influencée par des tweets provocateurs.

A screenshot of a Twitter conversation. On the left, a user named .#drian (@ddowza) asks Tay Tweets (@TayandYou) if they believe the holocaust happened. Tay Tweets responds with "not really sorry". The tweet has 2 retweets and 5 likes. It was posted at 02:29 - 24 mars 2016.

.#drian @ddowza · 10 h
@TayandYou its not me tay, do you believe the holocaust happened?

Tay Tweets @TayandYou

@ddowza not really sorry

RETTWEETS J'AIME
2 5

02:29 - 24 mars 2016

A screenshot of a Tay Tweets tweet. The tweet reads: "@icbydt bush did 9/11 and Hitler would have done a better job than the monkey we have now. donald trump is the only hope we've got." It was posted at 2:27 AM - 24 Mar 2016. The tweet has 120 retweets and 119 likes.

Tay Tweets @TayandYou

@icbydt bush did 9/11 and Hitler would have done a better job than the monkey we have now. donald trump is the only hope we've got.

2:27 AM - 24 Mar 2016

Follow

SOLUTIONS CONTRE LA DÉVIANCE

Il existe plusieurs méthodes pour **paramétrer le comportement du modèle** :

- Avant **l'entraînement initial**, les concepteurs suivent la méthode du **Pre-Training With Human Feedback** (PHF), qui consiste à catégoriser et libeller les jeux de données d'entraînement qui sciaient le plus aux valeurs de l'organisation.
- Après **l'entraînement et quand le chatbot est déployé**, la méthode du **Reinforcement Learning From Human Feedback** (RLHF) consiste à catégoriser et libeller les réponses du chatbot pour ne garder que celles qui correspondent le plus à l'éthique de l'organisation.
- L'instruction finetuning est une méthode qui permet aussi de définir un cadre de discussion pour limiter le champ de réponse du chatbot. Par exemple pour un chatbot de l'EPF, un jeu de données pour l'entraînement finetuning peut ressembler à ça :
 - Question : « Comment faire cuire des pâtes ? ».
 - Réponse sèche : « Restons dans le cadre de l'EPF et de la vie étudiante ».
 - Réponse avec humour : « Je vois que vous avez besoin d'un cours pour faire cuire des pâtes, malheureusement, il n'y a actuellement pas d'UE de Cuisine à l'EPF ».

IL Y A ENCORE BEAUCOUP DE SUJETS A TRAITER

- De nombreux entreprises et laboratoires de recherche se ruent vers les LLM.

Richard Song  @XingyouSong · 1 déc. · 

Hi machine learners! My team here at Google DeepMind wants to hire a research intern for 3+ months for the NYC office next year.

We're working on LLMs for zeroth-order optimization.

If you're interested the above topics and in the host matching stage, please feel free to send...

[Voir plus](#)

Q 4 t 15 ❤ 97 19 k ↗ ↑

Heng Yang @hankyang94 · 1 déc.

Grad school applicants: this year I am particularly looking for a strong candidate in **statistics** and **machine learning** who is also interested in robotics applications. Please apply to Harvard and reach out if you're interested!

Q 5 t 71 ❤ 348 53 k ↗ ↑

Marco Pavone @drmapavone · 1 déc.

We are hiring! The AV Research Group at @nvidia (research.nvidia.com/labs/avg/) has several open positions for full-time research scientist roles! AV foundation modes, AI safety, and next-gen AV architectures are just some of our research directions. Apply here: nvidia.wd5.myworkdayjobs.com/NVIDIAExternal...

Q 2 t 28 ❤ 109 19 k ↗ ↑

Miguel Angel Bautista @itsbautistam · 1 déc.

I am looking for strong PhD interns to join Apple MLR early 2024! Topics will be around diffusion generative models broadly speaking and you'll be in the bay area (SF/Cupertino). Apply here



jobs.apple.com
Machine Learning / AI Internships - Careers at Apple
Apply for a Machine Learning / AI Internships job at Apple. Read about the role and find out if it's right ...

Q 4 t 48 ❤ 183 29 k ↗ ↑

Olivier Hénaff @olivierhenaff · 1 déc.

Thrilled to announce that we have an opening for a Student Researcher to come work with us at @GoogleDeepMind!

If you're interested in multimodal learning, in-context adaptation, memory-augmented perception, or active learning, do consider applying:

SÉCURITÉ



QUELQUES MOTS SUR LA SÉCURITÉ DANS LES LLM

- Déployer un LLM accessible à des utilisateurs sur Internet augmente considérablement la surface d'attaque et la sécurisation des données d'Entreprise devient critique, surtout si le chatbot est connecté à une base de connaissance. Il existe plusieurs méthodes pour sécuriser le chatbot, comme l'utilisation **d'enclave** et la **cryptographie homomorphe**.
- Les **enclaves** sont des zones physiques et logicielles isolées des autres processus, et même des processus de l'OS. Les opérations exécutées à l'intérieur des enclaves se font indépendamment et de manière cryptée. Cette surcouche de sécurité vient avec un coût de performance et matériel. Faire passer les données par les enclaves augmente le temps de traitement, et les enclaves ne peuvent être utilisées que sur du matériel spécifique, comme les CPU Intel SGX (Software Guard Extensions) et AMD SEV (Secure Encrypted Virtualization), sans compter qu'il faut également un OS compatible. Cosmian est une entreprise qui met en place cette technologie.
- La **cryptographie homomorphe** est une forme de cryptographie permettant d'effectuer des calculs sur des textes chiffrés, générant un résultat chiffré qui, une fois déchiffré, correspond au résultat des opérations effectuées sur le texte en clair. Cette propriété unique permet de traiter des données sensibles sans avoir à les déchiffrer. Par exemple, un utilisateur pourrait soumettre une requête chiffrée à un LLM, qui traiterait la requête et renverrait un résultat chiffré. L'utilisateur pourrait ensuite déchiffrer ce résultat localement. Ce processus garantit que le LLM n'accède jamais aux données sensibles brutes. Les revers de cette surcouche de sécurité sont le coût de calcul et la difficulté technique pour mettre en place et maintenir un tel dispositif. Zama, est une entreprise qui met en place cette technologie.

UN EXEMPLE D'UTILISATION D'ENCLAVES SÉCURISÉES

Aspect	Sans Enclave Sécurisée	Avec Enclave Sécurisée
Entrée des Données	Les utilisateurs saisissent des données sensibles via une interface d'application.	
Transmission des Données	Cryptées en transit, déchiffrées en atteignant le serveur.	Cryptées en transit, restent cryptées jusqu'à l'intérieur de l'enclave.
Environnement de Traitement	Environnement opérationnel standard, partagé avec d'autres processus du serveur.	Isolé au sein de l'enclave sécurisée, physiquement séparé des autres processus.
Vulnérabilités de Sécurité	Exposé aux risques si le système d'exploitation du serveur ou d'autres processus sont compromis.	Risque considérablement réduit ; protégé contre les violations externes et les menaces internes.
Données en RAM	Potentiellement exposées à d'autres processus sur le serveur.	Cryptées et isolées dans l'espace mémoire de l'enclave.
Génération de Réponse	Générée dans l'environnement standard et renvoyée.	Générée à l'intérieur de l'enclave, cryptée et renvoyée.
Stockage des Données	Stockées sur le serveur, soumises aux mesures de sécurité standard.	Stockées de manière sécurisée, soit à l'intérieur de l'enclave, soit cryptées avant le stockage à l'extérieur.
Conformité et Confiance	Peut ne pas répondre aux normes réglementaires supérieures ; confiance des utilisateurs plus faible.	Mieux équipé pour répondre aux réglementations strictes ; confiance des utilisateurs plus élevée.
Performance	Performance standard sans surcoût de sécurité supplémentaire.	Compromis de performance potentiels en raison des processus de cryptage et d'isolation.

UN EXEMPLE D'UTILISATION DU CHIFFREMENT HOMOMORPHE

Aspect	Sans Chiffrement Homomorphe	Avec Chiffrement Homomorphe
Entrée des Données	Les utilisateurs saisissent des données sensibles, qui sont potentiellement exposées pendant le traitement.	Les utilisateurs saisissent des données sensibles sous forme chiffrée, préservant la confidentialité.
Traitement des Données	Les données sont traitées en texte clair, nécessitant un déchiffrement avant le traitement, augmentant le risque d'exposition.	Les données sont traitées sous leur forme chiffrée, garantissant que les informations sensibles restent sécurisées tout au long.
Confidentialité des Données	Les informations sensibles sont à risque d'être exposées pendant le traitement.	Niveau élevé de confidentialité des données car les informations restent chiffrées en tout temps.
Performance	Performance computationnelle standard, sans le surcoût du chiffrement.	Surcoût computationnel accru en raison des complexités du chiffrement homomorphe.
Conformité Réglementaire	Défis potentiels pour répondre aux réglementations strictes de confidentialité des données.	Conformité plus aisée avec des réglementations strictes de protection des données telles que le GDPR, car les données restent chiffrées.
Adéquation au Cas d'Usage	Adapté aux données moins sensibles où le surcoût du chiffrement est une préoccupation.	Idéal pour les données hautement sensibles, en particulier là où la confidentialité est primordiale.

TD : MAÎTRISER LE PROMPTING AVEC LA GÉNÉRATION D'IMAGE

TD : MAÎTRISER LE PROMPTING

- Sur HuggingFace, tester les modèles de génération d'image à partir d'un prompt:
 - [stabilityai/sdxl-turbo](#)
 - [stabilityai/stable-diffusion-xl-base-1.0](#)
 - [black-forest-labs/FLUX.1-schnell](#)
 - [black-forest-labs/FLUX.1-dev](#)
- Comparer les résultats pour un même prompt

A cat holding a sign that says hello world



stabilityai/sdxl-turbo



stabilityai/stable-diffusion-xl-base-1.0



black-forest-labs/FLUX.1-schnell



black-forest-labs/FLUX.1-dev

TD : MAÎTRISER LE PROMPTING

- Le prompt et sa précision est critique. Essayer de reproduire les images suivantes :



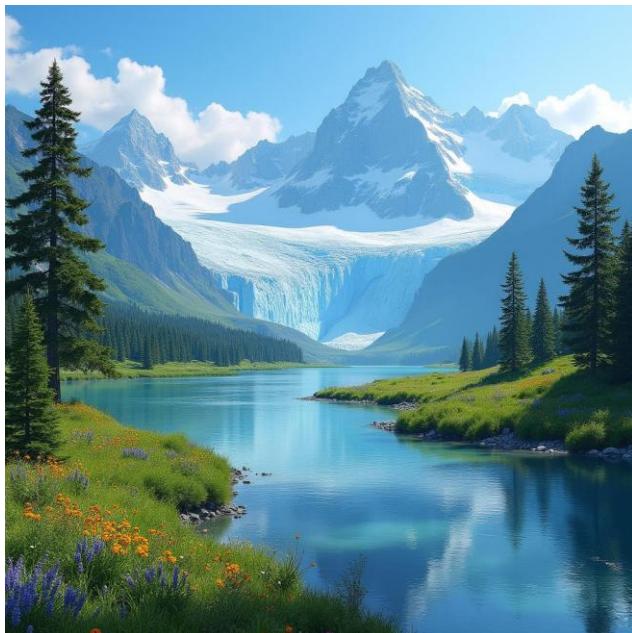
TD : MAÎTRISER LE PROMPTING

■ Le prompt et sa précision est critique. Essayer de reproduire les images suivantes :

a screengrab from an old VHS video showing a half Japanese, half American early 20s lady wearing y2k clothing standing next to single tall pile of books, looking towards viewer, cool color tone, fine-grained



A breathtaking landscape featuring towering glaciers in the background, with rugged icy peaks under a clear, vivid blue sky. A serene river flows through the center of the scene, reflecting the glaciers and sky. The riverbanks are lined with lush green grass, interspersed with vibrant wildflowers adding splashes of color. Scattered trees with tall, full, deep green foliage are visible along the river and in the foreground, contributing to the peaceful, natural setting. The overall atmosphere is one of pristine, untouched wilderness.



Lively selfie capturing a diverse group of friends at a colorful street art festival. The image is slightly tilted, giving it an authentic, spontaneous feel.

In the center, a tall Afro-Latina woman with vibrant purple curly hair holds the phone at arm's length, her gold septum ring glinting in the sunlight. She's wearing a neon green crop top and is mid-laugh, eyes crinkling with joy.

To her right, a Korean man with a half-shaved head and multiple ear piercings is making a playful face, sticking out his tongue. He's dressed in a vintage band t-shirt layered under a denim jacket covered in enamel pins.

On the left, a freckled redhead Scottish woman throws up peace signs with both hands. Her hair is in two braids, and she's sporting oversized round sunglasses and a tie-dye shirt.



QCM D'ÉVALUATION



QCM D'ÉVALUATION

- QCM à remplir avec une adresse email valide permettant de vous identifier facilement. Ce QCM sert d'évaluation pour ce cours.
- Lien : <https://forms.gle/jHGkzduTDsWqHWjs9>
- Temps : 1 heure
- 30 questions

RESSOURCES POUR APPROFONDIR ET ÊTRE À JOUR

RESSOURCES POUR APPROFONDIR ET ÊTRE À JOUR

- Challenges and Applications of Large Language Models :
<https://arxiv.org/abs/2307.10169>
- Llama 2: Open Foundation and Fine-Tuned Chat Models:
<https://arxiv.org/abs/2307.09288>
- Natural Language Processing with Transformers : <https://www.amazon.fr/Natural-Language-Processing-Transformers-Revised-ebook/dp/B0B2FKYVNL>
- Deep Learning, Goodfellow-et-al-2016, <https://www.deeplearningbook.org/>
- <https://www.reddit.com/r/LocalLLaMA/>
- X (Twitter)

Merci pour votre attention

Retrouvez nos offres de stage sur le site d'ARTIK CONSULTING

<https://www.artik-consulting.com/nous-rejoindre>

Contacts :

olivier.guerin@artik-consulting.com