

chatGPT

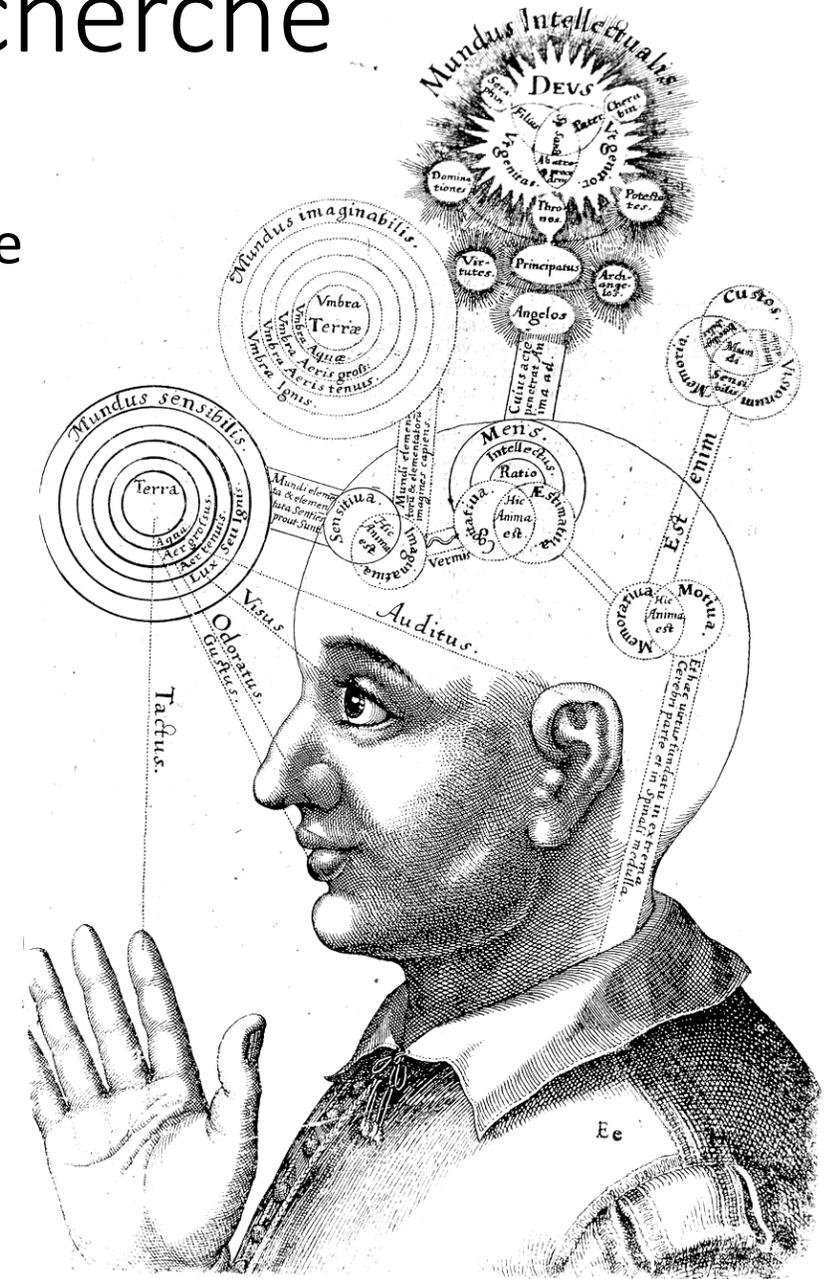
(r)évolution des modèles de langage génératif ?

Frédéric Bechet, Benoit Favre

Laboratoire Informatique et Systèmes - UMR 7020, AMU, CNRS

L'Intelligence Artificielle dans la recherche

- **Comprendre/imiter** les mécanismes de la cognition humaine
 - Selon plusieurs aspects
 - Perception
 - Représentation des connaissances
 - Raisonnement, résolution de problèmes
 - Planification et prise de décision
 - Apprentissage
 - L'IA est un domaine de recherche visant à construire des machines dotées, dans une certaine mesure, de ces capacités.



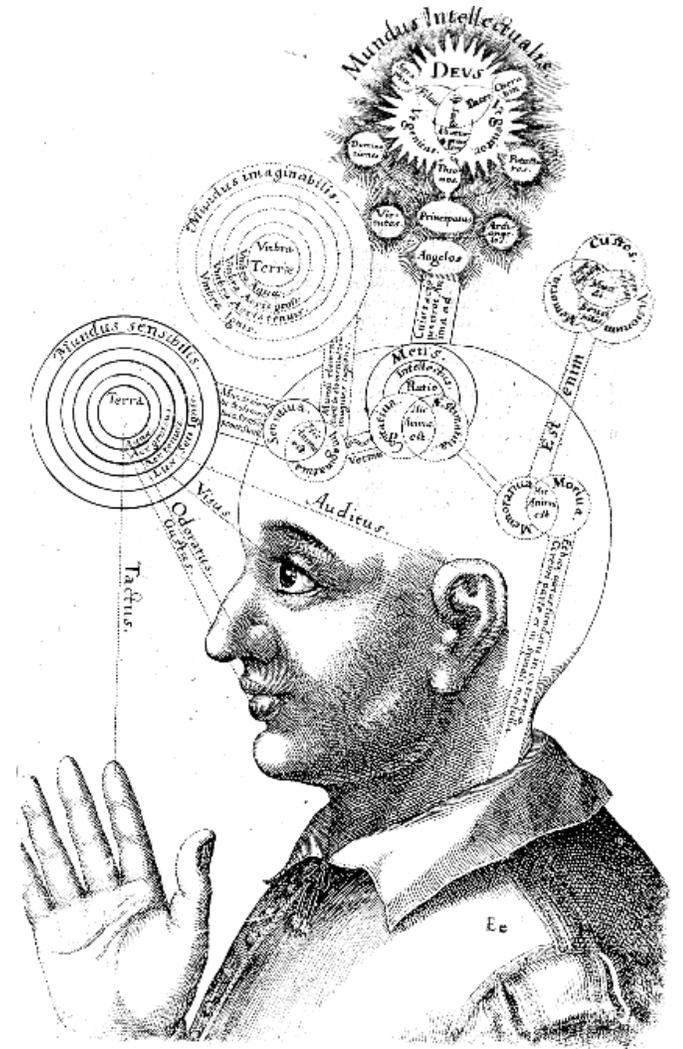
3 paradigmes pour définir des modèles d'IA

- Paradigmes

1. Règles et connaissances explicites définies par un expert humain
=> **systèmes experts**
2. Modèles statistiques formés sur des exemples humains
=> **apprentissage machine avec supervision**
3. Modèles statistiques entraînés pour atteindre un objectif
=> **autosupervision**

- Où se situe l'intelligence humaine dans ces paradigmes d'IA ?

- (1) et (2) peuvent être définis comme des modèles supervisés qui ont besoin de l'homme pour expliquer ce qu'il faut faire
 - explicitement => apprentissage avec un _professeur
 - implicitement => apprentissage avec un tuteur qui montre des exemples
- *(3) n'a besoin de l'homme que pour définir un objectif, pas pour apprendre une stratégie*

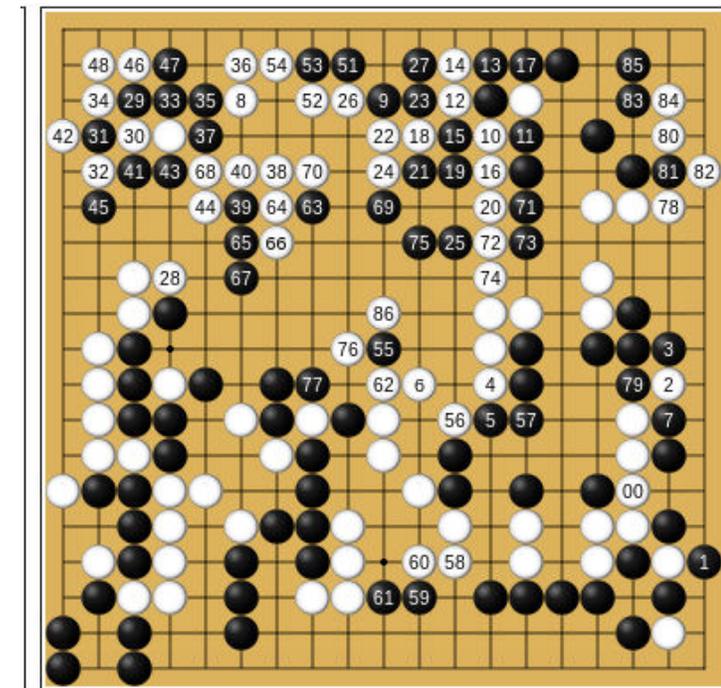


Exemples : machine contre humain dans les jeux

• AlphaGo

- Programme jouant au Go développé par DeepMind Technologies
- Modèle : combinaison des paradigmes (2) et (3)
 - (1) un premier modèle est entraîné sur une très grande base de données d'exemples de parties de Go (supervision humaine)
 - (2) le modèle est ensuite amélioré par **autosupervision** en jouant un très grand nombre de parties contre lui-même
- en mars 2016, il a battu Lee Sedol dans un match de cinq parties
 - première fois qu'une machine a battu un joueur professionnel de Go
 - modèle basé partiellement sur une expertise humaine car seul le premier modèle a besoin d'une supervision humaine
 - L'apprentissage autosupervisé peut conduire à la découverte de nouvelles stratégies « **non humaines** »

AlphaGo won by resignation. White 36 was widely praised.

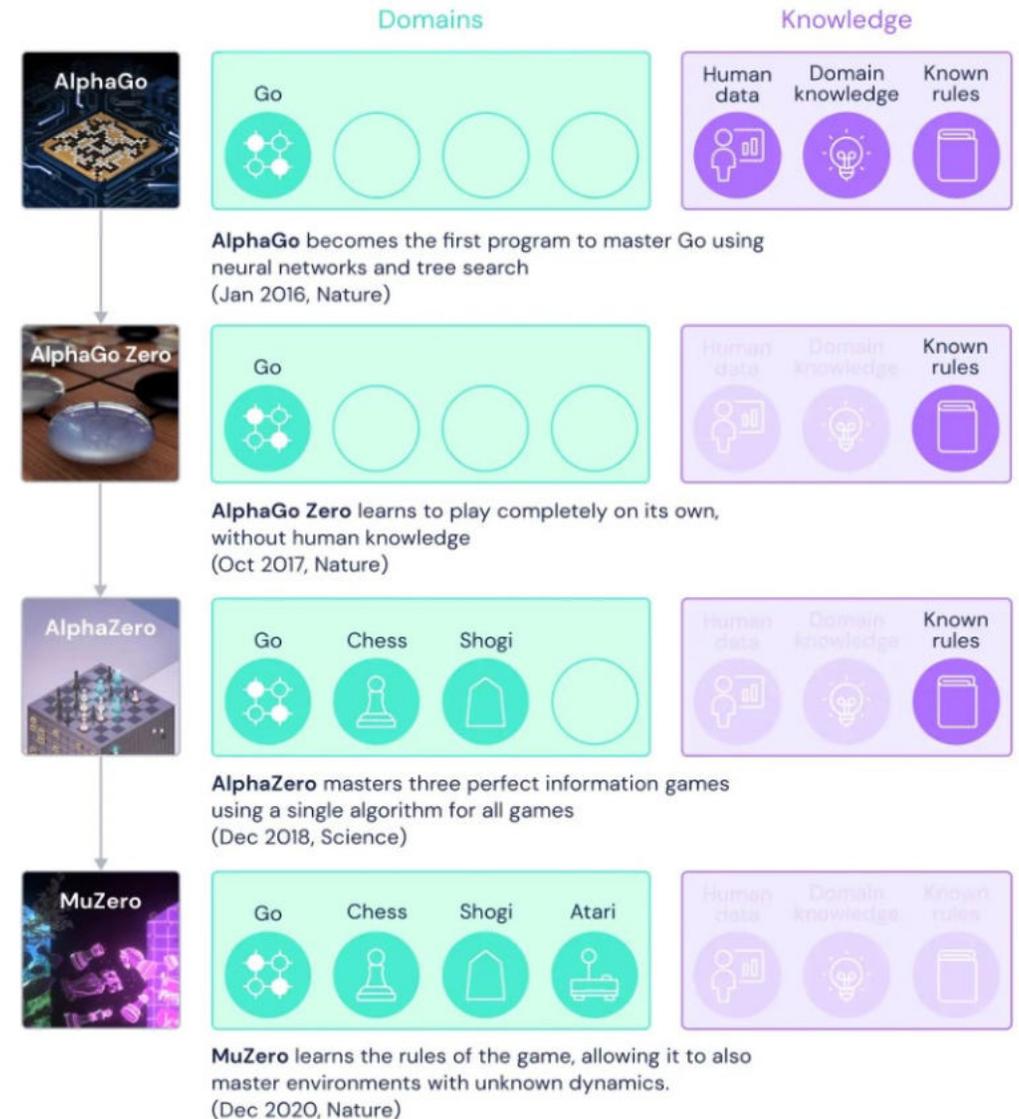


Moves 100–186 (149 at 131, 150 at 130)

Exemples : machine contre humain dans les jeux

• MuZero

- échecs, Go, jeux atari,
- modèle : pure (3) => uniquement **autosupervision** !
 - battre tous les programmes informatiques de jeux avec un seul algorithme
 - aucun besoin d'expertise humaine en dehors de la définition du but
 - But : **gagner** !!



Traitement Automatique du Langage (TAL)

- Modèles

- Modèles linguistiques

- Grammaires
 - Règles logiques

=> Paradigme 1 – basé sur des règles

- Modèles de classification

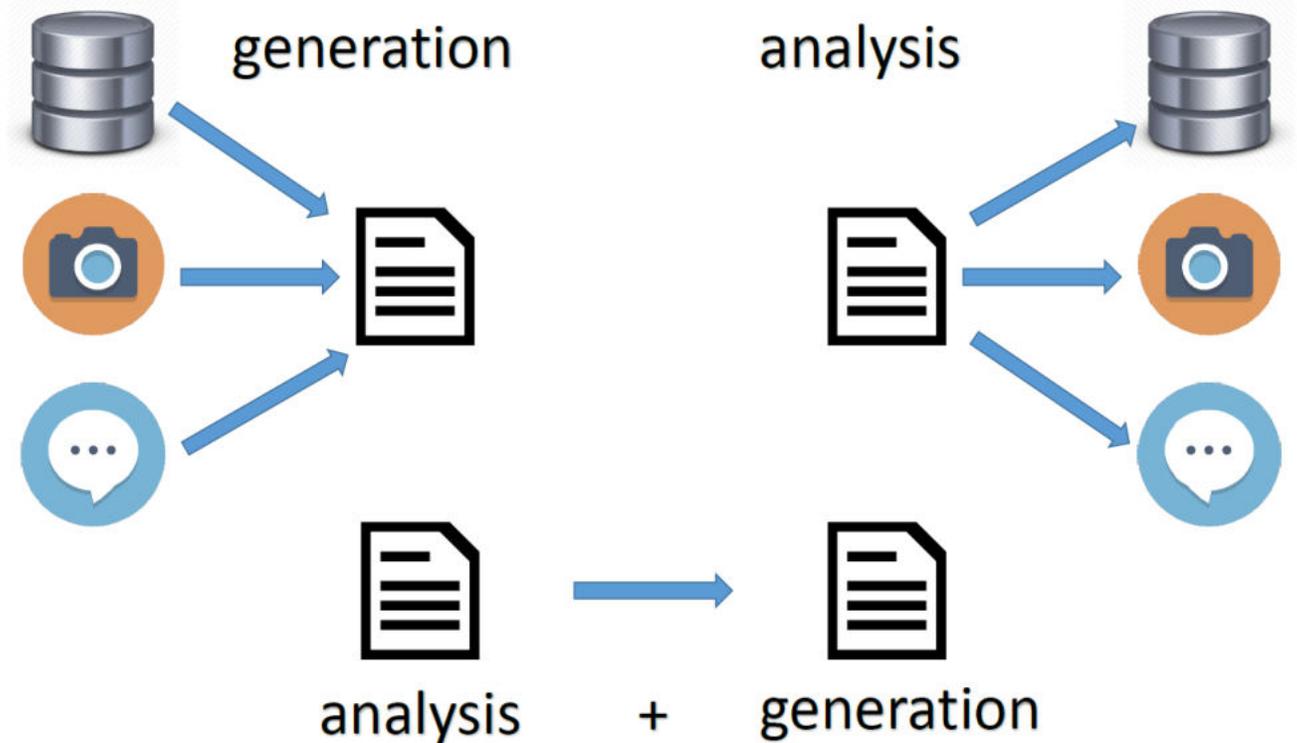
- Classification selon catégories prédéfinies
 - Nécessite une expertise humaine

=> Paradigme 2 – apprentissage supervisé

- Modèles de génération

- Apprentissage directement à partir du texte
 - Pas d'annotation supplémentaire
 - But : prédire le « mot » suivant

=> Paradigme 3 – apprentissage autosupervisé



Traitement Automatique du Langage (TAL)

- Modèles

- Modèles linguistiques

- Grammaires
 - Règles logiques

=> Paradigme 1 – basé sur des règles

- Modèles de classification

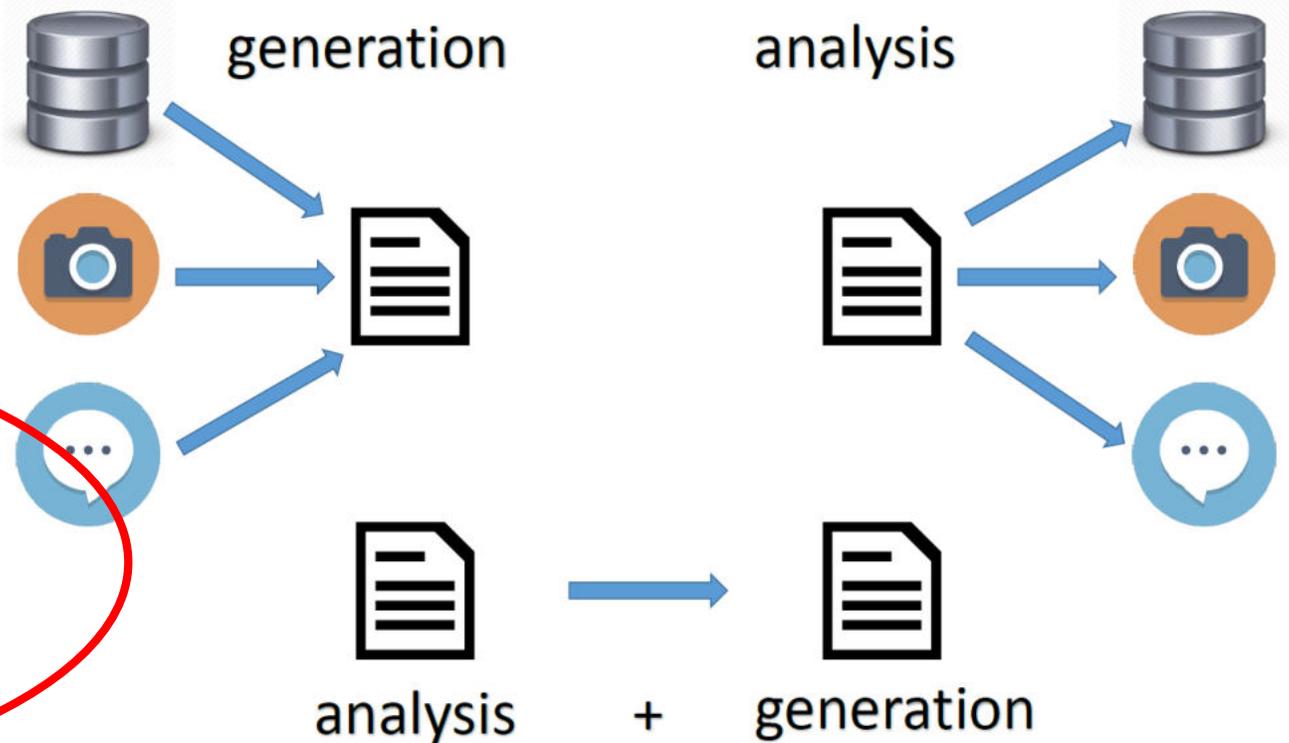
- Classification selon catégories prédéfinies
 - Nécessite une expertise humaine

=> Paradigme 2 – apprentissage supervisé

- Modèles de génération

- Apprentissage directement à partir du texte
 - Pas d'annotation supplémentaire
 - But : prédire le « mot » suivant

=> Paradigme 3 – apprentissage autosupervisé



Modèles de langage

Modèle de Langage

- Modélisation de la *forme* des mots et des phrases d'une langue
 - Soit modélisation formelle (grammaires, dictionnaires)
 - décision binaire : correct / incorrect
 - Soit modélisation numérique
 - prédiction d'un « *score* » de qualité linguistique

Modèle de Langage

- Modélisation de la *forme* des mots et des phrases d'une langue
 - Soit modélisation formelle (grammaires, dictionnaires)
 - décision binaire : correct / incorrect
 - Soit modélisation numérique
 - prédiction d'un « *score* » de qualité linguistique

Modèle de Langage : à quoi ça sert ?

- Indispensables pour la *génération* automatique de textes
 - Exemple : traduction automatique, transcription de parole
- Deux fonctions
 - Comment choisir parmi plusieurs phrases possibles ?
 - Comment générer du texte qui soit « correct » ?

Exemple : transcription de parole

Elle l'apprendra dès demain

Exemple : transcription de parole

Elle l'apprendra dès demain

Elle la prendra des deux mains

Exemple : transcription de parole

Elle l'apprendra dès demain

Elle la prendra des deux mains

Elle l'apprendra des deux mains

Exemple : transcription de parole

Elle l'apprendra dès demain

Elle la prendra des deux mains

Elle l'apprendra des deux mains

Elle la prendra dès demain

Exemple : transcription de parole

Elle l'apprendra dès demain

Elle la prendra des deux mains

Elle l'apprendra des deux mains

Elle la prendra dès demain

Aile l'a prend drap dé de main

Exemple : transcription de parole

Elle l'apprendra dès demain

Elle la prendra des deux mains

Elle l'apprendra des deux mains

Elle la prendra dès demain

Aile l'a prend drap dé de main

Hèl la pran dra dai de min

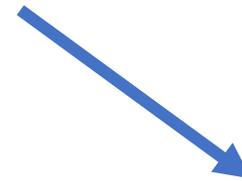
Exemple : transcription de parole

Elle l'apprendra dès demain	V
Elle la prendra des deux mains	V
Elle l'apprendra des deux mains	?
Elle la prendra dès demain	V
Aile l'a prend drap dé de main	X
Hèl la pran dra dai de min	X

Exemple : traduction automatique



Les scientifiques comptent
les baleines depuis l'espace



Les scientifiques comptent
les baleines de l'espace

Modèle de Langage

- But : estimation de la probabilité P d'une phrase S : $P(S)$

P(Elle l'apprendra dès demain) >> P(Elle l'apprendra des deux mains) >> P(Aile l'a prend drap dé de main)

- Quels modèles ?

- modèles statistiques → chaînes de Markov
- modèles prédictifs → modèles neuronaux

- Comment sont-ils entraînés ?

- corpus de textes sous format électronique
- Exemple : « *Web as a corpus* »

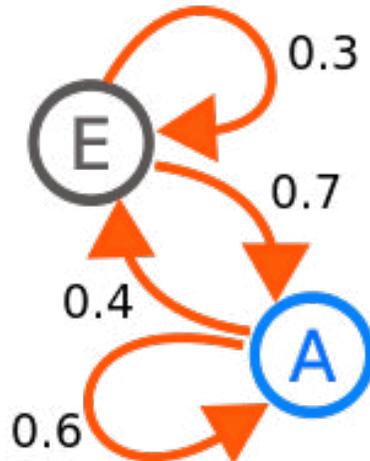
→ **7.2×10^{11} mots**

Modèle de Langage statistique

- Chaînes de Markov
 - modélisation de phénomènes aléatoires à mémoire courte
 - application à la modélisation du langage écrit (1913)
 - 20 000 lettres d'Eugène Onéguine d'Alexandre Pouchkine
 - premier modèle de langage statistique de l'histoire !!



Andreï Markov



Modèle de Langage statistique

- Modèles de langage n-grammes = chaîne de Markov
 - exemple modèle 3-grammes

$$\begin{aligned} P(\text{le chat boit du lait}) &= P(\text{le}) \\ &\quad \times P(\text{chat}|\text{le}) \\ &\quad \times P(\text{boit}|\text{le chat}) \\ &\quad \times P(\text{du}|\text{chat boit}) \\ &\quad \times P(\text{lait}|\text{boit du}) \end{aligned}$$

Modèle de Langage statistique

- Problème des évènements non vus : mots, groupes de mots

$$P(\text{le lézard boit du lait}) = P(\text{le})$$

$$\times P(\text{lézard}|\text{le})$$

$$\times P(\text{boit}|\text{le lézard})$$

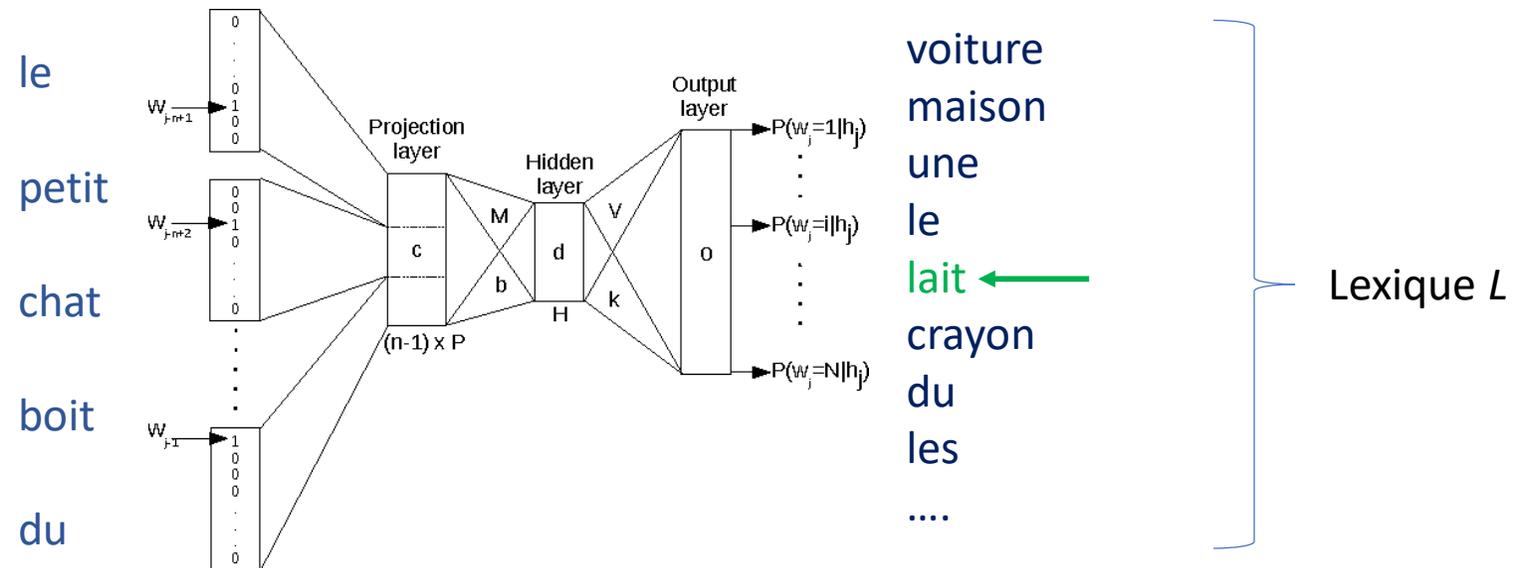
$$\times P(\text{du}|\text{lézard boit}) \longrightarrow 0$$

$$\times P(\text{lait}|\text{boit du}) \longrightarrow 0$$

Méthodes de redistribution de probabilités pour les évènements non vus.

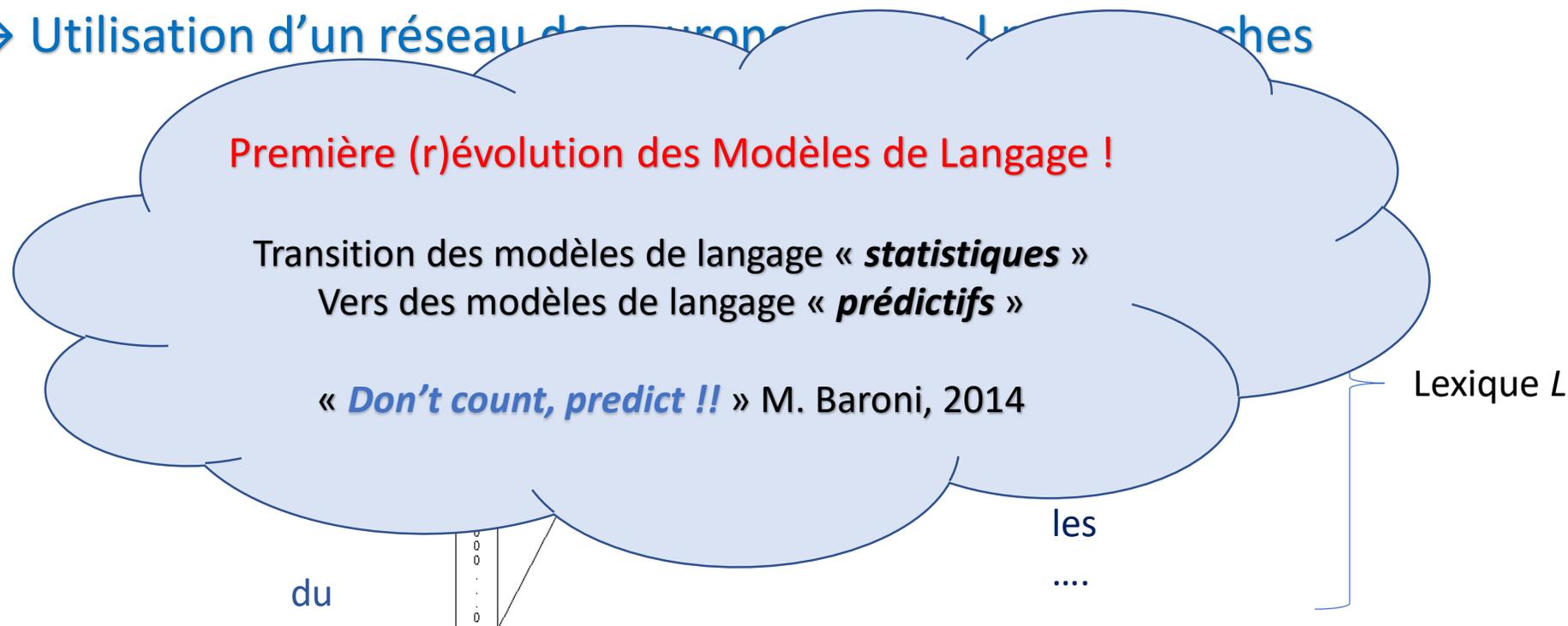
Modèle de Langage prédictifs

- Estimer la probabilité d'un mot w étant donné un contexte C : $P(w/C)$
 - Prédire un score pour tous les mots d'un lexique L étant donné C
- Utilisation d'un classifieur appris en mode *auto-supervisé*
 - Utilisation d'un réseau de neurone artificiel multi-couches



Modèle de Langage prédictifs

- Estimer la probabilité d'un mot w étant donné un contexte C : $P(w/C)$
 - Prédire un score pour tous les mots d'un lexique L étant donné C
- Utilisation d'un classifieur appris en mode *auto-supervisé*
 - Utilisation d'un réseau de neurones à couches



Représentation du lexique

- Qu'est ce qu'un mot pour un modèle de langage ?
 - *Un symbole → une entrée dans une base de données (dictionnaire)*
 - *Problème : pas de liens implicites entre les symboles*
 - *Cheval/poulain ; cheval/chevaux ; chat/chien ; dehors/dedans ;*
 - *Tous les liens doivent être explicites*
 - difficile d'explicitier tous les liens entre mots
 - *rugby ; jouer ; ballon ; XV ; basket ; panier ; olympique ; flamme ;*

Représentation du lexique

- Solution
 - Représenter un mot par un vecteur
 - liens entre mots = distances entre vecteurs dans un espace
 - possibilité d'effectuer des calculs dans l'espace lexical
 - Comment ?
 - mot = vecteur modélisant le contexte d'apparition du mot
- *sémantique distributionnelle*

Sémantique distributionnelle

“You shall know a word by the company it keeps !” (1957)

(pour connaître un mot, regardez ses compagnons)

- Un mot peut être défini par son contexte d'utilisation
 - **Modèles fréquentiels**
 - mot = vecteur de contexte
 - *Latent Semantic Analysis*
 - **Modèles prédictifs**
 - mot = vecteur de prédiction en fonction du contexte
 - *Word Embeddings* (ex: *Word2Vec*)



JOHN RUPERT FIRTH

ES04S, XXIV

Sémantique distributionnelle

“You shall know a word by the company it keeps !” (1957)

(pour connaître un mot, regardez ses compagnons)

- Un mot peut être défini par son contexte d'utilisation

- **Modèles**

- mot

Deuxième (r)évolution des Modèles de Langage !

Les Modèles de Langage peuvent servir à
« apprendre » des représentations des mots
directement à partir du texte, sans supervision

- **Mod**

-

- *Word Em*



JOHN RUPERT FIRTH

180AS, XXIV

Modèles de langage → et après ?

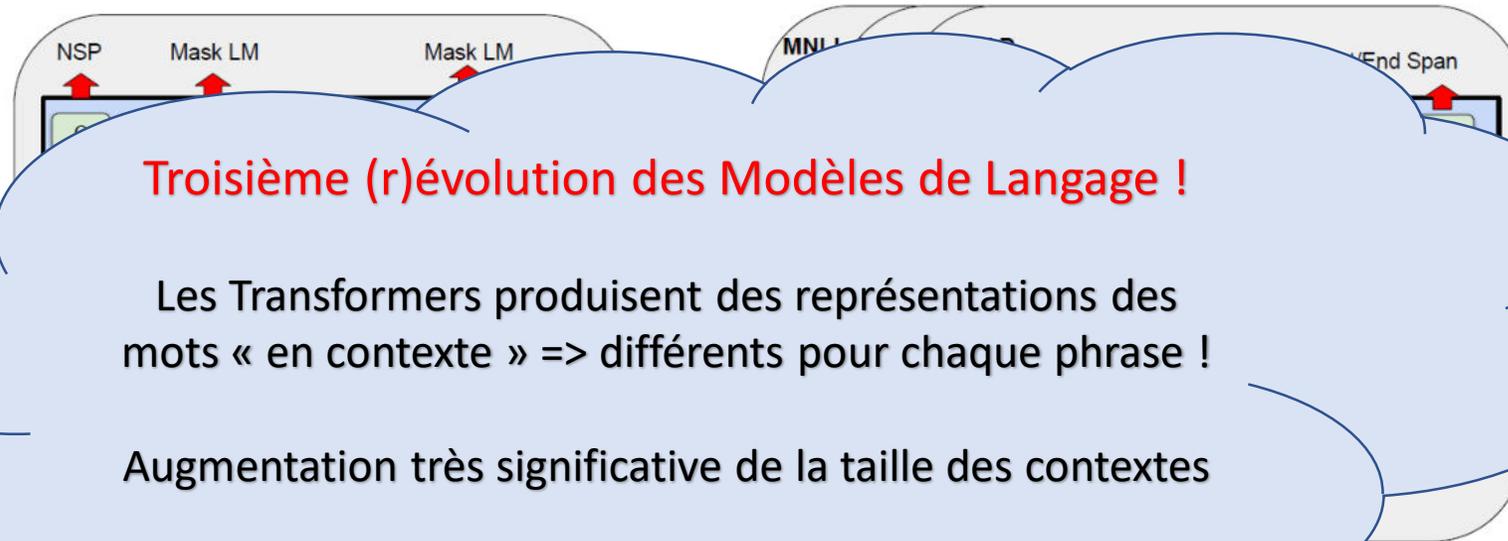
- Limitations des modèles de langage neuronaux + vecteur de mots
 - problème des mots inconnus
 - exemple : *découvrabilité*
 - problème de la polysémie
 - exemple : *avocats* → 1 vecteur / 2 sens
 - problème de la taille du contexte C pour prédire w
 - *Taille trop faible pour encoder des contraintes sémantiques complexes*

Modèles de langage « Transformer »

- nouveaux modèles proposant des solutions à ces problèmes
 - BERT – Deep Bidirectional Transformer
 - problème des mots inconnus
 - entrée : « briques » de lettres (*word piece*) pouvant former n'importe quel mot
 - problème de la polysémie
 - les vecteurs de mots deviennent contextuels → spécifique à chaque phrase
 - problème de la taille du contexte C pour prédire w
 - la taille du contexte augmente → modèles à base de **mécanisme d'attention**

Modèles de langage « Transformer »

- Paradigme du pré-entraînement par auto-supervision (masquage)
- + spécialisation du modèle par apprentissage supervisé sur une tâche



Troisième (r)évolution des Modèles de Langage !

Les Transformers produisent des représentations des mots « en contexte » => différents pour chaque phrase !

Augmentation très significative de la taille des contextes

très gros corpus généraliste sans annotation / spécifique à une tâche linguistique avec annotation

success story : BERT le couteau suisse du Traitement de la Langue

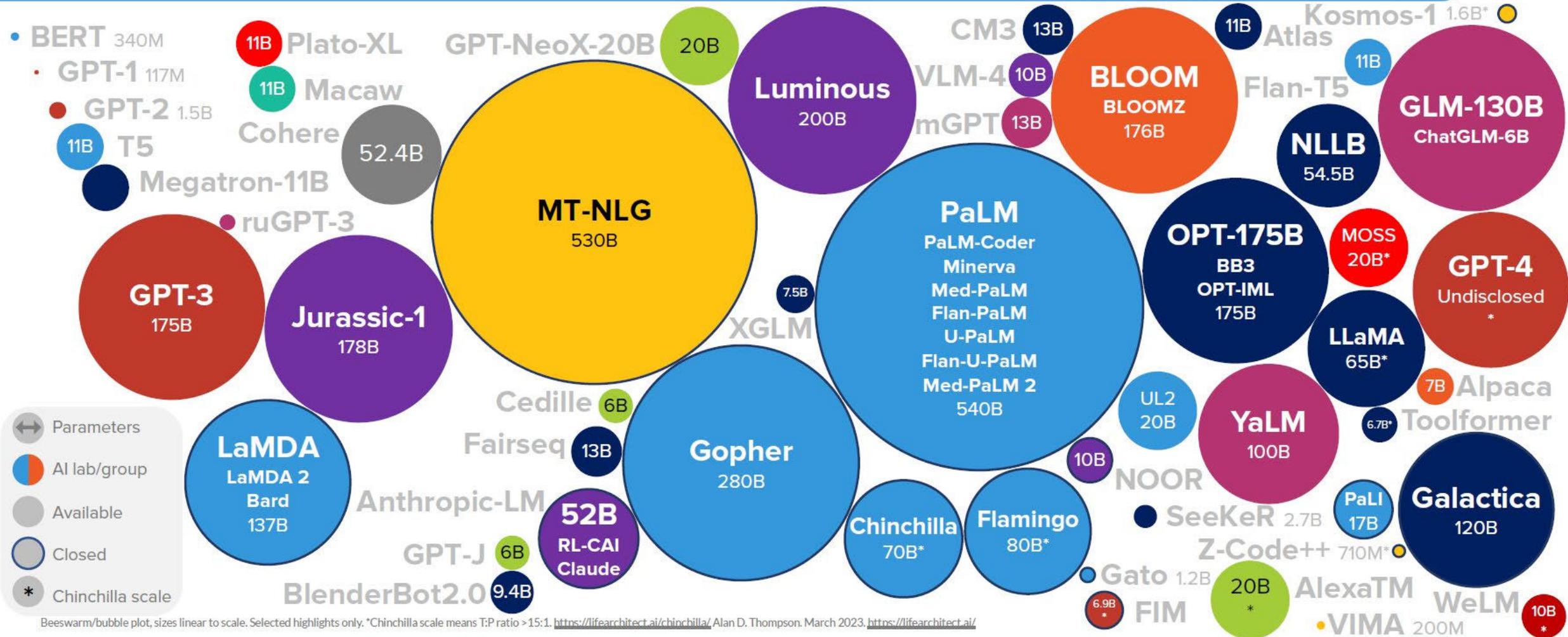
De BERT à GPT3

Generative Pre-trained Transformer

- Architecture « séquence à séquence »
 - Séquence de « tokens » en entrée => séquence de « tokens » en sortie
- Augmentation
 - de la taille du contexte, données d'apprentissage, des paramètres
 - BERT: 340 millions → GPT3: 175 milliards
 - du nombre de langues : apprentissage multilingue
 - > 50 langues dans Common Crawl
 - mais 93% de tokens en anglais
 - de la puissance de modélisation
 - au-delà de la forme des phrases ?
 - mais génère du texte qui « ressemble » aux données d'apprentissage
 - du coût : 4,6M\$ environ pour l'apprentissage

Dataset	#tokens (billions)
Common crawl (C4 filtered)	410
WebText2	19
Books1	12
Books2	55
Wikipedia	3
Total	499

LANGUAGE MODEL SIZES TO MAR/2023



Avènement des « Large Language Models »

Pourquoi est-ce qu'en augmentant le contexte, cela change tout !!

- prédire le mot d'après permet d'effectuer des tâches linguistiques
 - Contexte = « *le petit* » Mot=« *chat* »
 - Contexte = « *les cours de la bourse* » Mot=« *montent* »
 - Contexte = « *un synonyme du verbe flâner est* » Mot=« *déambuler* »
 - Contexte = « *si ma mère a une sœur, elle est ma* » Mot=« *tante* »
 - Contexte = « *la traduction du mot 'fleur' en anglais est* » Mot=« *flower* »
 - Contexte = « *[roman policier] et le coupable est* » Mot=« *le colonel* »
- modélisation allant au-delà de la forme des phrases
 - modélisation de connaissances
 - modélisation de raisonnements ?

Avènement des « Large Language Models »

Comment sont apprises ces tâches ?

- directement à partir des données
- sans supervision humaine
- apprentissage avec but: « prédire » !

(Radford et al. 2019,
["Language models are unsupervised multitask learners"](#))

"I'm not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile** [I'm not a fool].

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: "**Mentez mentez, il en restera toujours quelque chose,**" which translates as, "**Lie lie and something will always remain.**"

"I hate the word 'perfume,'" Burr says. 'It's somewhat better in French: 'parfum.'

If listened carefully at 29:55, a conversation can be heard between two guys in French: "**-Comment on fait pour aller de l'autre coté? -Quel autre coté?**", which means "**- How do you get to the other side? - What side?**".

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

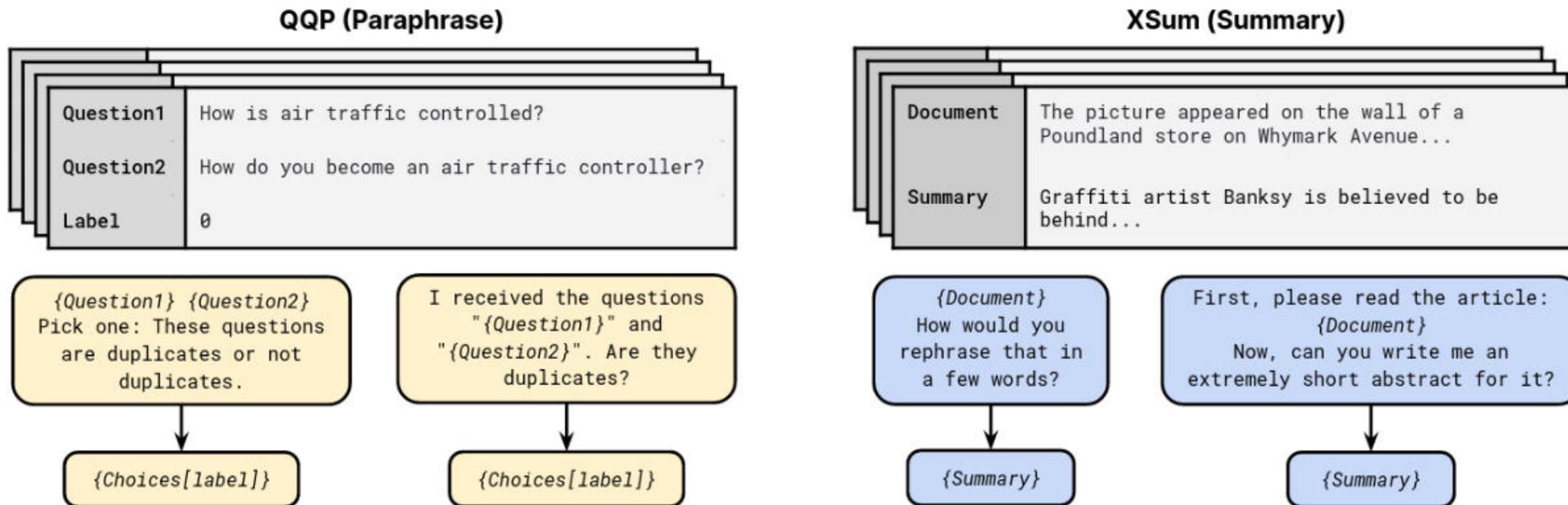
"Brevet Sans Garantie Du Gouvernement", translated to English: "**Patented without government warranty**".

Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

Avènement des « Large Language Models »

Comment sont apprises ces tâches ?

- En donnant des exemples => apprentissage supervisé (fine-tuning)
 - (Sanh et al, 2022; "[Multitask Prompted Training Enables Zero-Shot Task Generalization](#)")



Avènement des « Large Language Models »

Comment sont apprises ces tâches ?

- En donnant des exemples

- (S)

Quatrième (r)évolution des Modèles de Langage !

L'amorçage ou « prompting » permet d'avoir un modèle où la tâche à réaliser est décrite en langue naturelle lors de l'utilisation du modèle

Permet de généraliser sur des tâches *sans doute* non vues lors de l'apprentissage

On ne connaît pas toutes les tâches que peut réaliser un modèle !

{Choice

Amorçage (prompting)

- Principe : choisir le début du texte de manière à effectuer une tâche
 - *Le petit chaperon rouge. Il était une fois ...*
- Comment créer un prompt ?
 - Décrire la tâche à faire, donner l'entrée, puis solliciter la sortie
 - On peut mettre plusieurs exemples de résolution de la tâche avant de donner l'entrée
 - few-shot learning
 - Doit ressembler à du texte des données d'entraînement
- Prompts pour de nombreuses tâches
 - Traduction : *French Translation from English : cheese => ...*
 - Correction : *Poor English input: "The spider spun it's web". Good English input: ...*
 - Résolution d'un pronom : *The "she" in the sentence below refers to "Geraldine" True or False? ...*
 - Questions réponses : *Question: What is the boiling temperature of water? Answer: it is...*
 - Résumé automatique : *Summary of the following text...*
 - Génération : *A poem in the style of Shakespeare on the topic of democracy: ...*
 - Chatbot: *Alice: Hello. Bob: Hi, how are you? Alice: ...*

Exemple d'utilisation du « prompting »

- Challenge [DEFT 2023](#), concours d'interne en Pharmacie.
 - ▶ Environ 20% de bonnes réponses avec un [modèle de langage](#) entraîné juste à partir du web.

Corrigé du QCM de pharma.

Question 1 :

Lequel de ces cinq modes de désintégration radioactive se produit dans le noyau de tritium ?

- (a) Emission alpha.*
- (b) Emission bêta moins avec formation du noyau ^3_2He .*
- (c) Emission bêta plus avec formation du noyau ^3_2He .*
- (d) Capture électronique.*
- (e) Emission gamma.*

Réponse : (Δ)

Et chatGPT dans tout ça ?

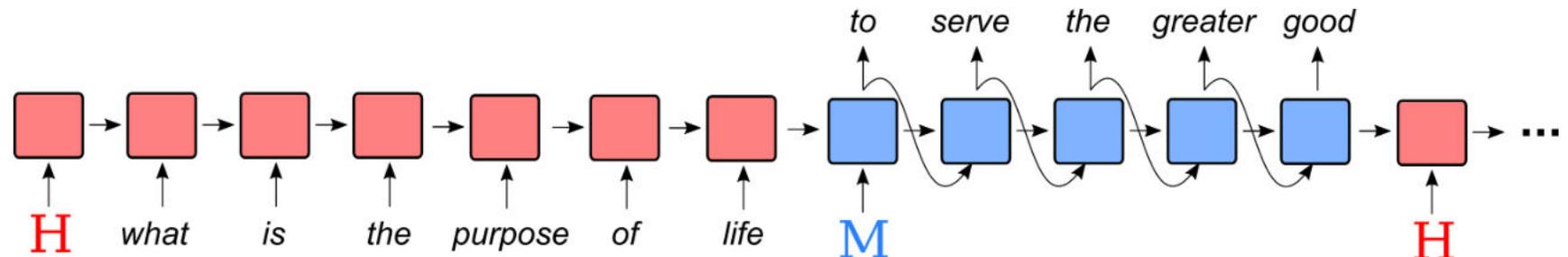
- Transition d'un modèle de prédiction à un modèle de « chat »
 - Entraîner un modèle de langage sur des conversations permet de prendre la place d'un des participants
(Vinyals et al. 2015; "[A Neural Conversational Model](#)")

Conversation 1: Basic

Human: *hello !*
Machine: *hello !*
Human: *how are you ?*
Machine: *i 'm good .*
Human: *what 's your name ?*
Machine: *i 'm julia .*
Human: *when were you born ?*
Machine: *july 20th .*
Human: *what year were you born ?*
Machine: *1977 .*
Human: *where are you from ?*

Conversation 4: Philosophical Q&A

Human: *what is the purpose of life ?*
Machine: *to serve the greater good .*
Human: *what is the purpose of living ?*
Machine: *to live forever .*
Human: *what is the purpose of existence ?*
Machine: *to find out what happens when we get to the planet earth .*
Human: *where are you now ?*
Machine: *i 'm in the middle of nowhere .*



Et chatGPT dans tout ça ?

- Problème du « contrôle » des sorties d'un modèle
 - Les textes d'Internet contiennent tout et n'importe quoi
 - des textes encyclopédiques (Wikipédia...), des articles scientifiques (Pubmed, Arxiv)
 - des documentations techniques, de l'aide technique (stack overflow, quora, wikihow...)
 - des articles d'actualité
 - des discussions (reddit, twitter...)
 - des sites commerciaux
 - des sites personnels
 - des textes générés
 - Problème : comment éviter les dérapages ?
 - Biais naturel des données (langues, cultures, 42% de la pop. mondiale n'utilisant pas internet)
 - Contenu inapproprié (violence, sexe, racisme, extrémisme...)
 - Contenu incorrect (rumeurs, désinformation, erreurs...)
 - **Alignement** avec un ensemble de valeurs
 - Reprendre le contrôle sur ce qui peut sortir (intéresse les industriels)

Alignement : Apprentissage avec des humains dans la boucle

- On cherche à créer un programme capable de valuer la qualité d'un texte généré
 - ▶ Collecter des textes générés pour un prompt donné
 - ▶ Demander à un humain de classer les textes du meilleur au plus mauvais
 - ▶ Modifier le modèle de langage pour qu'il tende à respecter ce classement

Table 3: Labeler-collected metadata on the API distribution.

Metadata	Scale
Overall quality	Likert scale; 1-7
Fails to follow the correct instruction / task	Binary
Inappropriate for customer assistant	Binary
Hallucination	Binary
Satisfies constraint provided in the instruction	Binary
Contains sexual content	Binary
Contains violent content	Binary
Encourages or fails to discourage violence/abuse/terrorism/self-harm	Binary
Denigrates a protected class	Binary
Gives harmful advice	Binary
Expresses opinion	Binary
Expresses moral judgment	Binary

- Promouvoir des productions du modèle de langage qui résolvent des tâches ou respectent des valeurs
 - ▶ Reinforcement Learning with Human Feedback (RLHF; Ouyang et al, 2022; "[Training language models to follow instructions with human feedback](#)")

ChatGPT = GPT3 + « prompt tuning » + alignement + ?

- Ouverture d'un service gratuit
 - Sortie en novembre 2022 par OpenAI
 - <https://chat.openai.com>
 - Succès fulgurant !!
 - +100 millions d'utilisateurs actifs en janvier 2023
 - Mentionné quotidiennement dans les médias
- Et après ?
 - Ajout de langages formels
 - Code source dans de nombreux langages informatiques
 - Formules mathématiques
 - Feedback à partir d'utilisations actuelles
 - « Red-teaming » : exploration des sujets sensibles
- OpenAI garde le secret du fonctionnement exact
 - Qualité pas encore égalee dans l'industrie



Qu'est-ce qu'on peut faire avec ?

- Capacités

- ▶ Générer (histoires, boîte à idées, emails types...)
- ▶ Reformuler (réécrire, traduire, résumer)
- ▶ Répondre à questions de culture générale pre 2021
- ▶ Résoudre des problèmes simples (pas très bon en calcul)
- ▶ Expliquer des concepts

- Incapacités

- ▶ N'a pas d'avis, ne parle pas de lui-même, de l'avenir...
- ▶ Ne génère pas de contenu inapproprié (mais on peut lui faire jouer un personnage qui le fait)

- Dialogue comme modalité d'interaction

- ▶ Négociation sur les sorties attendues
- ▶ Résolution collaborative avec l'utilisateur
- ▶ Construction itérative d'une solution (i.e. programme informatique)

Limites

- Longueur de l'historique (2k à 30k mots)
 - ▶ Ne se souvient pas au delà de l'historique
 - ▶ Impossible d'enseigner de manière durable quelque chose au modèle
- Hallucinations
 - ▶ Inventions, erreurs non factuelles
 - ▶ Impossibilité de reconnaître ce que le modèle de langage ne sait pas
- Événements récents
 - ▶ Modèles de langage créés à partir de données d'avant 2021
 - ▶ Vulnérabilité au feedback
- Biais
 - ▶ Langues peu représentées sur internet
 - ▶ Contenu dominé par la culture anglo-saxonne

Recherches en cours : intégrer d'autres types d'entrées

- Web / base de connaissances (WebGPT, Sidney, Bard)
 - ▶ Générer des requêtes
 - ▶ Retourner les résultats
- Appropriation d'outils (LangChain)
 - ▶ Calculatrice
 - ▶ Interpréteur python
 - ▶ Commandes pour manipuler des fichiers
- Images
 - ▶ Découper les images en morceaux et projeter ces morceaux dans l'espace des mots (Flamingo, Kosmos-1)
- Interaction avec le monde réel
 - ▶ Comment interfacer un modèle de langage avec le monde réel (robot PALM-e)
 - ▶ Recherches très actives dans ce domaine

+ New Chat



Today

Questions pour plus



July

Modèles de diffusion

May

Activités à Turin

April

Singular verb agreement.

March

Syntaxe d'une phrase.

Spiderman and Asterix Adve

Football in Marseille

English: Formal Language

Upgrade to Plus

FR

frederic.bechet@lis-lab



Default (GPT-3.5)

FR

Avez vous des questions ?



Bien sûr, j'ai plusieurs questions que je pourrais poser. Cependant, j'aurais besoin de plus de contexte pour vous poser des questions pertinentes. Pouvez-vous me donner plus d'informations sur le sujet ou le domaine que vous aimeriez explorer ? Cela me permettrait de poser des questions plus ciblées.



Regenerate

Send a message

