



# GPT - 3

Antoine Mairesse - IA Le Mans  
Janvier 2021



# Sommaire

## 1. Modèle de langage

## 2. De GPT à GPT-3

- a. GPT
- b. GPT-2
- c. GPT-3

## 3. Et la suite ?

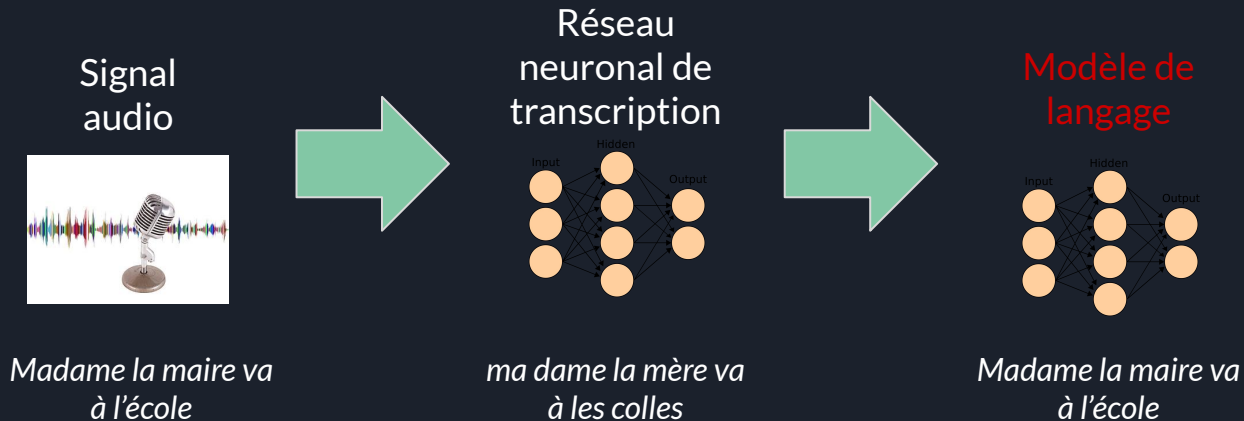
- a. DALL-E
- b. CLIP

Modèle de langage



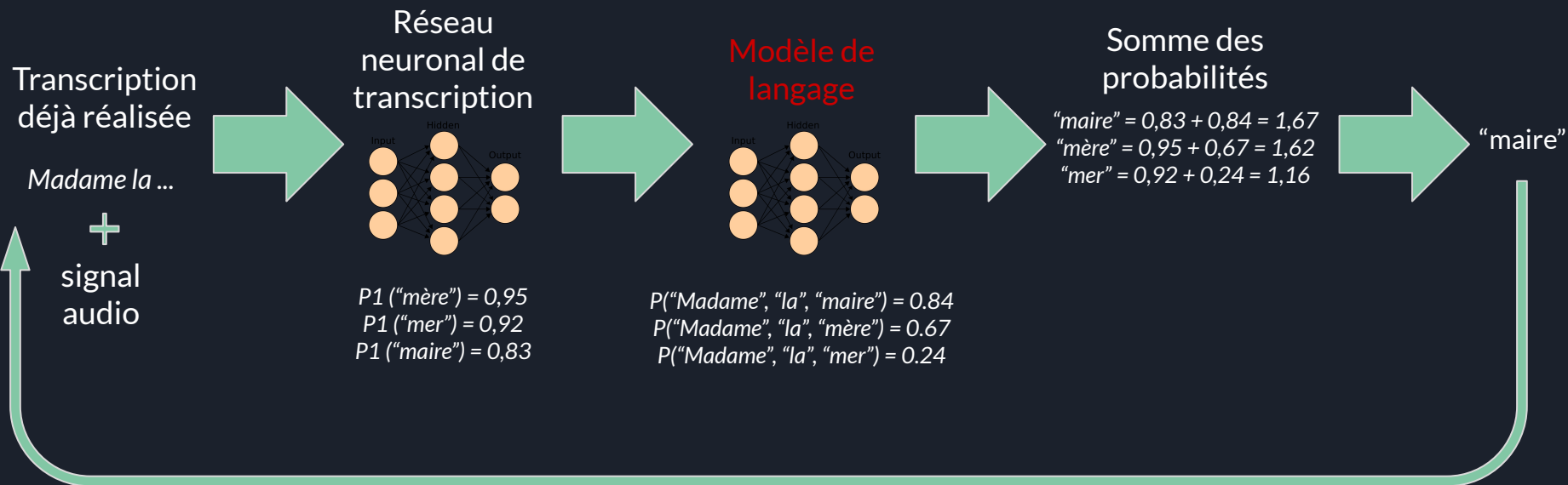
# Modèle de langage : usage

- Utilisé initialement en sortie d'un réseau neuronal de reconnaissance de la voix, pour fiabiliser la transcription
  - Notamment pour les transcriptions syllabiques pour choisir le bon "mot"
  - Mais aussi pour un modèle de mot pour construire les phrases les plus probables



# Modèle de langage : principe de fonctionnement

Un modèle de langage donne la probabilité d'une séquence de mots





# Modèle de langage : entraînement

- Extrêmement simple et on ne manque pas de données !
- On entraîne simplement un réseau neuronal à prédire le prochain mot de la phrase.
- On peut donc utiliser n'importe quelle source écrite
  - Wikipédia,
  - Les articles de presse,
  - Les romans,
  - Les blogs, les forums, etc.



# Modèle de langage : usage complémentaire

- Pour les anciens du Meetup nous avons déjà expérimenté avec un modèle de langage entraîné sur les textes de personnages anciens...
- Le principe
  - a. On choisit un mot au hasard
  - b. On applique le modèle pour prendre le mot le plus probable à suivre
  - c. On itère au point 'b' en partant de la phrase ainsi complétée pour construire des textes fictifs
- Ça génère du texte fictif avec le style de cette personne... C'était fun mais pas très réaliste et utile, mais ça c'était il y a 3 ans, une éternité dans ce domaine...

De GPT à GPT - 3







# De GPT à GPT - 3

- OpenAI, créateur de la série des “GPT” (Generative Pre-trained Transformer)

- GPT, premier du nom, sorti en Juin 2018

- 0,96 petaflops-days et 5 Go de données pour l'entraîner
- 117 millions de paramètres

- GPT-2, sorti en février 2019

- Plusieurs dizaines de petaflops-days et 40 Go de données pour l'entraîner
- 1,5 milliard de paramètres

- GPT-3, sorti en mai 2020

- Plusieurs milliers de petaflops-days et 45 To de données pour l'entraîner
- 175 milliards de paramètres

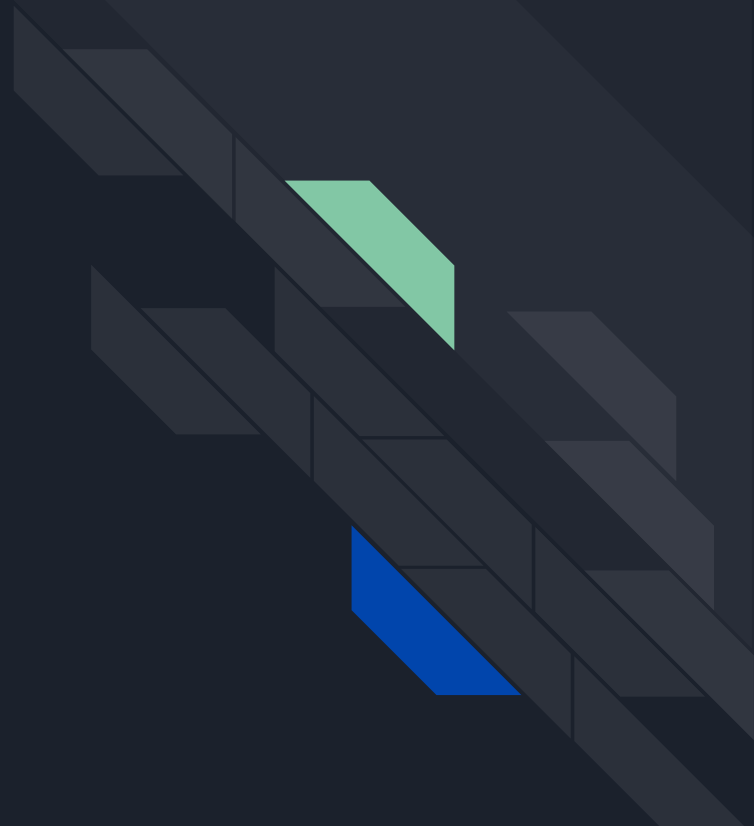
x10



x100



GPT - Juin 2018  
*(soit il y a 2,5 ans)*

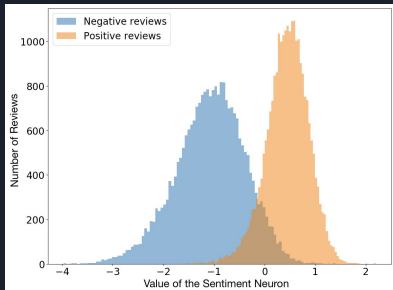


# GPT - L'inspiration

<https://openai.com/blog/language-unsupervised/>

- Précédemment les modèles étaient entraînés sur des corpus spécifiques à la tâche
  - Par exemple pour un modèle de langage pour une reconnaissance vocale en anglais on utilisait le corpus d'entraînement du texte transcrit, c'est tout !
  - Pour imiter Voltaire, on aurait entraîné le modèle sur les textes de Voltaire uniquement

- OpenAI avait détecté qu'une approche non-supervisée avec un volume suffisant de données permettait au modèle d'apprendre des features très pointues



- <https://openai.com/blog/unsupervised-sentiment-neuron/>
- Avec un modèle non-supervisé ils avaient découvert qu'un des neurones du réseau encodait systématiquement le "sentiment" du texte

# GPT - La révolution technique

<https://openai.com/blog/language-unsupervised/>

- La révolution était donc d'utiliser un pré-entraînement non-supervisé pour un modèle de langage plus traditionnel
  - Generative Pre-trained Transformer
- OpenAI a “juste” utilisé :
  - Un pré-entraînement sur du texte issu de différents livres
  - Puis un fine-tuning sur la tâche à réaliser, en gardant un double objectif cumulant la probabilité de la tâche à réaliser mais aussi du dataset de pré-entraînement pour ne pas sur-spécialiser le modèle

a. **Unsupervised Language Modelling** (Pre-training): For unsupervised learning, standard language model objective was used.

$$L_1(T) = \sum_t \log P(t_i | t_{i-k}, \dots, t_{i-1}; \theta) \quad (i)$$

where  $T$  was the set of tokens in unsupervised data  $\{t_1, \dots, t_n\}$ ,  $k$  was size of context window,  $\theta$  were the parameters of neural network trained using stochastic gradient descent.

b. **Supervised Fine-Tuning**: This part aimed at maximising the likelihood of observing label  $y$ , given features or tokens  $x_1, \dots, x_n$ .

$$L_2(C) = \sum_{x,y} \log P(y | x_1, \dots, x_n) \quad (ii)$$

where  $C$  was the labeled dataset made up of training examples.

Instead of simply maximising the objective mentioned in equation (ii), the authors added an **auxiliary learning objective** for supervised fine-tuning to get better generalisation and faster convergence. The modified training objective was stated as:

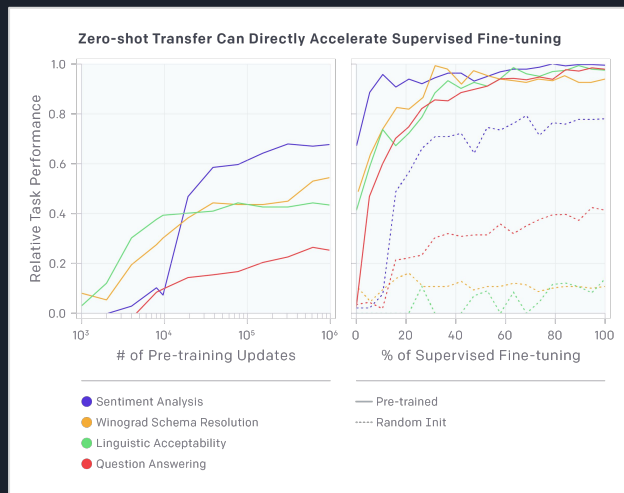
$$L_3(C) = L_2(C) + \lambda L_1(C) \quad (iii)$$

where  $L_1(C)$  was the auxiliary objective of learning language model and  $\lambda$  was the weight given to this secondary learning objective.  $\lambda$  was set to 0.5.

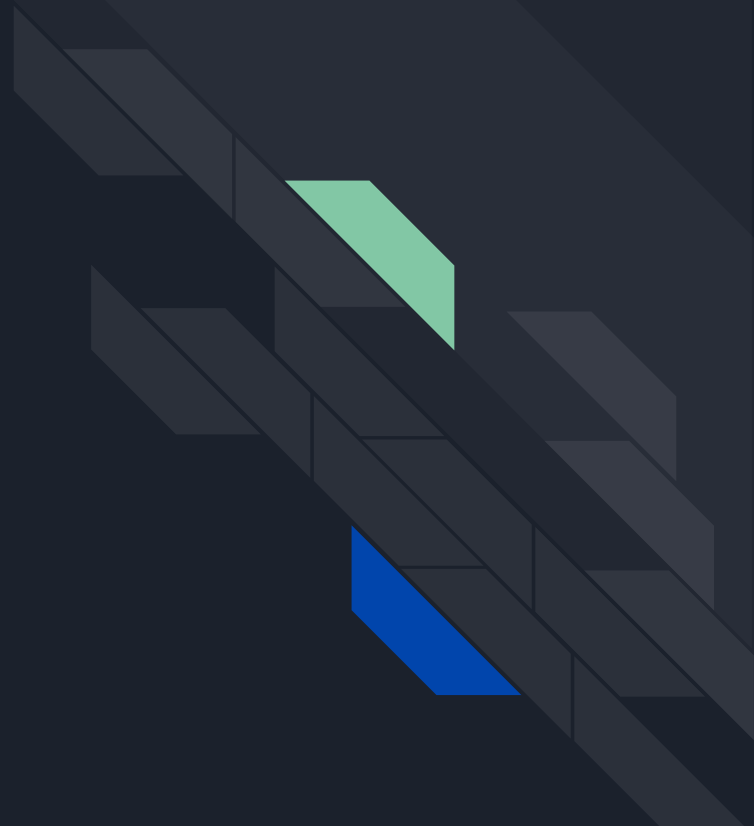
# GPT - Les résultats

<https://openai.com/blog/language-unsupervised/>

*GPT performed better than specifically trained supervised state-of-the-art models in 9 out of 12 tasks the models were compared on.*



GPT-2 - Février 2019  
*(soit il y a 2 ans)*





# GPT-2 - L'inspiration

<https://openai.com/blog/better-language-models/>

- Il s'agissait “simplement” de pousser l'expérimentation avec 10 fois plus de données
  - Aucun progrès significatif sur le modèle lui-même
  - L'envie de voir jusqu'où un entraînement non-supervisé pouvait donner de bons résultats sur des tâches spécifiques, sans faire de fine-tuning (“zero-shot learning”)
- Mais “pousser” l'expérimentation a nécessité un travail énorme et invisible
  - Pour être capable d'entraîner en masse sur des architectures multi-machines
    - + de 100 GPUs utilisés pendant des mois
    - Consolider un entraînement distribué sur plusieurs machines en quasi-temps réel
  - Pour trouver un moyen de créer un dataset aussi énorme et de qualité
    - Ils ont exclu Wikipedia pour privilégier les articles bien noté de Reddit par exemple
    - Ils ont développé une stratégie pour “nettoyer” les pages et obtenir du texte brut sans HTML mais en gardant le mieux possible le contenu
    - Etc.



# GPT-2 - La révolution technique

<https://openai.com/blog/better-language-models/>

- Le modèle n'a pas changé mais l'intelligence a été de réussir à le rendre "multi-tâche" sans en faire une donnée d'entrée encodée de manière explicite

$P(\text{output}|\text{input})$    $P(\text{output}|\text{input}, \text{task})$

- Si la taille de l'input est suffisamment "grand" alors on peut tout simplement l'utiliser pour expliciter la tâche
  - Ex. : si je veux traduire "children" en français avec un modèle qui ne fait que prédire le prochain mot d'une phrase, je peux utiliser différentes stratégies
    - INPUT : "The translation of children in french is :" ==> PREDICTED : "enfant"
    - INPUT : "sky = ciel ; dog = chien ; cat = chat ; children = " ==> PREDICTED : "enfant"

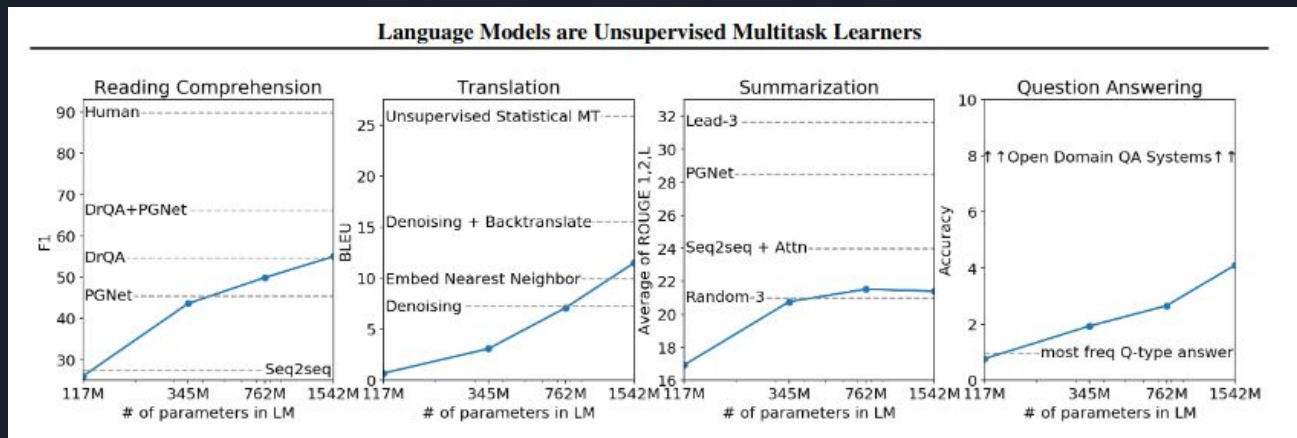


# GPT-2 - Les résultats

<https://openai.com/blog/better-language-models/>

*GPT-2 improved the then existing state-of-the-art for 7 out of 8 language modelling datasets in zero shot setting.*

*In French to English translation task, GPT-2 performed better than most unsupervised models in zero shot setting but did not outperform the state-of-the-art unsupervised model.*





# GPT-2 - Le buzz

<https://openai.com/blog/better-language-models/>

- La sortie de GPT-2 a fait du buzz !
  - Au-delà du résultat intéressant en multitâches qui était le coeur du papier publié...
  - ... la qualité de génération de texte était impressionnante
- La mise à disposition du modèle pré-entraîné a été progressive :
  - Février 2019 : modèle de 124 millions de paramètres
  - Mai 2019 : modèle de 355 millions de paramètres
  - Août 2019 : modèle de 774 millions de paramètres
  - Novembre 2015: modèle de 1,5 milliards de paramètres

SYSTEM PROMPT  
(HUMAN-WRITTEN)

*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*

MODEL COMPLETION  
(MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

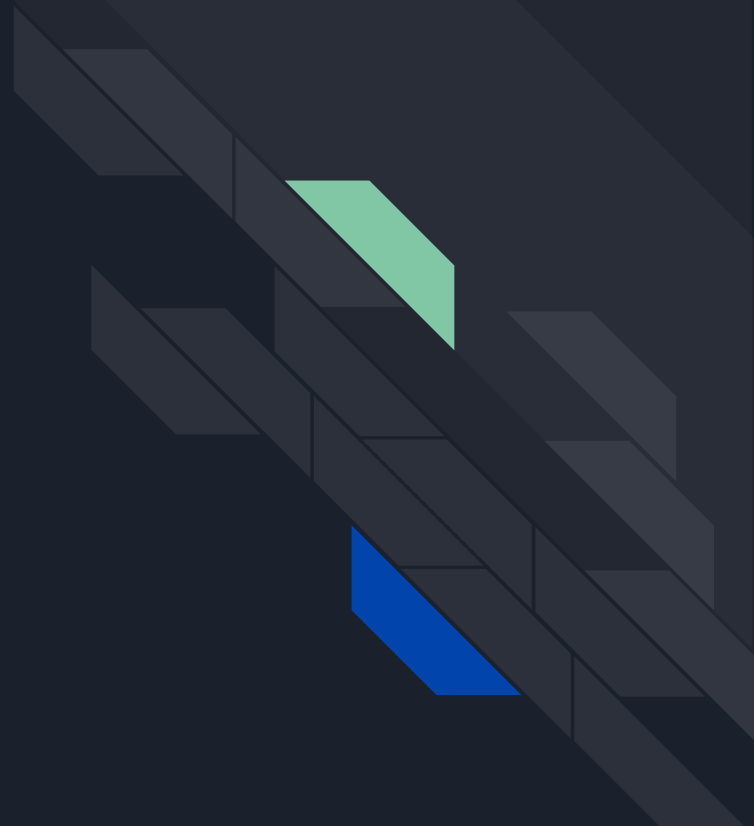
Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them - they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

GPT-3 - Mai 2020  
*(soit il y a 8 mois)*





# GPT-3 - L'inspiration

<https://arxiv.org/pdf/2005.14165.pdf>

- L'inspiration se résume assez rapidement avec cette phrase :

“A votre avis si on fait plus gros ça marche encore mieux ?”

- La capacité de génération de texte a fait le buzz avec GPT-2 mais la vraie surprise c'était tout ce que l'on pouvait finalement faire avec un modèle qui se contente de générer le mot suivant d'une phrase...
  - Traduire du texte
  - Résumer du texte
  - Répondre à des questions de compréhension sur un texte
  - ...



# GPT-3 - La révolution technique

<https://arxiv.org/pdf/2005.14165.pdf>

- Le modèle a
  - Été rendu plus “2 fois plus profond” : 96 layers et 96 attention heads (vs 48 pour le max de GPT-2)
  - Augmenté fortement son vocabulaire : 12.888 word embeddings (vs 1600 pour GTP-2)
  - Reçu un “prompt” 2 fois plus gros : 2.048 tokens (vs 1.024 pour GPT-2)
- Mais surtout le dataset et le modèle sont incroyablement massifs...
  - 45 To de texte...
  - 3,114 E23 FLOPS estimés pour l'entraîner, soit environ 4,6 millions de \$
  - Le modèle entraîné pèse 700 Go
    - À charger en mémoire du GPU pour faire uniquement l'inférence...

# GPT-3 - Les résultats

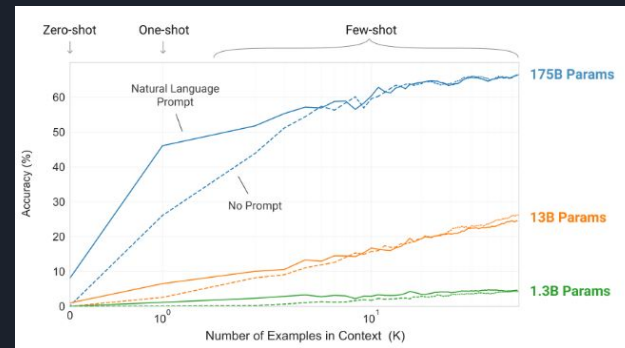
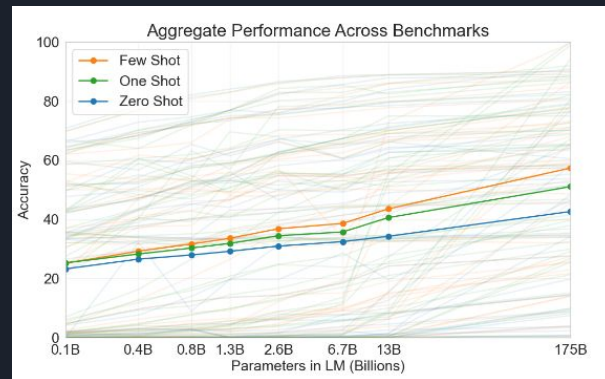
<https://arxiv.org/pdf/2005.14165.pdf>

Des modèles de différentes tailles pour pouvoir comparer la progression (comme GPT-2)

Pas de plafonnement de la performance en fonction de la taille du modèle

Gain de performance important sur une tâche avec le "few-shot"

*Its performance continues to scale with the # of parameters (with no end in sight).*



# GPT-3 - “One-shot”, “Few-shot” ?

- Apporte quasiment le même gain de performance qu’un fine-tuning !
  - Mais ici aucun fine-tuning, le modèle reste complètement générique...
- Et GPT-3 accepte un input de 2.048 mots avec un dictionnaire de 12.888 mots (\*)
  - Ça laisse de la place pour le contexte...

(\*) : les caractères unitaires et les sigles (=, +, -, /, >, <, ...) sont aussi dans le dictionnaire des “mots”

## The three settings we explore for in-context learning

### Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French: task description
2 cheese => prompt
```

### One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1 Translate English to French: task description
2 sea otter => loutre de mer example
3 cheese => prompt
```

### Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1 Translate English to French: task description
2 sea otter => loutre de mer examples
3 peppermint => menthe poivrée
4 plush girafe => girafe peluche
5 cheese => prompt
```

# GPT-3 - Produit commercial

- OpenAI a décidé de ne pas fournir le modèle mais plutôt de le commercialiser
  - Pour financer la recherche, pour maîtriser l'usage
  - Le modèle est uniquement accessible en webservice pour de l'inférence : <https://beta.openai.com/>

```
bash
$ curl -s -u $OPENAI_API_KEY -H 'Content-Type: application/json' https://api.openai.com/v1/completions -d '{
  > "model": "davinci",
  > "temperature": 0,
  > "stop": "\n",
  > "prompt": "Q: What is human life expectancy in the United States?
\nA: Human life expectancy in the United States is 78 years.\n\nQ: What
is the meaning of life?\n"
}' | jq
{
  "id": "cmpl-RRcXqSMfadADL1tst9gxPTJ9",
  "object": "text_completion",
  "created": 1591865181,
  "model": "davinci:2020-05-03",
  "choices": [
    {
      "text": "A: The meaning of life is 42.",
      "index": 0,
      "logprobs": null,
      "finish_reason": "stop"
    }
  ]
}
```

```
import openai
stop = "\n"
```

```
prompt = """Q: What is human life expectancy in the United States?
A: Human life expectancy in the United States is 78 years.
```

```
Q: Who was president of the United States in 1955?
A: Dwight D. Eisenhower was president of the United States in 1955.
```

```
Q: What party did he belong to?
A: He belonged to the Republican Party.
```


```
Q: Who was president of the United States before George W. Bush?
A: Bill Clinton was president of the United States before George W.
Bush.
```

```
Q: Who won the World Series in 1995?
A: The Atlanta Braves won the World Series in 1995.
```

```
Q: What year was the first fax sent?
A: ""
```

```
response = openai.Completion.create(model="davinci",
prompt=prompt, stop=stop, temperature=0)
```

```
print(response)
```

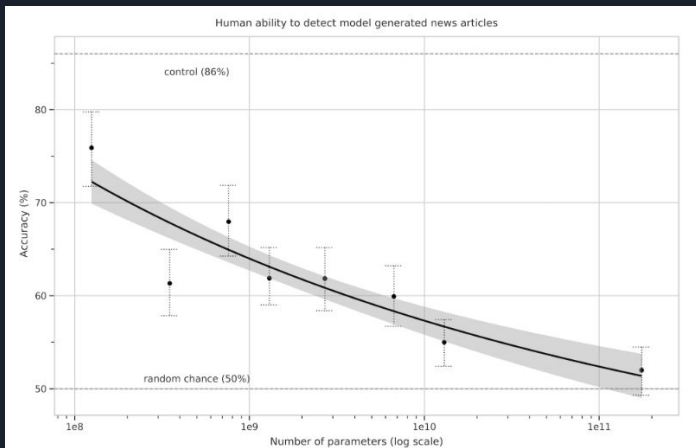


Q: What year was the first fax sent? A: The first fax was sent in 1843.

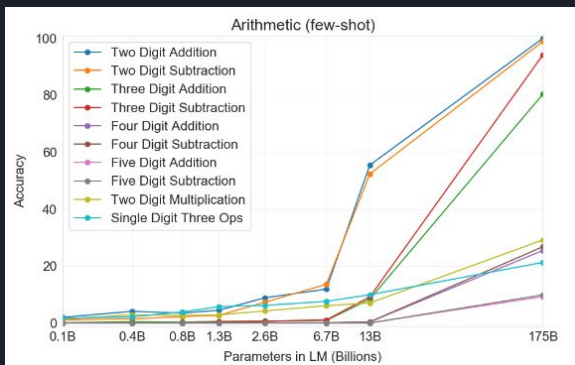


# GPT-3 - Que peut-il faire ?

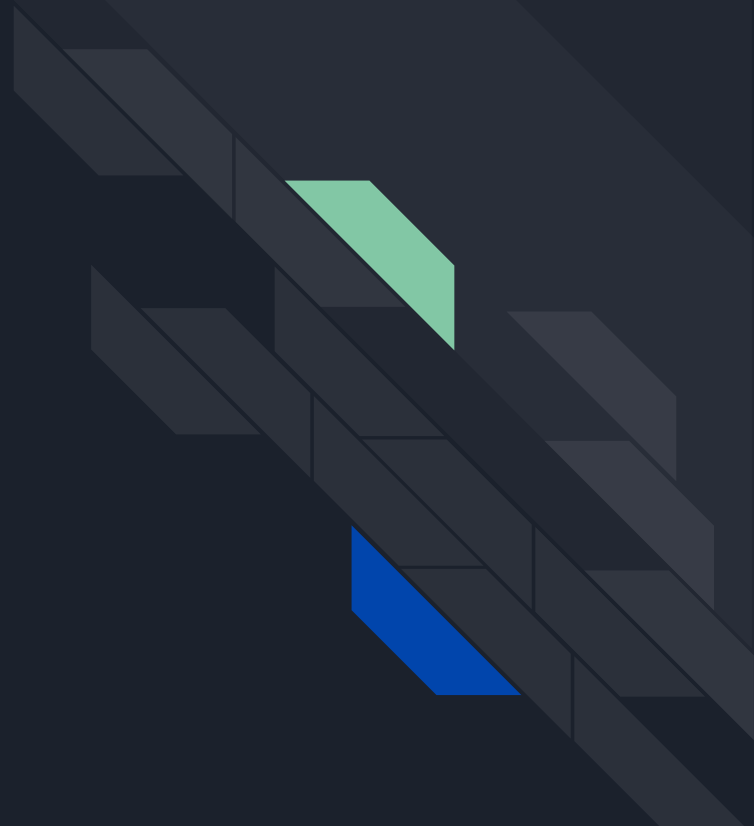
- Les “classiques” :
  - Traduction (multi-langues et dans tout sens), résumé de texte, réponse à des questions de connaissance générale, génération de texte, ...
  - Mais beaucoup mieux !
- Mais aussi :
  - Faire des opérations mathématiques
  - Reformuler simplement un passage juridique
  - Un assistant de saisie de code (code completion)
  - Un assistant de commande shell, de regex ou de sql en langage naturel
  - Un correcteur grammatical et de style



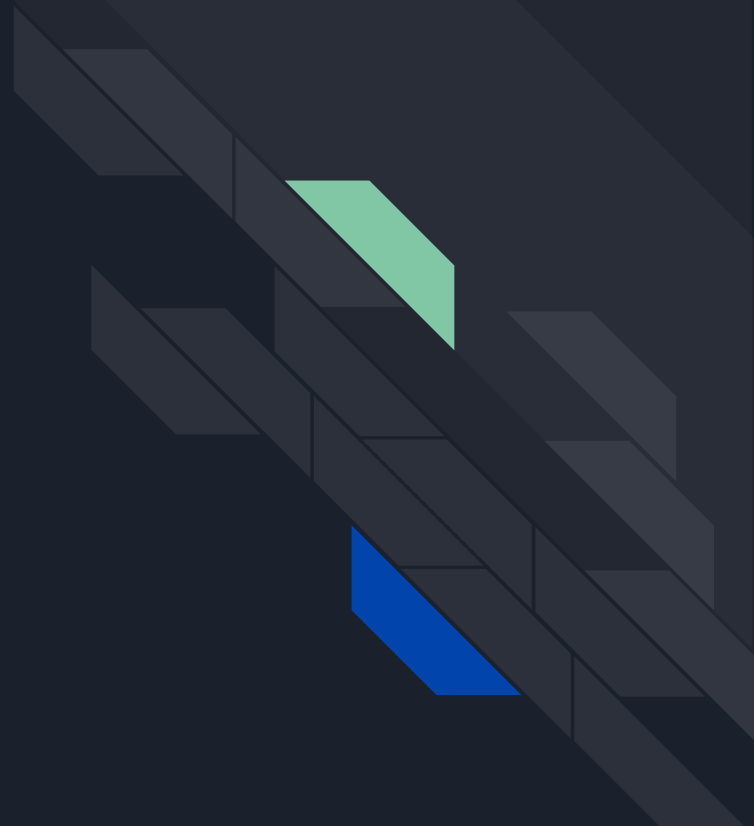
**Figure 3.13:** People's ability to identify whether news articles are model-generated (measured by the ratio of correct assignments to non-neutral assignments) decreases as model size increases. Accuracy on the outputs on the deliberately-bad control model (an unconditioned GPT-3 Small model with higher output randomness) is indicated with the dashed line at the top, and the random chance (50%) is indicated with the dashed line at the bottom. Line of best fit is a power law with 95% confidence intervals.



Et la suite ?



DALL-E - 5 janvier 2021  
*(soit il y a moins de 20 jours)*



# DALL-E : Creating Images from Text

## <https://openai.com/blog/dall-e/>

*DALL-E is a 12-billion parameter version of GPT-3 trained to generate images from text descriptions, using a dataset of text-image pairs.*

- L'input contient toujours du texte mais on y ajoute si nécessaire une image (partielle ou complète)
- L'output est une image

TEXT PROMPT an armchair in the shape of an avocado, an armchair imitating an avocado.

AI-GENERATED  
IMAGES



In the preceding visual, we explored DALL-E's ability to generate fantastical objects by combining two unrelated ideas. Here, we explore its ability to take inspiration from an unrelated idea while respecting the form of the thing being designed, ideally producing an object that appears to be practically functional. We found that prompting DALL-E with the phrases "in the shape of," "in the form of," and "in the style of" gives it the ability to do this.

When generating some of these objects, such as "an armchair in the shape of an avocado", DALL-E appears to relate the shape of a half avocado to the back of the chair, and the pit of the avocado to the cushion. We find that DALL-E is susceptible to the same kinds of mistakes mentioned in the previous visual.

TEXT PROMPT a male mannequin dressed in an orange and black flannel shirt and black jeans

IMAGE PROMPT



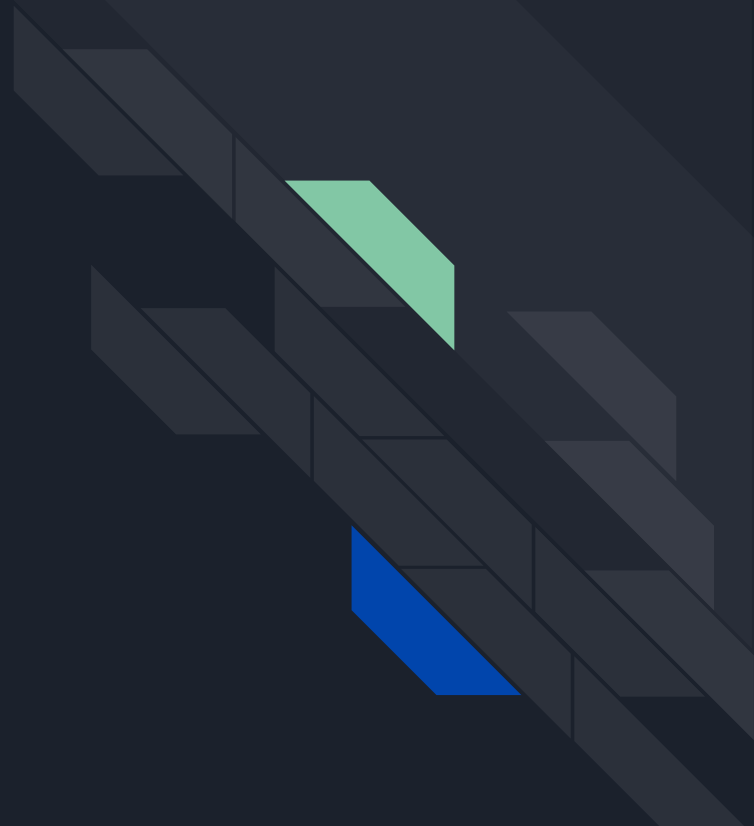
AI-GENERATED  
IMAGES



We explore DALL-E's ability to render male mannequins in a variety of different outfits. When prompted with two colors, e.g., "an orange and white bomber jacket" and "an orange and black turtleneck sweater" DALL-E often exhibits a range of possibilities for how both colors can be used for the same article of clothing.

DALL-E also seems to occasionally confuse less common colors with other neighboring shades. For example, when prompted to draw clothes in "navy", DALL-E sometimes uses lighter shades of blue, or shades very close to black. Similarly, DALL-E sometimes confuses "olive" with shades of brown or brighter shades of green.

CLIP - 5 janvier 2021  
*(soit il y a moins de 20 jours)*

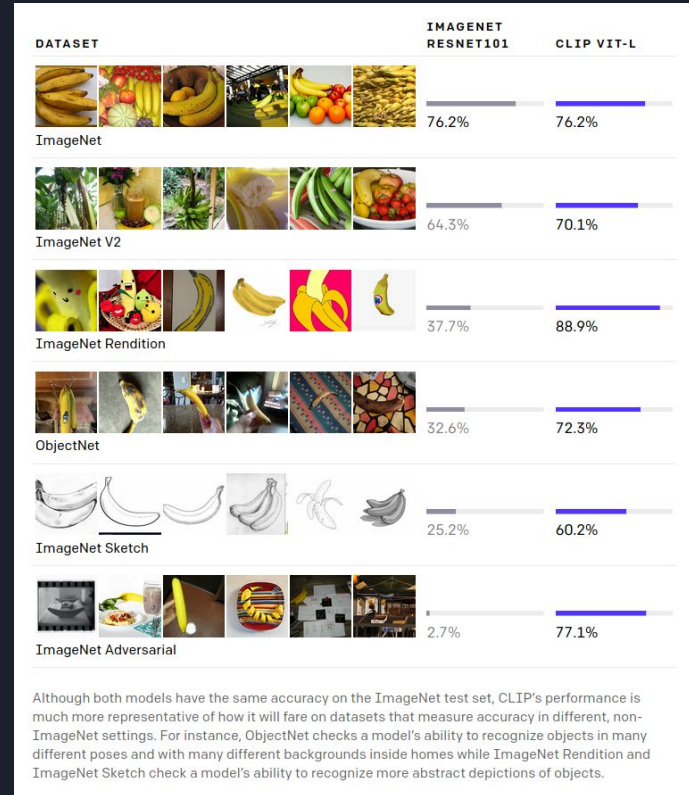


# CLIP : Connecting Text and Images

<https://openai.com/blog/clip/>

*CLIP (Contrastive Language-Image Pre-training) can be applied to any visual classification benchmark by simply providing the names of the visual categories to be recognized, similar to the “zero-shot” capabilities of GPT-2 and 3.*

Plutôt que de construire un modèle de classification avec un dictionnaire de classe figé, le modèle reçoit une image et un prompt textuel à compléter




# CLIP : Connecting Text and Images

<https://openai.com/blog/clip/>

**FOOD101**


**guacamole (90.1%)** Ranked 1 out of 101 labels



- ✓ a photo of **guacamole**, a type of food.
- ✗ a photo of **ceviche**, a type of food.
- ✗ a photo of **edamame**, a type of food.
- ✗ a photo of **tuna tartare**, a type of food.
- ✗ a photo of **hummus**, a type of food.

**SUN197**


**television studio (90.2%)** Ranked 1 out of 397



- ✓ a photo of a **television studio**.
- ✗ a photo of a **podium indoor**.
- ✗ a photo of a **conference room**.
- ✗ a photo of a **lecture room**.
- ✗ a photo of a **control room**.

**YOUTUBE-BB**

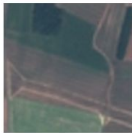
**airplane, person (89.0%)** Ranked 1 out of 23



- ✓ a photo of a **airplane**.
- ✗ a photo of a **bird**.
- ✗ a photo of a **bear**.
- ✗ a photo of a **giraffe**.
- ✗ a photo of a **car**.

**EUROSAT**

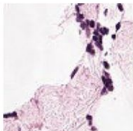
**annual crop land (12.9%)** Ranked 4 out of 10



- ✗ a centered satellite photo of **permanent crop land**.
- ✗ a centered satellite photo of **pasture land**.
- ✗ a centered satellite photo of **highway or road**.
- ✓ a centered satellite photo of **annual crop land**.
- ✗ a centered satellite photo of **brushland or shrubland**.

**PATCHCAMELYON (PCAM)**


**healthy lymph node tissue (22.8%)** Ranked 2 out of 2



- ✗ this is a photo of **lymph node tumor tissue**
- ✓ this is a photo of **healthy lymph node tissue**

**IMAGENET-A (ADVERSARIAL)**

**lynx (4.2%)** Ranked 5 out of 200



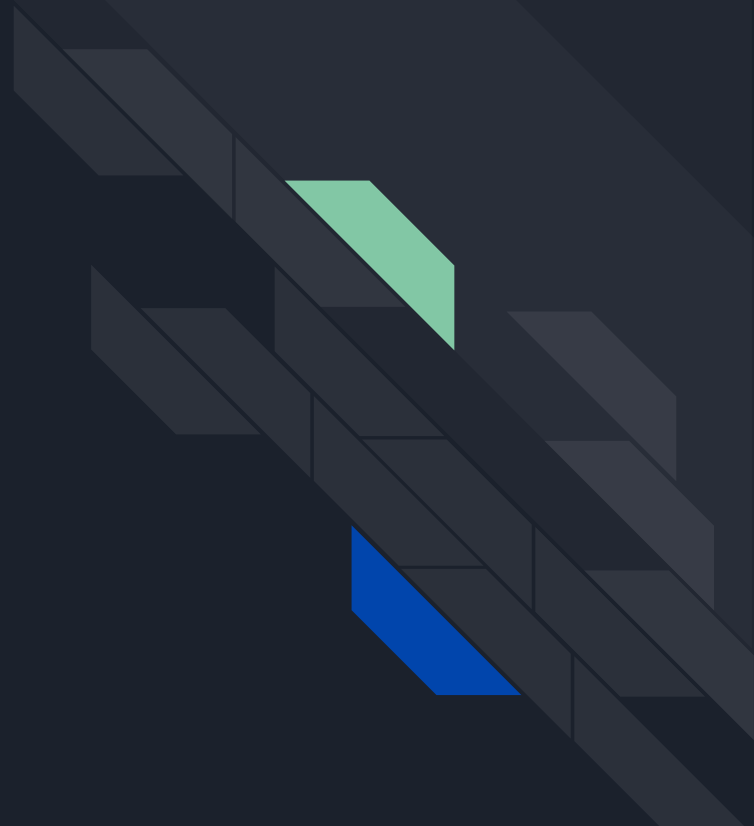
- ✗ a photo of a **fox squirrel**.
- ✗ a photo of a **mongoose**.
- ✗ a photo of a **skunk**.
- ✗ a photo of a **red fox**.
- ✓ a photo of a **lynx**.

Questions ?





# Références





# References

OpenAI's GPT-3 Language Model - A Technical Overview : <https://lambdalabs.com/blog/demystifying-gpt-3/>

The Journey of Open AI GPT models : <https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2>

How to Build OpenAI's GPT-2: <https://blog.floydhub.com/gpt2/>

OpenAI's GPT-3 Language Model: A Technical Overview : <https://lambdalabs.com/blog/demystifying-gpt-3/>

The GPT-3 economy : <https://bdtechtalks.com/2020/09/21/gpt-3-economy-business-model/>

GPT-3: A Hitchhiker's Guide : <https://lambdalabs.com/blog/gpt-3/>

GPT-3: The good, the bad and the ugly : <https://towardsdatascience.com/gpt-3-the-good-the-bad-and-the-ugly-5e2e5b7f0f66>