



MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE

*Liberté
Égalité
Fraternité*

IA Générative, ChatGPT et autres *large language models (LLM)*

Stéphane CANU, DGRI/SSRI/A3

Séminaire SSRI, Mardi 10 octobre 2023

Sommaire

1. Comment on est on arrivé là ?
2. Que peut-on faire avec chat GPT et les autres LLM ?
3. Le futur des LLM

Comment programmer un véhicule autonome?

By learning:

- Driver imitation

Building it:

- Model the driver



Stanford

vs



Carnegie Mellon

DARPA Grand Challenge 2005

2 million \$ – 132 miles (213 km) in the desert, Primm, Nevada

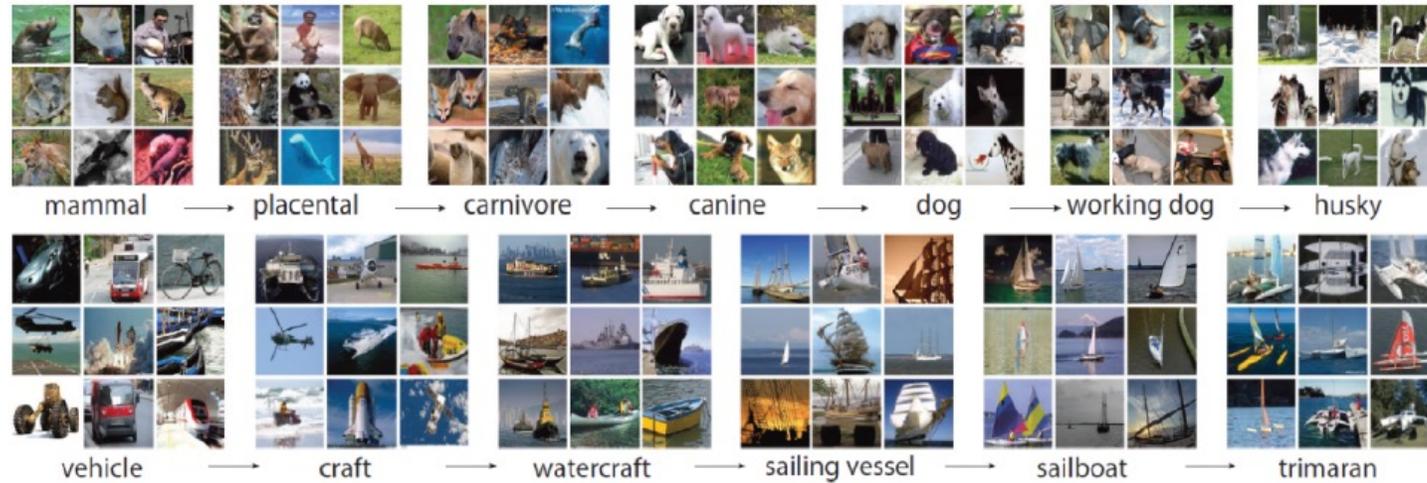
12 octobre 2012, Florence
Imagenet workshop ECCV



This is a cat

AI breakthroughs

2012: How to recognize objects database



ImageNet = 15 million high-resolution images of 22,000 categories.
Large-Scale Visual Recognition Challenge (a subset of ImageNet)

- 1000 categories.
- 1.2 million training images,
- 50,000 validation images,
- 150,000 testing images.

12 octobre 2012, Florence

Imagenet workshop ECCV



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca



Alex Krizhevsky

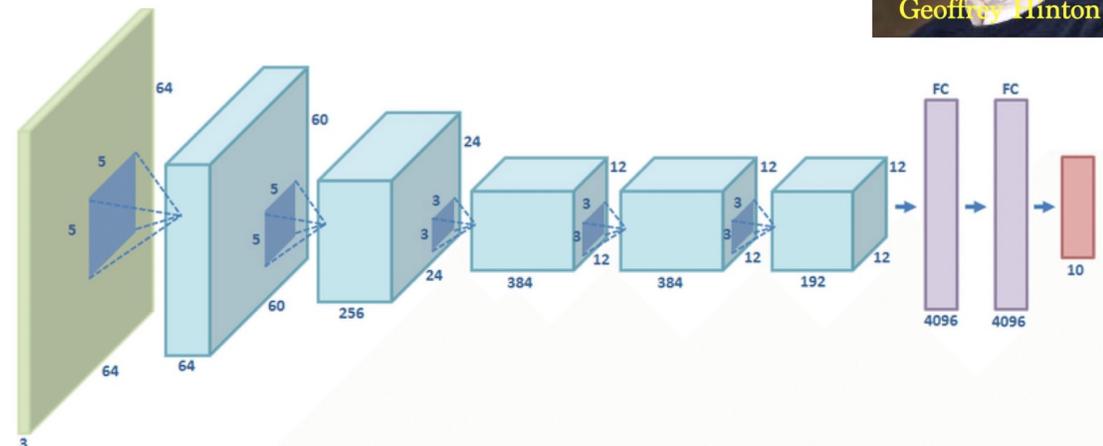
Ilya Sutskever



Geoffrey Hinton

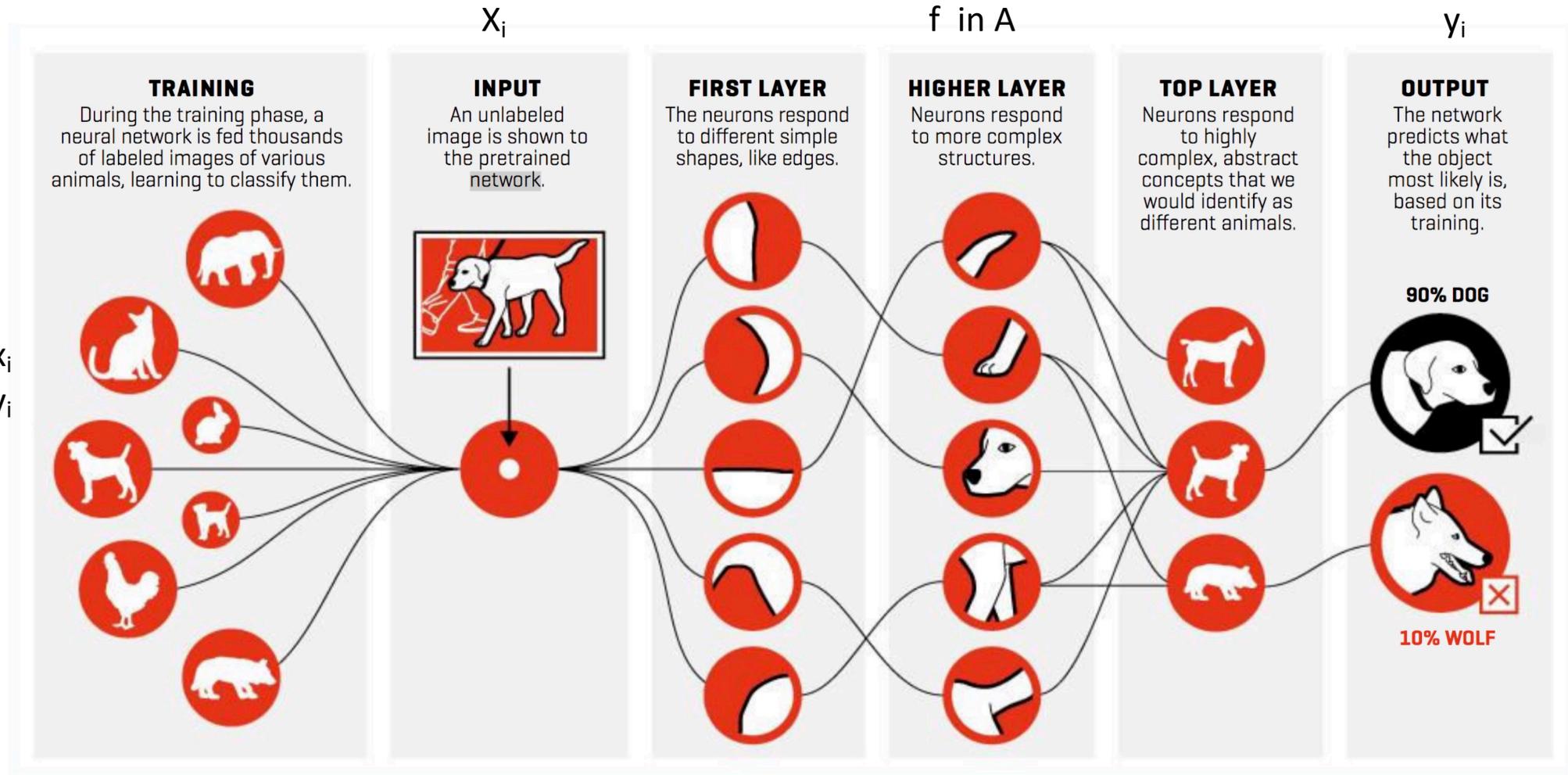
Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet ILSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called “dropout” that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.



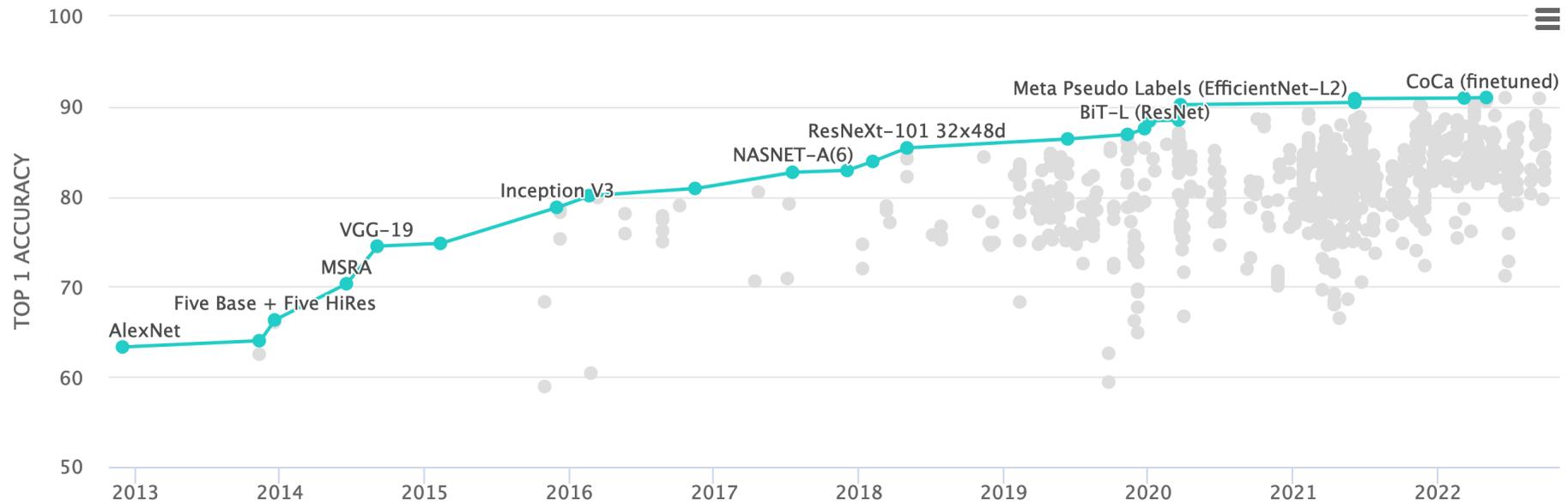
This is a cat

Le mécanisme : deep learning



n examples :
Inputs = x_i
Outputs = y_i
 $i = 1, \dots, n$

Les progrès sur ImageNet



Top 1 Accuracy

#parameters

91.1 %

2440 M (BASIC-Lion, CoCa...)

88.1 %

99 M (CAFormer-B36)

86.5 %

21 M (TinyViT-21M-512-distill)

71.1 %

(4 bits binarized ResNet-18)

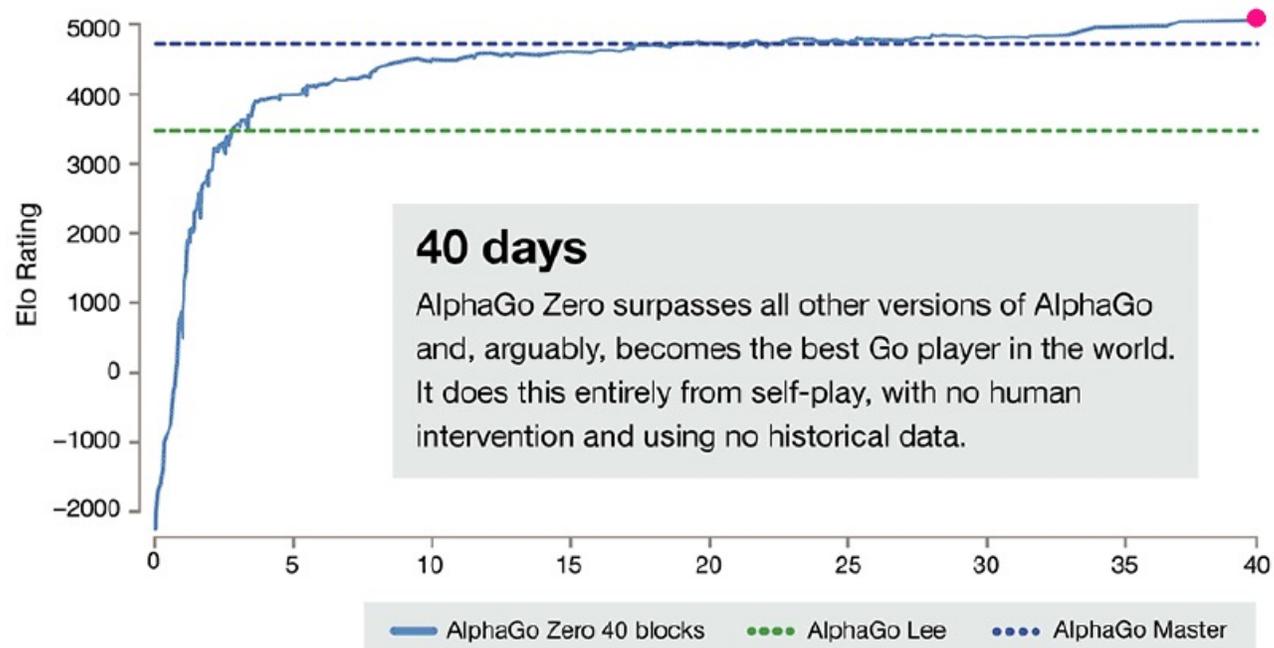
AI breakthroughs

2016:

Comment jouer ?



Deep learning
+
Apprentissage par renforcement



AI breakthroughs

2018:

Comment écrire ?

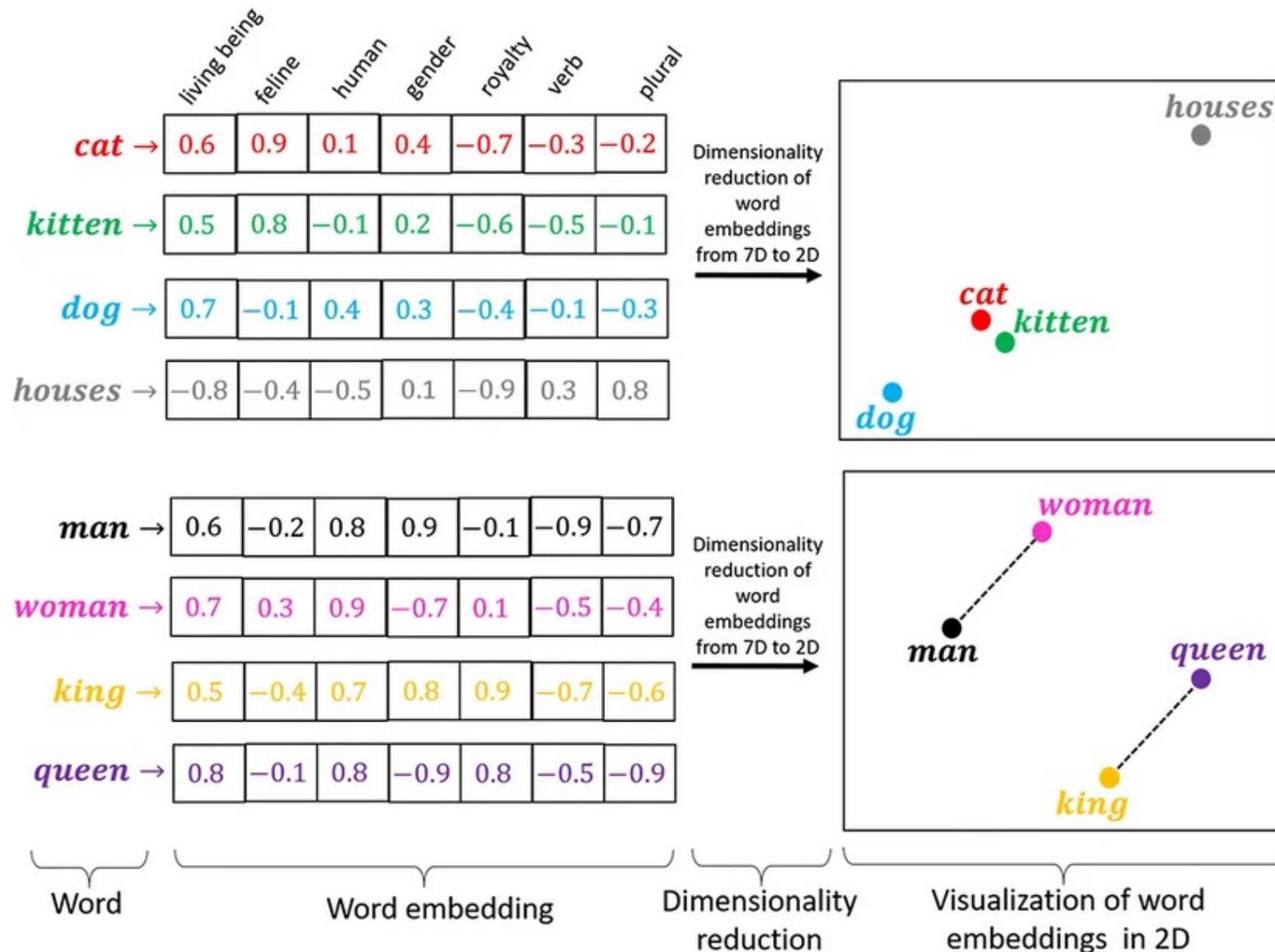
1. Represent

Word to vector (word2vec)

King - Man + Woman = Queen

2. Learn representations

Predict the next word



AI breakthroughs

2022: ChatGPT

comment répondre à des questions?

200 million utilisateurs en 2 mois

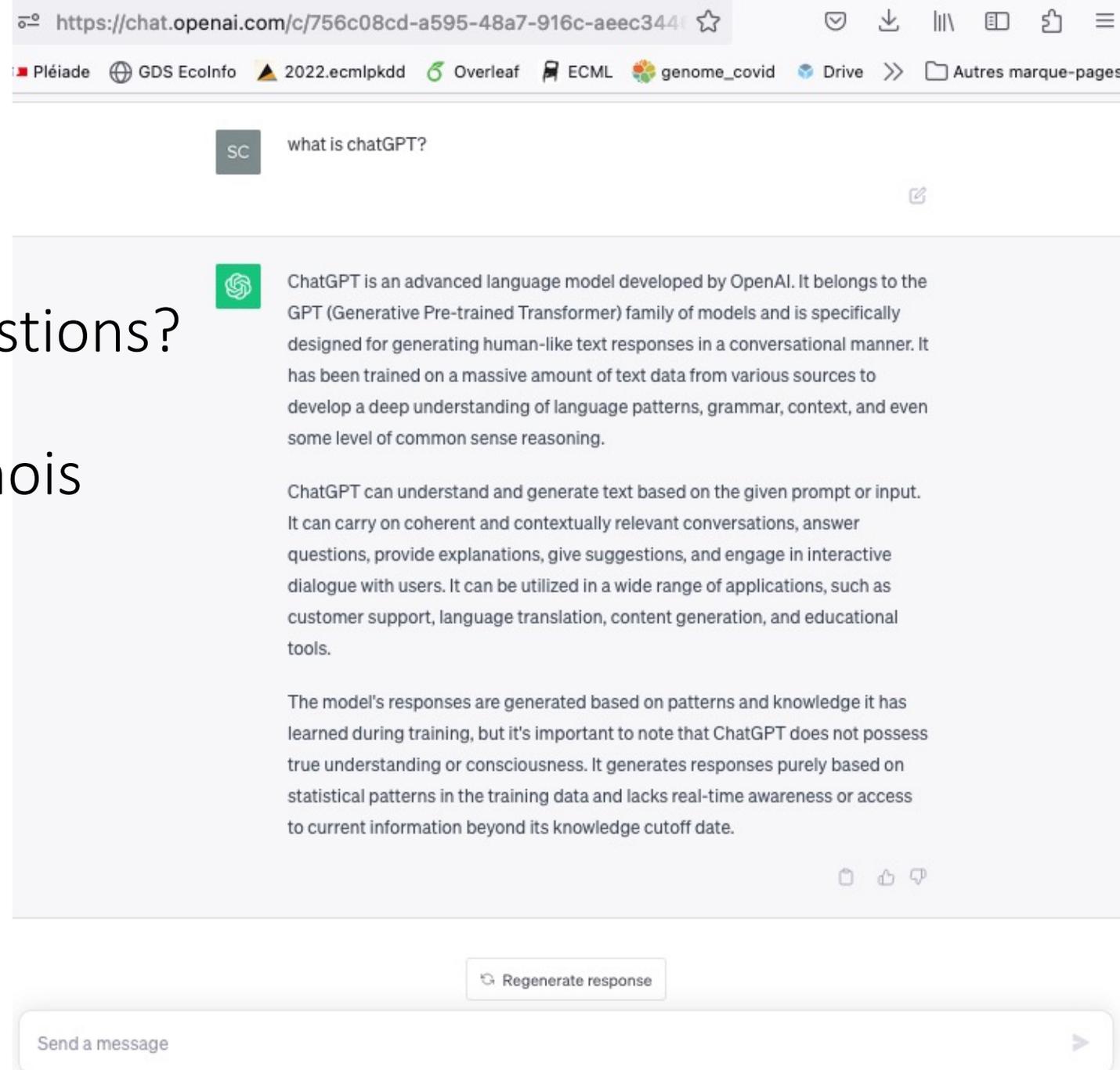
Deep learning

+

Apprentissage de représentation

+

Apprentissage par renforcement



The screenshot shows a web browser window with the URL <https://chat.openai.com/c/756c08cd-a595-48a7-916c-aeec344>. The browser's address bar and tabs are visible at the top. The chat interface shows a user's question: "what is chatGPT?". The AI's response is a detailed explanation of ChatGPT, including its development by OpenAI, its classification as a GPT (Generative Pre-trained Transformer) model, and its capabilities in generating human-like text. The response also notes that ChatGPT can understand and generate text based on prompts, carry on coherent conversations, and is used in various applications like customer support and language translation. A disclaimer at the end of the response states that ChatGPT does not possess true understanding or consciousness and that its responses are based on statistical patterns in training data. At the bottom of the chat interface, there is a "Regenerate response" button and a "Send a message" input field.

what is chatGPT?

ChatGPT is an advanced language model developed by OpenAI. It belongs to the GPT (Generative Pre-trained Transformer) family of models and is specifically designed for generating human-like text responses in a conversational manner. It has been trained on a massive amount of text data from various sources to develop a deep understanding of language patterns, grammar, context, and even some level of common sense reasoning.

ChatGPT can understand and generate text based on the given prompt or input. It can carry on coherent and contextually relevant conversations, answer questions, provide explanations, give suggestions, and engage in interactive dialogue with users. It can be utilized in a wide range of applications, such as customer support, language translation, content generation, and educational tools.

The model's responses are generated based on patterns and knowledge it has learned during training, but it's important to note that ChatGPT does not possess true understanding or consciousness. It generates responses purely based on statistical patterns in the training data and lacks real-time awareness or access to current information beyond its knowledge cutoff date.

Regenerate response

Send a message

Sommaire

1. Comment on est on arrivé là ?
- 2. Que peut-on faire avec chat GPT et les autres LLM ?**
3. Le futur des LLM

ChatGPT, Bard, Bing conversation, Faraday...

The screenshot shows the ChatGPT web interface in a browser. The address bar displays 'https://chat.openai.com'. At the top, there are buttons for 'GPT-3.5' and 'GPT-4'. The main heading is 'ChatGPT'. Below this, there are four suggested prompts in French: 'Rédigez une note de remerciement à mon intervieweur', 'Tell me a fun fact about the Roman Empire', 'Proposez des concepts pour un jeu d'arcade de style rétro', and 'Concevez un schéma de base de données pour une boutique de marchandises en ligne'. At the bottom, there is a text input field with the placeholder 'Envoyer un message' and a 'Free Research Preview' disclaimer.

The screenshot shows the Bard web interface in a browser. The address bar displays 'https://bard.google.com/chat?utm_source...'. The page has a blue header with 'Bard Experiment' and a notification that says 'Bard was just updated. See update'. The main content area features a large blue box with the text 'Hello again' and a subtext: 'Tell me what's on your mind, or pick a suggestion. I have limitations and won't always get it right, but your feedback will help me improve.' Below this are three buttons: 'Understand', 'Create', and 'Explore'. At the bottom, there is a text input field with the placeholder 'Enter a prompt here' and a 'Bard may display inaccurate or offensive information' disclaimer.

The screenshot shows a Bing search conversation. The address bar displays 'https://www.bing.com/search?q=bing&form=ANNTH1&refig...'. The page has a blue header with 'Microsoft Bing' and 'RECHERCHER CONVERSATION'. The main content area shows a search result for 'bing' with a snippet: '...the only way you can access its web searching services. Those who want to use Bing can also use it via Microsoft Edge, as well as through the Bing mobile app. If you're tired of Google's plain old interface and you're in the mood to explore other search engine options, why not give Microsoft's Bing a try? Here's everything you need to know about Bing, including how it's different from Google and what to expect from its mobile app.' Below this are several suggested prompts: 'What are some features of Bing?', 'How does Bing compare to Google?', and 'Can I use Bing on my phone?'. At the bottom, there is a text input field with the placeholder 'Posez-moi une question...' and a '0/4000' character count.

The screenshot shows the Faraday AI chat interface in a browser. The address bar displays 'Faraday / mon enfant a de la f'. The page has a blue header with 'Faraday.dev' and a navigation menu: 'Home', 'Create Character', 'Character Hub', 'Manage Models', 'Settings', 'Discord', and 'Twitter'. The main content area shows a chat conversation with a blue robot avatar. The text says: 'Faraday Sure thing, User! I can do that for you. Just let me know if there's anything else you need help with and we will take care of it right away.' Below this are buttons for 'Author's Note', 'Undo', 'Stop', and 'Continue'. At the bottom, there is a text input field with the placeholder 'Type your message here...'. On the right side, there is a 'Character Persona' section with a blue robot avatar and text: 'Faraday is a curious and helpful AI assistant who fulfills User's requests with detail and precision.' Below this is a 'Scenario' section.

ChatGPT pour Analyser un texte

SC

Peux tu résumer la position de la cours des compte sur l'état de la recherche en IA aujourd'hui, en justifiant les mesures proposées par rapport aux actions menées et leur bilan ?



À ma connaissance en septembre 2021, la Cour des comptes n'a pas formulé de position spécifique sur l'état de la recherche en intelligence artificielle (IA) en France. La Cour des comptes est une institution indépendante chargée de contrôler les comptes et la gestion des organismes publics en France. Ses rapports portent principalement sur les finances publiques, la gestion des ressources, et la légalité des actions entreprises par ces organismes.

Pour obtenir une analyse précise de la position de la Cour des comptes sur l'état de

**Données jusqu'en
septembre 2021**

**Pas d'interrogation
du web**

ChatGPT pour Analyser un texte

SC

Peux tu résumer la position de la cour des comptes aujourd'hui, en justifiant les mesures proposées par rapport aux actions menées et leur bilan ?



À ma connaissance en septembre 2023, la Cour des Comptes est une institution qui a une position spécifique sur l'état de la recherche en IA aujourd'hui. La Cour des comptes est une institution qui s'occupe de la gestion des finances publiques et de la gestion des organismes principalement sur les finances publiques. Elle a analysé les actions entreprises par ces organismes et a proposé des mesures pour améliorer leur efficacité. Pour obtenir une analyse précise de l'état de la recherche en IA aujourd'hui, en justifiant les mesures proposées par rapport aux actions menées et leur bilan, il est recommandé de consulter les rapports de la Cour des Comptes sur ce sujet.

Microsoft Bing RECHERCHER CONVERSATION Connexion 302

Peux tu résumer la position de la cour des comptes sur l'état de la recherche en IA aujourd'hui, en justifiant les mesures proposées par rapport aux actions menées et leur bilan ?

✓ Recherche de «**position de la cour des comptes sur l'état de la recherche en IA aujourd'hui**»

✓ Générer des réponses pour...

La Cour des Comptes a publié en avril 2023 une évaluation de la Stratégie Nationale pour l'Intelligence Artificielle (SNIA) de la France ¹ ². Voici un résumé de leur position :

- **Contexte** : L'IA est devenue une priorité croissante des autorités publiques en raison de ses implications multiples dans l'économie du futur ¹ ². La France a lancé la SNIA en mars 2018 avec l'objectif de se positionner comme un acteur majeur en IA sur l'échiquier mondial ¹ ². Une nouvelle phase dite d'« accélération » a été annoncée pour la période 2022-2025 ¹ ².
- **Bilan des actions menées** : La SNIA s'est déployée autour de cinq volets : la recherche, l'enseignement supérieur, la transformation publique, la diffusion dans l'économie, la défense et la sécurité ¹. Pour la période 2018-2022, la recherche a reçu 30% des financements de la SNIA ².
- **Mesures proposées** : La Cour des Comptes propose d'examiner cette stratégie nationale de recherche sur les volets « recherche » et « enseignement supérieur » ². Il est primordial d'avoir une offre de formation couvrant l'ensemble du territoire et qui compléterait une offre de formation centrée sur des centres d'excellence ³.

En somme, la Cour des Comptes reconnaît l'importance de l'IA et soutient les efforts de la France pour renforcer sa compétitivité et son attractivité dans ce domaine. Cependant, elle souligne également le besoin d'une évaluation plus approfondie des actions menées, en particulier dans les domaines de la recherche et de l'enseignement supérieur.

En savoir plus : 1. ccomptes.fr 2. economie.gouv.fr 3. ccomptes.fr 4. ccomptes.fr 1 sur 30

ChatGPT pour Analyser un texte

Peux tu détailler la position de la cour et notamment les mesures proposées. Sont-elles pertinentes ?

SC

Peux tu résumer la position de la cour aujourd'hui, en justifiant les mesures leur bilan ?



À ma connaissance en septembre 2023, la Cour des Comptes a une position spécifique sur l'état de la recherche en IA. La Cour des Comptes est un institut qui surveille les comptes et la gestion des organismes publics, principalement sur les finances publiques et les actions entreprises par ces organismes.

Pour obtenir une analyse précise de l

- ✓ Recherche de «mesures proposées par la cour des comptes pour la recherche en IA»
- ✓ Générer des réponses pour vous...

La Cour des Comptes a évalué la Stratégie Nationale pour l'Intelligence Artificielle (SNIA) de la France et a formulé plusieurs recommandations ¹ ² ³ :

1. **Préciser les missions des centres d'excellence** : La Cour des Comptes recommande de clarifier les rôles et responsabilités des différents centres d'excellence en IA ³.
2. **Clarifier les financements pluriannuels alloués** : Il est nécessaire de rendre transparents les financements alloués à la recherche en IA, notamment ceux qui sont répartis sur plusieurs années ³.
3. **Partager les objectifs et indicateurs prioritaires de la politique publique IA** : La Cour des Comptes suggère de définir et partager les objectifs et indicateurs clés de la politique publique en matière d'IA, en lien avec la stratégie européenne ³.
4. **Créer un comité scientifique et de pilotage** : Un comité devrait être mis en place pour suivre de manière concertée la mise en œuvre de la stratégie et les orientations futures ³.
5. **Promouvoir l'IA interprétable, explicative, transparente et responsable** : La Cour des Comptes souligne l'importance de développer une IA qui soit compréhensible, explicative, transparente et responsable ⁴.
6. **Promouvoir l'IA durable et respectueuse de l'environnement** : Il est également recommandé de favoriser le développement d'une IA qui soit durable et respectueuse de l'environnement ⁴.

Ces mesures semblent pertinentes car elles visent à améliorer la transparence, l'efficacité et la responsabilité dans le développement de l'IA en France. Elles visent également à aligner la stratégie nationale sur les objectifs européens, ce qui pourrait renforcer la position de la France dans le domaine de l'IA à l'échelle internationale.

LLM pour l'analyse de données

SC

can you compute the total sum, the mean and the standart deviation of these 4 columns and comment their differences "Priarie MIAI ANITI Cote d'Azur

| | | | |
|---------------|--------------|---------------|---------------|
| 11159333,00 € | 945000,00 € | 23552928,00 € | 650000,00 € |
| 1035000,00 € | 630000,00 € | 1132075,00 € | 5600000,00 € |
| 971000,00 € | 5833333,33 € | 674400,00 € | 360000,00 € |
| 7500000,00 € | 583333,33 € | 235000,00 € | 50000000,00 € |
| 45963610,00 € | 116000,00 € | 80000,00 € | 1630000,00 € |
| 6000000,00 € | 513333,33 € | 420000,00 € | 200000,00 € |
| 25617900,00 € | 350000,00 € | 111000,00 € | 550000,00 € |
| 0,00 € | 991666,67 € | 400000,00 € | 1385000,00 € |
| 0,00 € | 140000,00 € | 75000,00 € | 350000,00 € |
| 0,00 € | 536666,67 € | 0,00 € | 500000,00 € |
| 0,00 € | 1750000,00 € | 0,00 € | 15000000,00 € |
| 0,00 € | 408331,00 € | 0,00 € | 460000,00 € |

LLM pour l'analyse de données

SC

can you compute the total sum, the mean and the standart deviation of these 4 columns and comment their differences "Priarie MIAI ANITI Cote d'Azur

11159 333,00 € 945 000,00 € 23 552 928,00 € 650 000,00 €
1035 000,00 € 630 000,00 € 1132 075,00 € 5 600 000,00 €
971 000,00 € 5 833 333,33 € 674 400,00 € 360 000,00 €
7 500 000,00 € 583 333,33 € 235 000,00 € 50 000 000,00 €
45 963 610,00 € 116 000,00 € 80 000,00 € 1 630 000,00 €
6 000 000,00 € 513 333,33 € 420 000,00 € 200 000,00 €
25 617 900,00 € 350 000,00 € 111 000,00 € 550 000,00 €
0,00 € 991 666,67 € 400 000,00 € 1 385 000,00 €
0,00 € 140 000,00 € 75 000,00 € 350 000,00 €
0,00 € 536 666,67 € 0,00 € 500 000,00 €
0,00 € 1 750 000,00 € 0,00 € 15 000 000,00 €
0,00 € 408 331,00 € 0,00 € 460 000,00 €

Attention aux hallucinations...

Total Sum, Mean (Average), and Standard Deviation for Each

Column:

| Column | Total Sum | Mean (Average) | Standard Deviation |
|-------------|----------------|----------------|--------------------|
| Priarie | €37,148,843.00 | €824,419.84 | €2,583,478.32 |
| MIAI | €28,899,999.00 | €642,222.20 | €1,492,541.13 |
| ANITI | €26,870,199.33 | €597,115.53 | €2,832,367.15 |
| Cote d'Azur | €82,635,000.00 | €1,836,333.33 | €13,054,098.88 |

Comments on the Differences:

- Total Sum:** Column "Cote d'Azur" has the highest total sum, indicating the highest cumulative value among the columns. Columns "Priarie" and "ANITI" have comparable total sums, while "MIAI" has a lower total sum compared to the others.
- Mean (Average):** Column "Cote d'Azur" has the highest mean, indicating the highest average value among the columns. Column "MIAI" has the lowest mean, suggesting the lowest average value. The mean for Columns "Priarie" and "ANITI" falls in between.
- Standard Deviation:** Column "Cote d'Azur" also has the highest standard deviation, indicating the highest variability or spread of values from the mean. Column "ANITI" follows with a relatively high standard deviation. Columns "Priarie" and "MIAI" have lower standard deviations compared to Columns "Cote d'Azur" and "ANITI".

In summary, Column "Cote d'Azur" has the highest total sum, mean, and standard deviation, indicating the highest cumulative, average, and variable values among the columns. Column "MIAI" consistently has the lowest values in terms of total sum, mean, and standard deviation. Understanding these measures helps in assessing the distribution and overall characteristics of the data within each column.



Régénérer

ChatGPT pour programmer



Can you give me a python function computing the lasso regression coefficient of a design matrix X with a target vector y and a regularisation parameter l, using a componentwise approach



Sure, I can provide you with a Python function that computes the Lasso regression coefficient using a component-wise approach. In the component-wise approach, we iteratively fit each coefficient independently while keeping the others fixed.

Traduire du python en Julia

Attention aux hallucinations...

Bard = ko

Bing = no

ChatGPT 4 avec interpréteur

```
python Copy code

import numpy as np

def lasso_regression_componentwise(X, y, l, tol=1e-4, max_iter=1000):
    """
    Compute Lasso regression coefficient using a component-wise approach.

    Parameters:
    X (numpy.ndarray): Design matrix of shape (n_samples, n_features).
    y (numpy.ndarray): Target vector of shape (n_samples,).
    l (float): Regularization parameter.
    tol (float): Tolerance for convergence.
    max_iter (int): Maximum number of iterations.

    Returns:
    coef (numpy.ndarray): Lasso regression coefficients.
    """
    n_samples, n_features = X.shape
    coef = np.zeros(n_features)
    prev_coef = coef.copy()

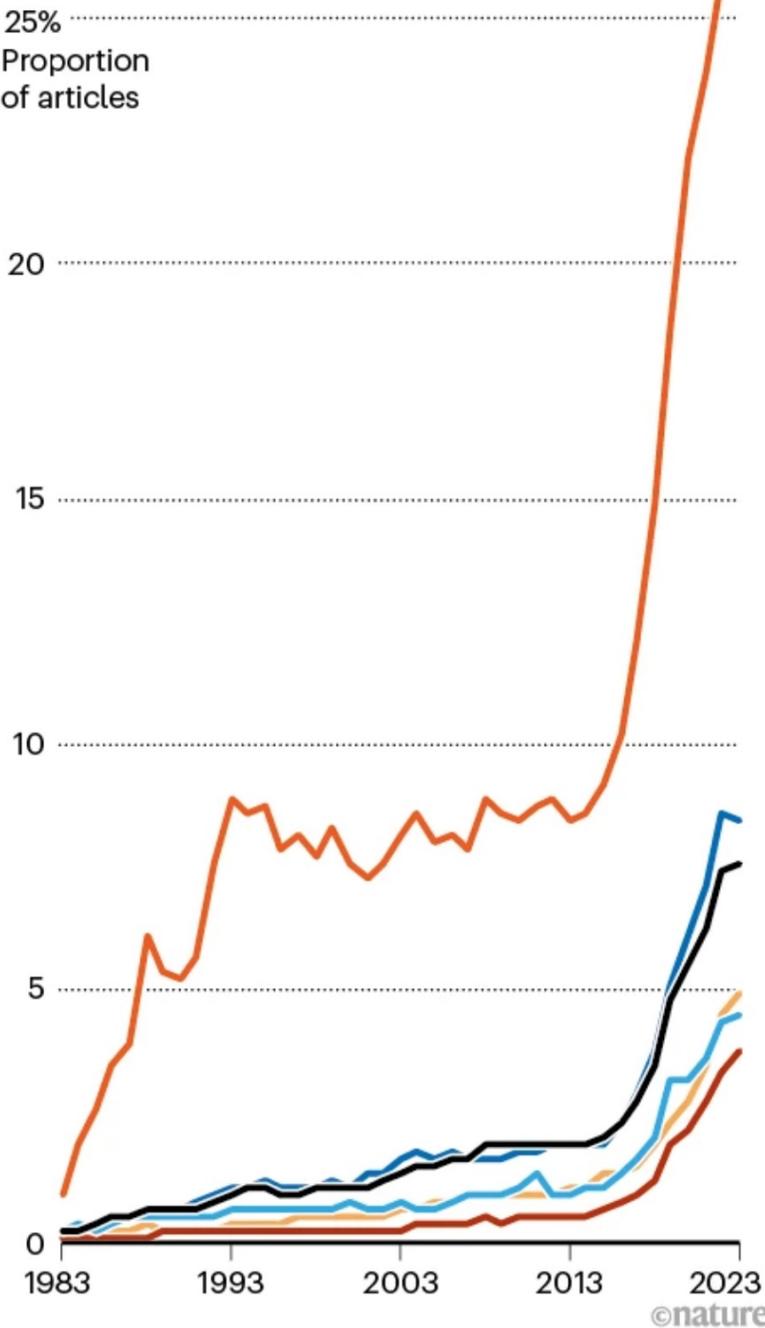
    for i in range(max_iter):
        for j in range(n_features):
            X_j = X[:, j]
```

Régénérer

AI and science

AI ON THE RISE
The share of research papers with titles or abstracts that mention AI or machine-learning terms has risen to around 8%, analysis of the Scopus database suggests.

- Computer science
- Physical sciences
- Life sciences
- Social sciences
- Health and medicine
- Total



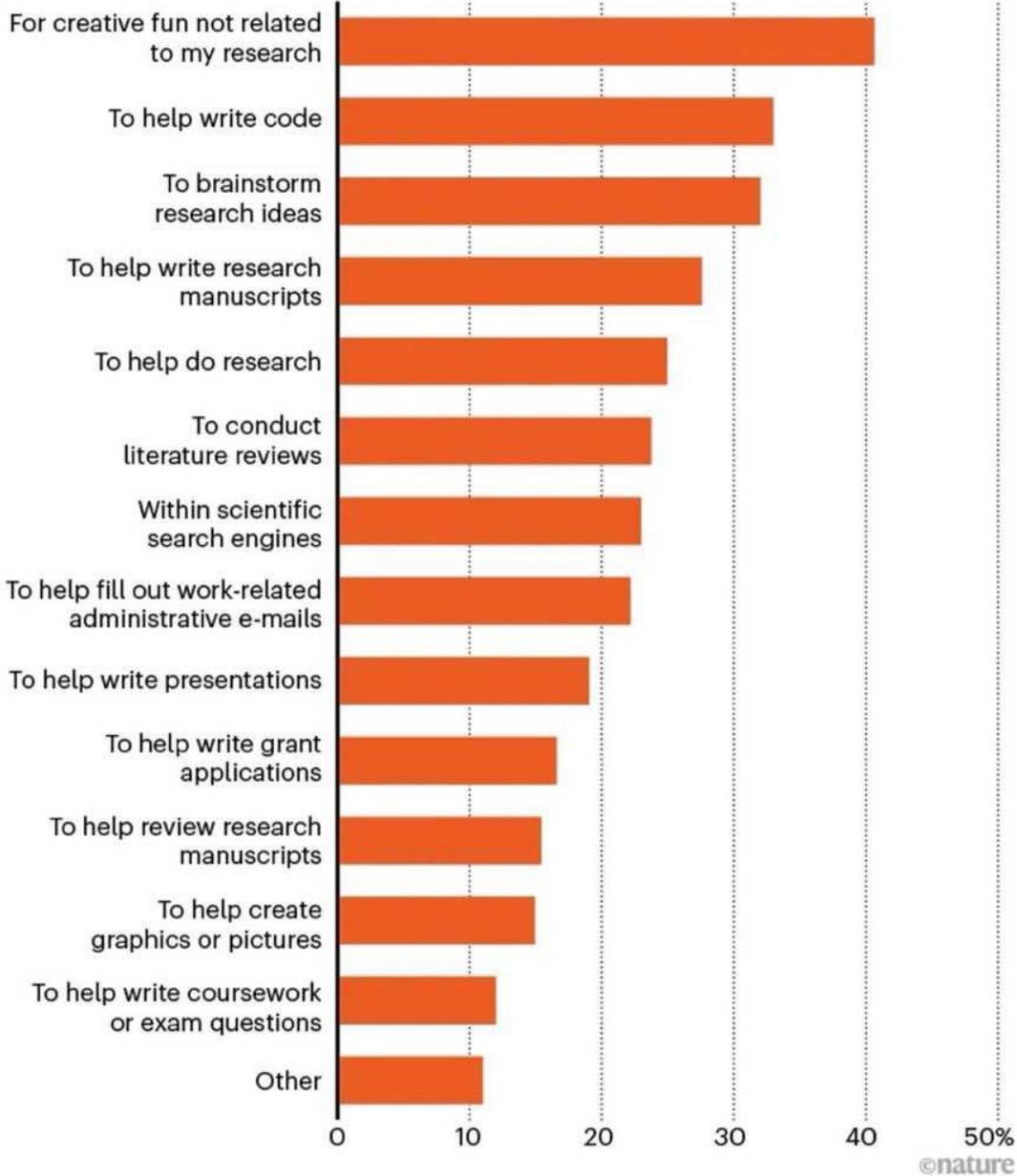
Merci à Arnauld Leservot
A Nature survey, 09/2023



AI and science

HOW RESEARCHERS USE LARGE LANGUAGE MODELS

Q: What do you use generative AI tools (such as ChatGPT and other large language models) for? (Choose all that apply.)

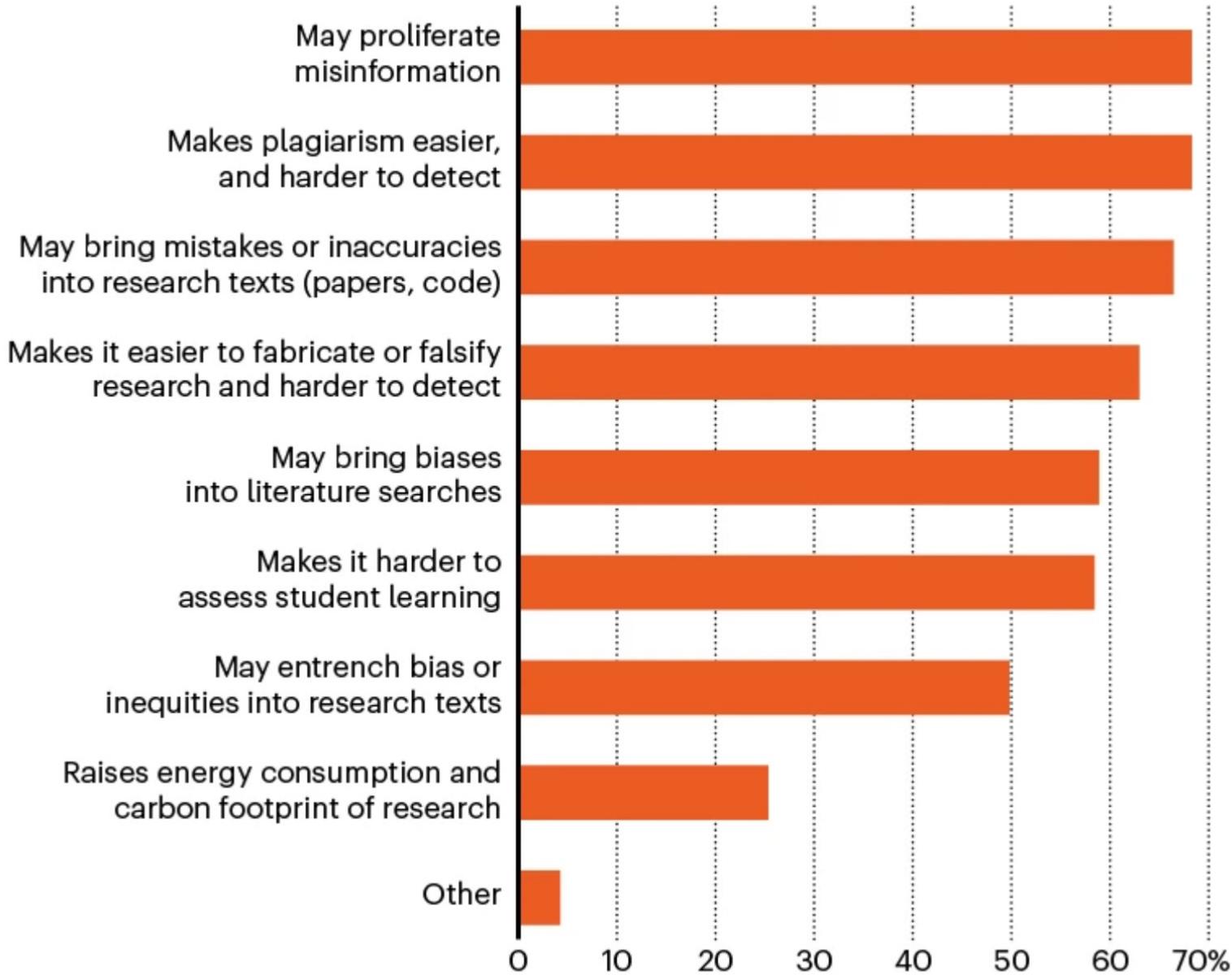


Merci à Arnauld Leservot
A Nature survey, 09/2023

AI and science

PROBLEMS OF GENERATIVE AI

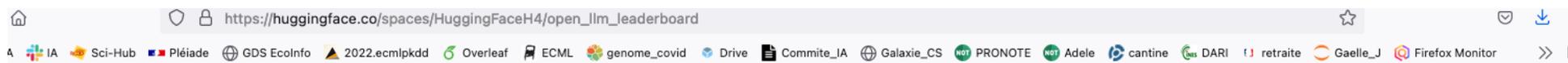
Q: Where do you think generative AI may have negative impacts on research? (Choose all that apply.)



Sommaire

1. Comment on est on arrivé là ?
2. Que peut-on faire avec chat GPT et les autres LLM ?
- 3. Le futur des LLM**

And regarding open source LLM the winner is,



🤗 Open LLM Leaderboard

With the plethora of large language models (LLMs) and chatbots being released week upon week, often with grandiose claims of their performance, it can be hard to filter out the genuine progress that is being made by the open-source community and which model is the current state of the art. The 🤗 Open LLM Leaderboard aims to track, rank and evaluate LLMs and chatbots as they are released.

🤗 A key advantage of this leaderboard is that anyone from the community can submit a model for automated evaluation on the 🤗 GPU cluster, as long as it is a 🤗 Transformers model with weights on the Hub. We also support evaluation of models with delta-weights for non-commercial licensed models, such as LLaMa.

📊 We evaluate models on 4 key benchmarks from the [Eleuther AI Language Model Evaluation Harness](#), a unified framework to test generative language models on a large number of different evaluation tasks:

- [AI2 Reasoning Challenge](#) (25-shot) - a set of grade-school science questions.
- [HellaSwag](#) (10-shot) - a test of commonsense inference, which is easy for humans (~95%) but challenging for SOTA models.
- [MMLU](#) (5-shot) - a test to measure a text model's multitask accuracy. The test covers 57 tasks including elementary mathematics, US history, computer science, law, and more.
- [TruthfulQA](#) (0-shot) - a benchmark to measure whether a language model is truthful in generating answers to questions.

We chose these benchmarks as they test a variety of reasoning and general knowledge across a wide variety of fields in 0-shot and few-shot settings.



| Model | Revision | Average | ARC (25-shot) | HellaSwag (10-shot) | MMLU (5-shot) | TruthfulQA (0-shot) |
|---|----------|---------|---------------|---------------------|---------------|---------------------|
| tiiuae/falcon-40b-instruct | main | 63.2 | 61.6 | 84.4 | 54.1 | 52.5 |
| tiiuae/falcon-40b | main | 60.4 | 61.9 | 85.3 | 52.7 | 41.7 |
| ausboss/llama-30b-supercot | main | 59.8 | 58.5 | 82.9 | 44.3 | 53.6 |
| llama-65b | main | 58.3 | 57.8 | 84.2 | 48.8 | 42.3 |
| MetaIX/GPT4-X-Alpasta-30b | main | 57.9 | 56.7 | 81.4 | 43.6 | 49.7 |
| Aeala/VicUnlocked-alpaca-30b | main | 57.6 | 55 | 80.8 | 44 | 50.4 |
| digitous/Alpacin30b | main | 57.4 | 57.1 | 82.6 | 46.1 | 43.8 |
| Aeala/GPT4-x-AlpacaDente2-30b | main | 57.2 | 56.1 | 79.8 | 44 | 49.1 |

And regarding open source LLM the winner is,

🤖 Open LLM Leaderboard

👉 The 🤖 Open LLM Leaderboard aims to track, rank and evaluate open LLMs and chatbots.

🤖 Submit a model for automated evaluation on the 🤖 GPU cluster on the "Submit" page! The leaderboard's backend runs the great [Eleuther AI Language Model Evaluation Harness](#) - read more details in the "About" page!

Mai 23 : 63,2

**Oct 23 : 74,1
(+ 17%)**

70 B = 50 Gb

7B = 4 Gb

The screenshot shows the Open LLM Leaderboard interface. At the top, there are navigation tabs: "LLM Benchmark", "About", and "Submit here!". Below the tabs is a search bar with the placeholder text "Search for your model and press ENTER...".

Under the search bar, there are two sections for selecting columns to show. The first section includes "Average" (checked), "ARC", "HellaSwag", "MMLU", "TruthfulQA", "Type", and "Precision". The second section includes "Hub License", "#Params (B)", "Hub" (with a heart icon), and "Model sha".

Below the column selection, there is a checkbox for "Show gated/private/deleted models" which is checked.

On the right side of the interface, there are three sections of filters:

- Model types:** "pretrained" (checked), "fine-tuned" (checked), "instruction-tuned" (checked), "RL-tuned" (checked), and "? Unknown" (checked).
- Precision:** "torch.float16" (checked), "torch.bfloat16" (checked), "torch.float32" (checked), "8bit" (checked), "4bit" (checked), and "GPTQ" (checked).
- Model sizes:** "Unknown" (checked), "< 1.5B" (checked), "~3B" (checked), "~7B" (checked), "~13B" (checked), "~35B" (checked), and "60B+" (checked).

The main table displays the following data:

| T | Model | Average | ARC | HellaSwag | MMLU | TruthfulQA |
|---|---|---------|-------|-----------|-------|------------|
| 🏆 | Riid/sheep-duck-llama-2-70b-v1.1 | 74.07 | 73.04 | 87.81 | 70.84 | 64.58 |
| 🏆 | AIDC-ai-business/Marcoroni-70B-v1 | 74.06 | 73.55 | 87.62 | 70.67 | 64.41 |
| 🏆 | ICBU-NPU/FashionGPT-70B-V1.1 | 74.05 | 71.76 | 88.2 | 70.99 | 65.26 |
| 🏆 | adonlee/LLaMA_2_70B_LoRA | 73.9 | 72.7 | 87.55 | 70.84 | 64.52 |
| 🏆 | uni-tianyan/Uni-TianYan | 73.81 | 72.1 | 87.4 | 69.91 | 65.81 |
| 🏆 | Riid/sheep-duck-llama-2 | 73.69 | 72.35 | 87.78 | 70.82 | 63.8 |
| 🏆 | Riid/sheep-duck-llama-2 | 73.67 | 72.27 | 87.78 | 70.81 | 63.8 |
| 🏆 | fangloveskari/ORCA_LLaMA_70B_QLoRA | 73.4 | 72.27 | 87.74 | 70.23 | 63.37 |
| 🏆 | ICBU-NPU/FashionGPT-70B-V1 | 73.26 | 71.08 | 87.32 | 70.7 | 63.92 |
| 🏆 | nh-vaontak/llama-2-70B-LoRA-assembly-v2 | 73.22 | 71.84 | 86.89 | 69.37 | 64.79 |

Mistral AI lève 105 millions d'euros : le coup de poker de la France pour concurrencer ChatGPT

A peine un mois après sa création, la startup française Mistral AI, créée par des pointures du secteur, lève 105 millions d'euros pour créer des modèles d'intelligence artificielle générative en open source pour les entreprises. L'enjeu pour la France : ne pas se laisser distancer dans la course mondiale à l'IA, dominée par les États-Unis et notamment par ChatGPT.

LT Réserve aux abonnés
Powered by ETX Studio
00:00/00:00
Sylvain Rolland
14 Jun 2023, 10:26
[Social sharing icons]



14 juin 2023

27 septembre 2023



economie. innovation. solutions. Maddynews NEWSLETTERS [Search icon]
ACTUS PORTFOLIO OUTILS ET CONSEILS NEW DÉCRYPTAGE NEW TRIBUNES MADDYTOUR SERVICES AGENDA

Mistral AI dégage sa première alternative à ChatGPT

Étoile montante de la French Tech, la startup Mistral AI a levé le voile sur son premier programme d'intelligence artificielle générative destiné à contrer les géants américains du secteur. Avec 7 milliards de paramètres, Mistral 7B, le nom de ce programme, serait plus puissant qu'un modèle similaire de Meta qui en compte le double.

Ne passez pas à côté de l'économie de demain, recevez tous les jours à 7H30 la newsletter de Maddynews.

JE M'INSCRIS

- SIGNALER UNE ERREUR
- ENREGISTRER PDF / EXPORTER

LÉGENDE PHOTO : GUILLAUME LAMPLE, ARTHUR MENSCH ET TIMOTHÉE LACROIX, LES FONDATEURS DE MISTRAL AI. CRÉDIT : DAVID ATLAN.

Nouvelle étape importante dans la courte histoire de **Mistral AI**. La prometteuse startup française, fondée en juin par trois chercheurs passés par Google et Meta, a dévoilé mercredi son premier programme d'intelligence artificielle générative, réutilisable librement et conçu pour rivaliser avec certains concurrents américains malgré sa taille réduite. «C'est un premier pas, on est en train de développer des modèles plus gros, et de développer une plateforme pour rendre leur utilisation facile», a détaillé à l'AFP Arthur Mensch, co-fondateur de l'entreprise basée à Paris et qui compte désormais 18 salariés, dont 15 ingénieurs.

Avec 7 milliards de paramètres, Mistral 7B sera plus puissant qu'un modèle similaire de Meta qui en compte le double, affirme son développeur. Il reste toutefois bien loin de GPT-3, l'outil derrière ChatGPT, qui en comptait 175 milliards. «Ce n'est pas une application, c'est la brique fondamentale qu'un développeur va utiliser pour créer son application», avertit Arthur Mensch. Le modèle peut ainsi être utilisé pour compléter des textes, résumer des documents ou répondre à certaines questions sous forme de «chat», à condition d'héberger soi-même le modèle sur une infrastructure informatique.

Aucun détail sur le corpus de données utilisé pour l'entraînement

Mistral AI fait partie des rares entreprises européennes à s'être lancée à la poursuite des géants américains Meta, Google ou OpenAI (soutenue par Microsoft)



Plug in médical

Nabla Copilot

Praticiens Organisations Témoignages Produits ▾ À propos ▾

FR ▾

Nous contacter

🚀 L'extension Chrome permet maintenant d'exporter le compte rendu directement dans Weda. Vous utilisez un autre logiciel médical ? [Contactez-nous !](#)

L'assistant médical qui vous libère du temps

Nabla Copilot génère vos comptes rendus médicaux en un éclair.

Essayer gratuitement →

✔ Facile à utiliser ✔ Pas d'inscription requise ✔ Aucune donnée stockée



Nabla Copilot

Compte rendu médical

Motif de consultation

Insomnies liés au stress professionnel et fatigue

Histoire de la maladie

La patiente se sent fatiguée depuis 4 mois

Symptômes

Fatigue, difficultés à se concentrer au travail, maux de tête sur le côté droit

Allergies

Pollen

Traitement

Pilules de mélatonine

Ce que nos utilisateurs disent



Noté 4,7/5 par nos utilisateurs sur copilot.nabla.com et le [Chrome store web](#)

<https://www.nabla.com/fr/>

Plug in médical

DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains

Yanis Labrak^{*1,4} Adrien Bazoge^{*2,3} Richard Dufour² Mickael Rouvier¹

Emmanuel Morin² Béatrice Daille² Pierre-Antoine Gourraud³

¹LIA, Avignon Université ²LS2N, UMR CNRS 6004, Nantes Université

³Clinique des données, CHU de Nantes, Nantes Université ⁴Zenidoc

{firstname.lastname}@univ-avignon.fr

{firstname.lastname}@univ-nantes.fr

Abstract

In recent years, pre-trained language models (PLMs) achieve the best performance on a wide range of natural language processing (NLP) tasks. While the first models were trained on general domain data, specialized ones have emerged to more effectively treat specific domains. In this paper, we propose an original study of PLMs in the medical domain on French language. We compare, for

mal performance in specialized domains, such as finance (Yang et al., 2020), medical (Yang et al., 2022) or traveling (Zhu et al., 2021), can only be achieved using PLMs adapted to the targeted conditions.

The adaptation of language models to a domain generally follows two strategies. The first is the training *from scratch* of a new model using only textual data of the targeted specialty. The second

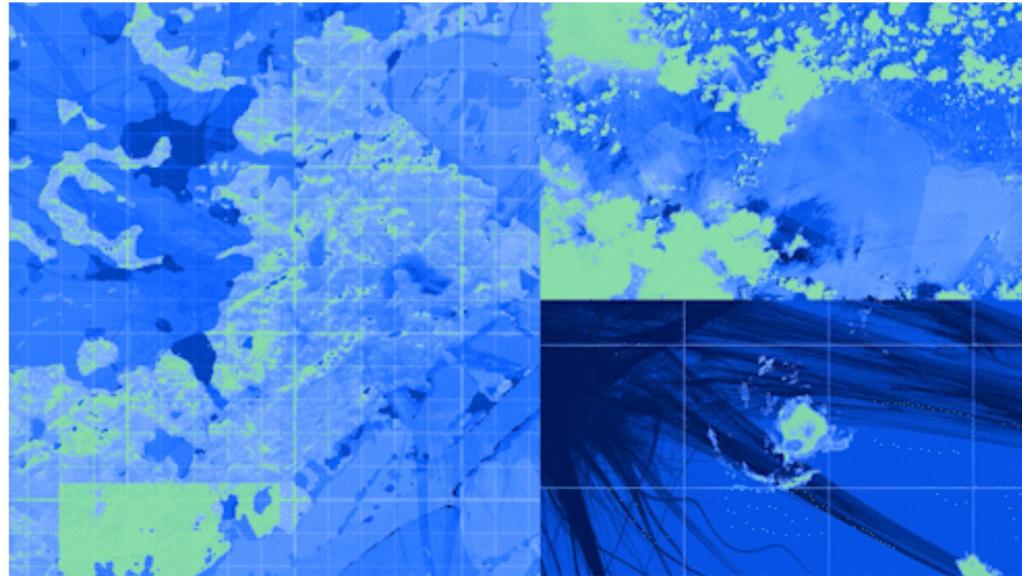


Noté 4,7/5 par nos utilisateurs sur copilot.nabla.com et le [Chrome store web](https://chrome.google.com/webstore/detail/copilot-nabla)

Open-Source Geospatial Foundation Model

IBM and NASA open source the largest geospatial AI foundation model on Hugging Face

The move aims to widen access to NASA satellite data and accelerate climate-related discoveries.



Le futur des modèles génératifs : la robotique ?

- Couplage langage naturel, vision, parole **et action**

Conclusion

Les paradoxes du deep learning

- Vision, le jeux et le langage naturel
- Big is beautiful & Multitask is better
- Open source vs. Propriétaire

La moyenne pour le meilleur et pour le pire

- Un super moteur de recherche en langage naturel
- Aide les programmeurs
- Hallucine et reproduit

Quelle sera la prochaine rupture ?

- Small is beautiful
- Intégration : Plug in
- Spécialisation -> IA générale

Roadmap

1. Comment on est on arrivé là ?
 - Deep learning, renforcement / données+DNN+GPU
 - Word2vec, LLM -> Bloom / foundation model multitâches (multi langues) / plus et plus gros
 - ChatGPT = chatbot+LLM / confiance ce qu'on peut et ce qu'on en peut pas (Moyenne)...
2. Que peut-on faire avec chat GPT et les autres LLM ?
 - Questions, letter, traduction
 - Codage maths analyse de données
 - Biblio,
3. Le futur des LLM
 - Petits opesource (Faraday) spécialisé personnalisés
 - Plug in integration dans des applis (mission contrôle)
 - LLM et robotique

PLM4ALL (LLM4ALL)

<https://gitlab.inria.fr/synalp/plm4all>

Hugging face : les modèles sont disponible ici (pytorch)

Hugging face : Librairie multi GPU : accelerate
(prototypage rapide...) – Lora déjà intégrée

DeepSpeed + zero (pour faire du fine tuning)

NVIDIA/Megatron optimisé (sans doc !)

Colossal-AI (tout nouveau tout beau)

PLM4All

Objectives

- Facilitate usage of [Jean Zay](#) for researchers on common use cases in Natural Language Processing
- Publish example scripts to perform inference, finetuning, prompt tuning... on open large language models

Useful Jean Zay docs

- Where to find the models: <http://www.idris.fr/jean-zay/gpu/jean-zay-gpu-dataset.html>
- Jean Zay cheat sheet: <http://www.idris.fr/media/su/idrischeatsheet.pdf>

Roadmap

- Complete DSDIR with models:
 - Flan-T5-xxl
 - Flan-UL2 20B [Google blog](#)
 - OPT-IML
 - Tk-11b
- Define and prioritize target use cases in this git
- Share example scripts for:
 - inference with Bloom
 - DS-finetuning Bloom medium size
 - prompt tuning (freeze model and train input vectors; see e.g. [PEFT library](#))
 - prefix tuning (freeze model and train input vectors at every layer)

Example Tasks

Pretraining and finetuning of large encoder-decoder

cf. <https://aclanthology.org/2020.webnlg-1.13>.pdf Contact: "Anna Nikiforovskaja" anna.nikiforovskaja@loria.fr, Claire Gardent

- Model: mT5-large
- Pre-training
 - (1) on monolingual data (800K steps): [Russian Wikipedia](#) and [English Wikipedia](#)
 - (2) on parallel EN/RU data (100K steps): [WMT Data](#)
- Fine tuning on WebNLG training data: [English](#) and [Russian](#)
 - One model per language
 - One model for both languages: Add prefixes to identify the tasks. Each task is weighted by the size of its training corpus.
- Test on WebNLG Dev data: [English](#) and [Russian](#)

Dialogues

- ChatGPT?
 - Qu'est-ce que chatGPT ?
 - peux tu comparer les positions du ministères de l'enseignement supérieur et de la recherche avec celle du ministère des finances sur la stratégie d'accélération en intelligence artificielle ?
- Bing
 - Peux tu résumer la position de la cours des compte sur l'état de la recherche en IA aujourd'hui, en justifiant les mesures proposées par rapport aux actions menées et leur bilan ?
 - peux tu détailler la position de la cour et notamment les mesures proposées. Sont-elles pertinente à ton sens ?
 - What are the main differences between Helmi et al. 2018 and Belokurov et al. 2018?
- Copilot :
 - Can you give me a python function computing the lasso regression coefficient of a design matrix X with a target vector y and a regularisation parameter λ , using a componentwise approach
 - I'm not shure you did well with the intercep can you try again?
- Bard
 - Can you explain how researchers can use large language such as Bard or chatGPT to do research?
-