[Dossier IA générative] - ChatGPT : un beau parleur bien entraîné

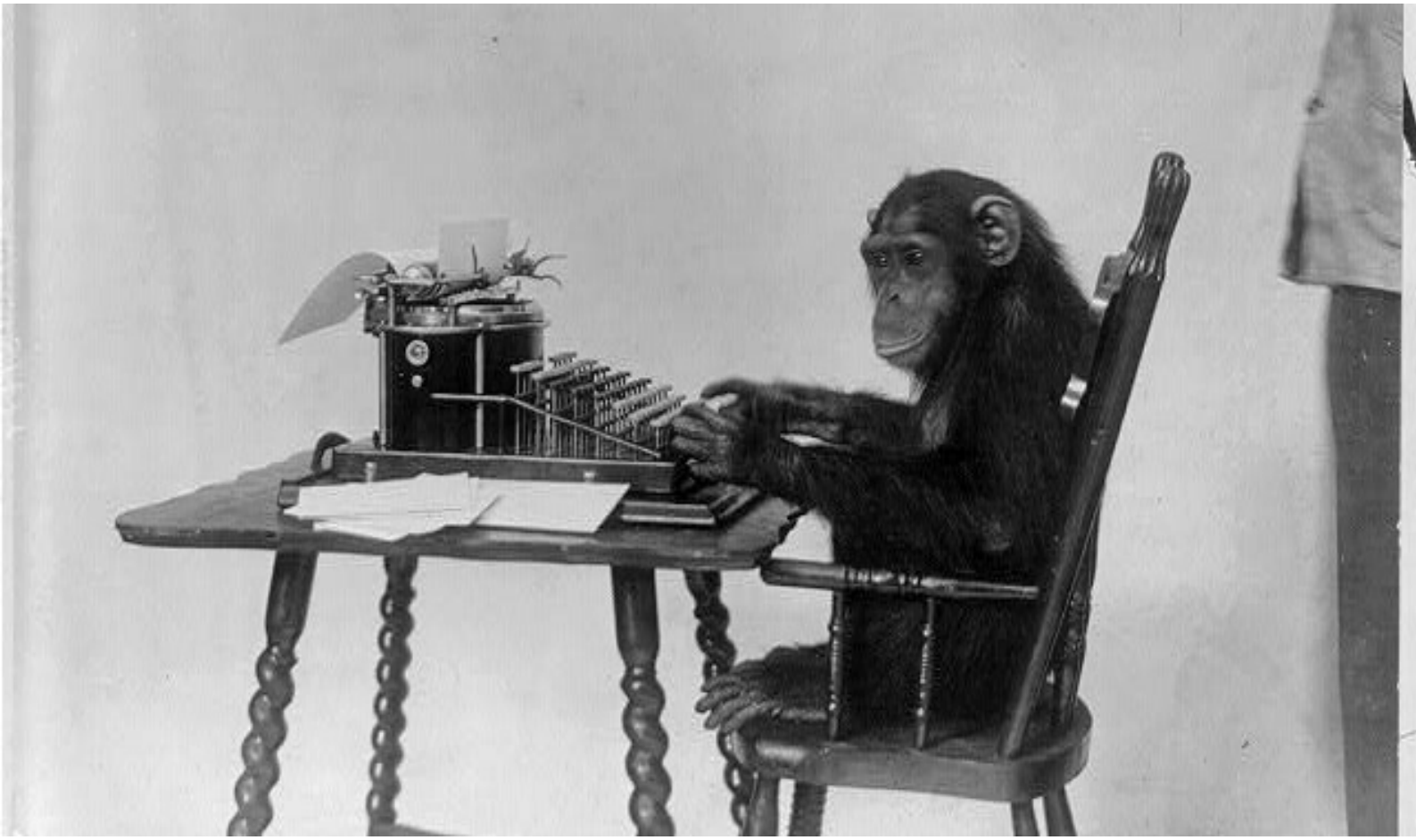
Rédigé par Alexis Léautier

-

26 avril 2023

ChatGPT, Bard ou encore Ernie, les chatbots alimentés à l'IA sont présents dans tous les journaux qui en vantent les prouesses ou en soulignent les enjeux. Les grandes entreprises de la Tech, à l'origine de ces créations, semblent se lancer dans une course à la médiatisation et aux parts d'un marché qui était encore inexistant en fin d'année 2022. Cette bataille médiatique en IA n'est bien sûr pas la première, les entreprises de la tech joutent régulièrement dans des domaines tels que les véhicules autonomes, la génération de contenus, l'analyse d'images ou encore les assistants vocaux.

Quels sont les risques techniques et juridiques liés à ces technologies et aux nouveaux usages auxquels elles ouvrent la voie ? Cet article, premier d'un dossier sur les IA génératives, revient sur les méthodes de traitement du langage et d'apprentissage utilisées pour le fonctionnement d'un chatbot en prenant l'exemple de ChatGPT.



Sommaire :

- ChatGPT : un beau parleur bien entraîné (1/4)

- **Quelle régulation pour la conception des IA génératives ? (2/4)**

- De l'entraînement à la pratique : l'IA générative et ses usages (3/4)

- **[Exploration LINC] - Les travaux d'IAsterix : les systèmes d'IA mis à l'épreuve (4/4)**

Si les textes produits par cette nouvelle génération de robots sont troublants par leur vraisemblance, les techniques qu'ils emploient n'ont rien d'inédit : le traitement du langage est loin d'être une discipline nouvelle.

Dans ce domaine, l'objectif est d'abord de parvenir à une compréhension grammaticale (sujet/verbe/complément) mais aussi sémantique des phrases (par exemple, comprendre que « auto », « voiture » et « bagnole » désignent le même objet d'une part, dans des registres de langage différents d'autre part). Parmi les autres défis connus figurent la distinction des homonymes (« avocat », « sous », « Paris »), le dialogue (qui suppose d'être capable de conserver une mémoire de la conversation) ou la détection de l'ironie.

Dans les années 1950, des algorithmes de traduction reposant sur des modèles symboliques (qui reposent sur des règles préétablies plutôt que sur un apprentissage) ont vu le jour avant d'être peu à peu remplacés par des algorithmes issus d'approches statistiques à partir des années 1980. En effet grâce à l'augmentation des capacités de calcul des ordinateurs et de serveurs, il est devenu possible de manipuler de grandes quantités de texte et d'en tirer des analyses statistiques telles que la probabilité d'occurrence du mot suivant dans une phrase. Enfin, l'arrivée du Web dans les années 2000 ouvrit la voie aux approches non-supervisées ou semi-supervisées en palliant la difficulté d'extraire des informations sémantiques des textes d'entraînement par la très grande quantité d'information maintenant disponible. Un historique plus complet des techniques, françaises notamment, de traitement automatique du langage peut être trouvé dans **le livre blanc de la CNIL sur les assistants vocaux**.

Toutefois, depuis les années 2010, différentes tentatives peu fructueuses ont eu lieu avant d'atteindre des résultats satisfaisants lorsqu'il s'agit d'effectuer des tâches diversifiées. En effet, les données étaient disponibles et les capacités de calcul semblaient suffisantes, mais l'algorithme qui permettrait d'exploiter ces ressources manquait encore. C'est en 2017 qu'une équipe de Google Brain introduisit les modèles de type Transformers dans **Vaswani et al., 2017** (dont le titre, « *Attention is all you need* », possède son importance et sur lequel nous reviendrons plus bas). La capacité de ces modèles à résoudre **le problème de la disparition du gradient** (voir plus bas) d'une part, et la possibilité de paralléliser leur apprentissage d'autre part, en font aujourd'hui les algorithmes de référence pour le traitement automatique du langage, à la suite des GRU, LSTM et autres RNN qui semblent avoir atteint leurs limites[1].

Cet article, qui se concentrera sur le fonctionnement et les enjeux techniques de l'outil ChatGPT développé par l'entreprise OpenAI, est le premier d'un dossier dont l'objectif est d'explorer les enjeux soulevés par les algorithmes d'IA génératifs comme GPT-3 et GPT-4 (génération de texte), Stable Diffusion (génération d'images), ou MusicLM (génération de sons) et par leur utilisation. Pour mieux cerner les capacités de ChatGPT, certains mécanismes du fonctionnement technique de cet outil seront expliqués, avant de répertorier les traitements de données personnelles ayant lieu lors de sa conception, et de son utilisation. Il s'agira ensuite de s'interroger sur les enjeux juridiques et éthiques de tels systèmes dans le reste du dossier.

ChatGPT : recette étape par étape

Le chatbot ChatGPT est une interface qui effectue en réalité plusieurs tâches. Il traite les informations renseignées dans l'invite, ou prompt en anglais, par l'utilisateur, alimente un algorithme de traitement du langage (appelé « Transformeur ») de ces données et présente la sortie de l'algorithme sous un format compréhensible pour l'utilisateur. L'algorithme de traitement du langage est donc l'ingrédient principal du fonctionnement de l'outil, mais nous verrons que la recette pour obtenir un chatbot est plus complexe.

Faut-il comprendre une question pour y répondre ?

Du langage humain au langage machine

Le texte inséré dans l'invite d'un chatbot doit tout d'abord être encodé dans un format lisible par une machine. Lors de cette étape appelée « vectorisation », une nouvelle représentation est associée à chacun des mots du corpus d'entraînement. Celle-ci se présente sous la forme d'un vecteur ou plus simplement d'une liste de valeurs numériques. L'ensemble de ces représentations, qui correspond d'une certaine manière à un nouveau langage interprétable par la machine, est appelé « l'espace latent ». Le site **Embedding Projector** offre une visualisation d'un tel espace latent (d'autres ressources sur le sujet sont à trouver **sur le site de la CNIL**).

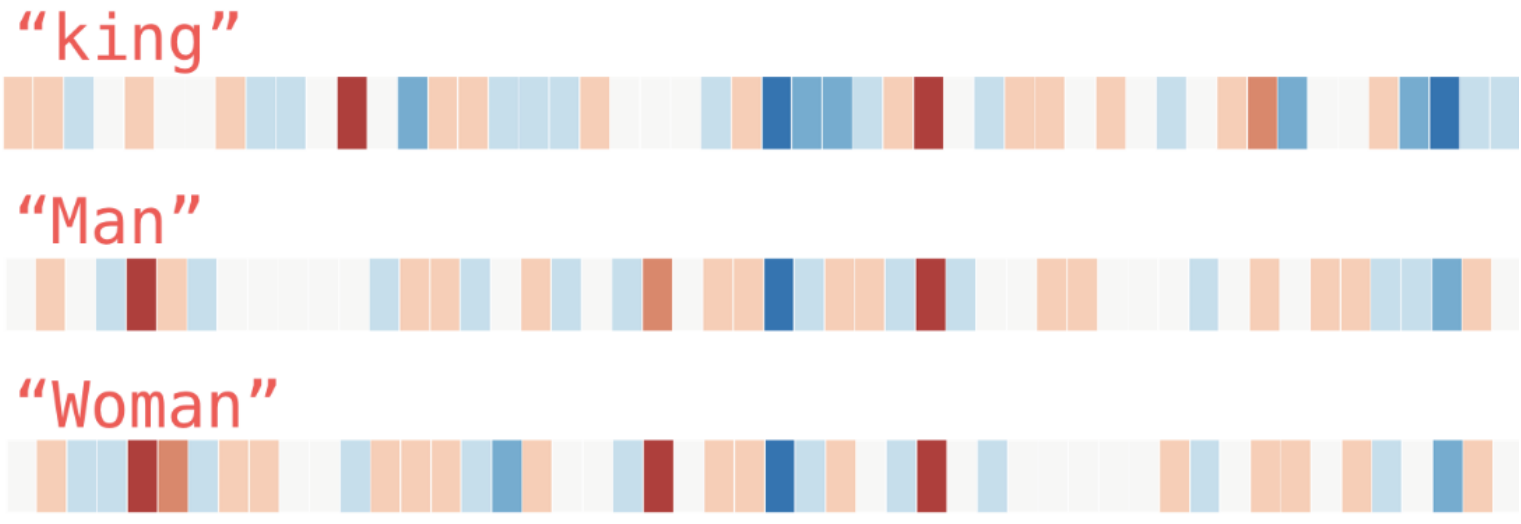


*Figure 1 : Représentation vectorielle des mots « king » ou roi, « Man » ou Homme et « Woman » ou Femme, où chaque case correspond à une entrée du vecteur. Une couleur chaude représente une valeur élevée, et une couleur froide, une valeur faible (source : **Jay Alammar, Illustrated Word2Vec**)*

Instinctivement, **ces vecteurs représentent le sens d'un mot décliné sous différents axes d'analyse** correspondant aux différentes valeurs numériques du vecteur. Toutefois, ces axes d'analyse ne sont jamais définis, c'est-à-dire que des mots tels que « homme », et « femme » peuvent avoir des valeurs proches en 5ème position de leur représentation comme c'est le cas dans l'exemple en figure 1. Pourtant, il n'a jamais été indiqué que la 5ème position correspondait à une notion comme « humain », « personne », « genre » ou « individu » : cela a été induit par l'algorithme de vectorisation grâce au contexte dans lequel les mots sont retrouvés dans le corpus d'entraînement.

Ainsi pour comprendre la représentation d'un mot, il faut la comparer avec celle d'autres mots, selon l'idée que deux mots sémantiquement proches auront des valeurs numériques proches pour certaines composantes des listes qui leur correspondent. On voit naître ici une première problématique : toute association de mots dans le corpus d'entraînement sera interprétée comme une proximité sémantique lors de la vectorisation. La figure 2 compare les représentations des mots « homme », « femme », « roi » et « reine ». En particulier, elle fait l'analogie entre la représentation du calcul « roi - homme + femme » et celle de « reine » (puisqu'on peut réaliser des opérations mathématiques sur les vecteurs). Si les vecteurs obtenus sont très proches dans l'ensemble, il reste certaines différences. **Celles-ci pourraient-elles être dues à un biais, ou en être la cause si on utilisait cette représentation ?**

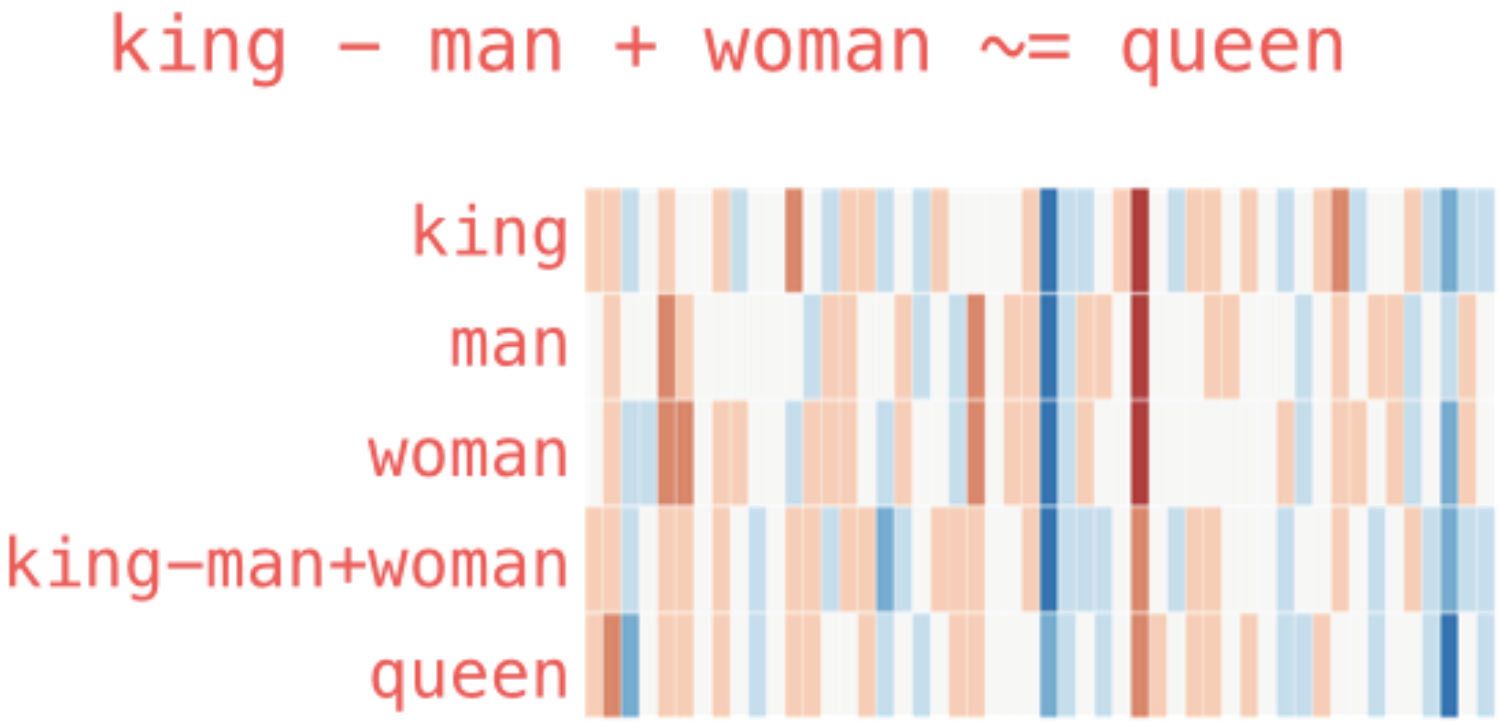


*Figure 2 : illustration de la représentation vectorielle des mots king (roi), man (homme), woman (femme), queen (reine) et de l'opération « king-man+woman ». Des couleurs similaires représentent des valeurs similaires dans la représentation (source : **Jay Alammar, Illustrated Word2Vec**)*

Les biais en IA : l'ombre qui plane

L'intelligence artificielle appliquée au traitement du langage pose le risque, comme tout système d'IA, de présenter des biais. Ces biais, de différentes natures, peuvent entraîner des discriminations comme rappelé dans **le rapport sur les enjeux éthiques des algorithmes et de l'IA publié par la CNIL**, ainsi que dans le rapport « **Algorithmes : prévenir l'automatisation des discriminations** » publié avec le Défenseur des Droits. Si ces biais ne peuvent être complètement évités en pratique, l'enjeu est alors de diminuer leurs conséquences pour les personnes, notamment en rendant les systèmes d'IA explicables, ce pourquoi plaide le professeur de mathématiques Philippe Besse **dans son entretien avec le LINC**.

Dans le cas du traitement de texte, l'enjeu des biais est particulièrement important car le langage possède certaines caractéristiques propres :

- Il possède des ambiguïtés complexes (l'utilisation du mot « nègre » par exemple est encore courante pour désigner un prête-plume) ;
- Il est imprégné des biais de la société (certains mots sont plus utilisés au féminin qu'au masculin par exemple) ;
- Il peut être très différent d'une communauté à l'autre, ou d'un milieu social à l'autre ;
- Il évolue au cours du temps.

Il est ainsi garanti qu'un modèle de langage possèdera des biais puisqu'il reproduira ce qui lui a été enseigné, même si certaines précautions peuvent être prises (cf. le paramétrage décrit plus bas).

Plus d'informations parmi **les ressources utiles sur l'IA** recensées sur le site de la CNIL.

De plus, le choix de **la longueur des vecteurs** représentant les mots aura aussi son importance. Le mot « roi » par exemple correspond au titre d'une personne de sexe masculin dirigeant une société qui reconnait la royauté, ce qui était courant au moyen-âge en particulier en France. On peut imaginer que sa représentation intègre toutes ces proximités sémantiques si l'algorithme était entraîné sur des manuels d'histoire occidentale. En revanche, en entraînant le même algorithme sur des manuels d'histoire asiatique, où l'histoire des Rois et Reines de France sera certainement présentée différemment, on obtiendra certainement des nuances différentes, comme « étranger » ou « occident » à intégrer dans la représentation vectorielle. Dans un autre cas où l'histoire des Rois et Reines de France serait simplement absente du corpus d'entraînement, le mot « roi » pourrait être lié à des notions assez aléatoires pour le modèle. En l'interrogeant sur le sujet, on risquerait alors d'obtenir des informations inexactes, fréquemment appelées « hallucinations », qu'il est difficile de corriger en aval (à l'instar de **Cobus Greyling** qui a tenté de rendre ChatGPT « poli » en lui indiquant de répondre « Sorry, I don't know » plutôt que d' « halluciner »). Dans le cas des Transformeurs, comme celui utilisé par ChatGPT, la longueur des vecteurs est de 512 valeurs. Si un vecteur plus long pourrait permettre de contenir plus d'information, il nécessite également d'augmenter le nombre de paramètres du modèle et augmente ainsi le coût de l'apprentissage. Une seconde problématique émerge : **comment trouver la taille de vecteurs réalisant le meilleur compromis entre l'atteinte du niveau de nuance sémantique voulu et l'augmentation du temps de calcul ?**

Enfin, avant d'utiliser les représentations vectorielles des mots pour traiter un texte, une dernière opération a lieu afin de tenir compte de la position du mot dans la phrase. La représentation vectorielle finale du mot tient ainsi compte du sens du mot, mais également de sa position.

Attention au contexte

La vectorisation est une première étape permettant à la machine d'interpréter les mots isolés, avec un certain niveau de subjectivité, comme nous l'avons vu. La seconde étape consiste en l'interprétation du contexte pour chacun des mots : **on parle de mécanisme de l'attention**.

Pour réaliser un lien entre les mots d'un même texte, et ainsi construire un contexte donnant une compréhension générale, les représentations vectorielles de chacun des mots du texte sont comparées entre elles. Par une opération vectorielle classique (un produit scalaire), on attribue à chacune des paires de mots dans le texte un nombre représentant leur proximité dans l'espace vectoriel. Intuitivement, cela devrait revenir à comparer la proximité sémantique entre chacun des mots du texte, ce qui permet d'établir des liens entre eux. Ce calcul est réalisé pour un même mot avec tous les mots l'entourant : on obtient une nouvelle représentation du mot intégrant son lien avec le contexte[2].

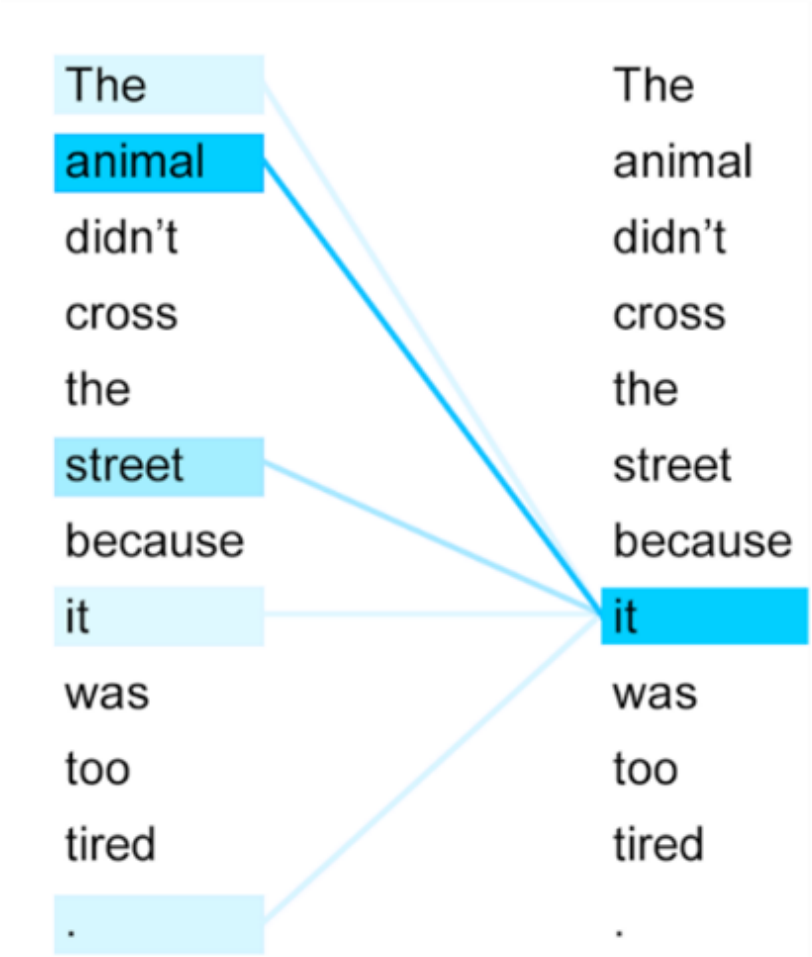


*Figure 3 : Proximité entre le mot « it » et son contexte pour la phrase « L'animal n'a pas traversé la rue car il était trop fatigué ». Une couleur intense représente une grande proximité entre deux mots. (source : **Uszkoreit, 2017**)*

Dans l'exemple en figure 3, les valeurs indiquant la proximité sémantique avec le mot « it » sont représentées par des couleurs plus intenses quand la proximité est grande (note : toutes les valeurs ne sont pas représentées, certaines étant trop faibles pour être pertinentes).

En pratique, cette étape visant à utiliser le contexte est répétée plusieurs fois, liant ainsi les mots un à un, grâce à des connexions croisées. Ces itérations permettent de distinguer des phrases similaires dont le sens n'est toutefois pas le même, comme le montre l'exemple donné en figure 4. Dans ces deux exemples, seul le lien entre « animal » et « tired » dans le premier cas, et entre « street » et « wide » dans le second permet de savoir si « it » se rapporte à « animal » ou à « street ». Les sens de « animal » et de « street » doivent donc d'abord être déterminés grâce au contexte avant de déterminer le sens de « it ».

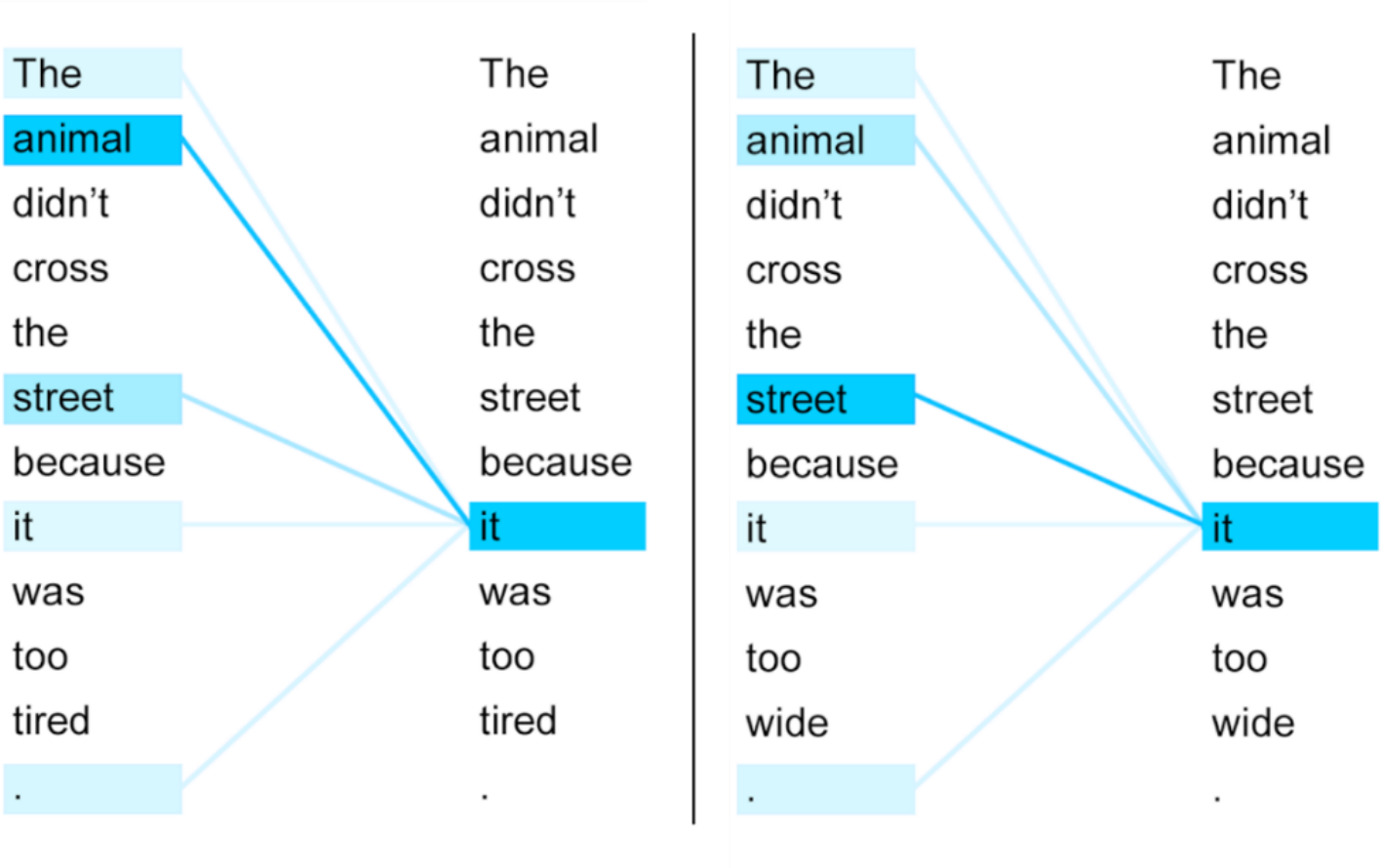


*Figure 4 : Proximité entre le mot « it » et son contexte pour deux phrases similaires. Une couleur intense représente une grande proximité entre deux mots. (source : **Uszkoreit, 2017**)*

Cette capacité des Transformeurs à utiliser l'ensemble du contexte pour déterminer le sens d'un mot est l'un des avantages majeurs de cette catégorie de réseaux. Les autres types de réseaux d'apprentissage profond, comme les réseaux de neurones récurrents (RNN), sont généralement moins performants pour cette tâche en raison d'un problème appelé la disparition du gradient.

La disparition du gradient

Lors de l'apprentissage, les paramètres d'un réseau de neurone sont corrigés en fonction de l'erreur de prédiction sur les données d'entraînement, cette erreur est quantifiée par le gradient. On parle ainsi de descente du gradient lorsqu'on tente de réduire sa valeur, c'est-à-dire de réduire l'erreur.

Dans le cas d'un réseau de neurone récurrent, le gradient prend en compte la contribution d'une séquence de données, comme une phrase entière. Cependant, tous les mots dans la phrase n'ont pas le même poids dans le calcul du gradient : plus ces derniers remontent vers le début de la phrase, plus leur poids décroit (de manière exponentielle). Cette décroissance étant exponentielle, les premiers mots d'une phrase n'ont généralement plus aucun impact dans le calcul du gradient, et donc ils ne sont quasiment pas utilisés pour corriger les paramètres du modèle. C'est ce qu'on appelle la disparition du gradient : en remontant dans une séquence, la contribution des éléments diminue de manière exponentielle et leur contribution à l'apprentissage « disparaît ».

De l’apprentissage du langage au chatbot

Encoder, décoder

Le procédé vu dans la partie précédente permet de représenter chacun des mots dans l’espace latent. Cette première étape, qui correspond à la partie de l’algorithme appelée « encoder », ne permet pas de générer du texte. C’est ici que la seconde partie, appelée « decoder » intervient.

Selon la fonctionnalité prévue pour l’outil, le decoder intervient pour la mettre en œuvre. En utilisant la représentation des mots d’une phrase qu’on souhaite traduire ou compléter par exemple, le decoder vient générer des mots selon la probabilité de leur occurrence. Le calcul de cette probabilité repose sur les méthodes vues plus haut et tient ainsi compte de la proximité sémantique du mot avec son contexte, tout comme de sa position dans la phrase.

Dans le cas de ChatGPT, ce qui est entré par l’utilisateur dans l’invite est utilisée comme une donnée d’entrée et encodée de façon à créer un contexte initial. Sur cette base, l’algorithme Transformeur sur lequel repose le chatbot parvient à générer du texte en sélectionnant un à un les mots les plus probables pour construire une réponse. Cet ensemble « encoder-decoder » constitue le modèle algorithmique, nommé GPT^3 dans le cas de ChatGPT, sur lequel repose le « chat ». Néanmoins, la génération de texte mot à mot pourrait difficilement donner les performances observées de ChatGPT. Une étape d’adaptation du modèle semble nécessaire pour que le « chat » produise des réponses claires, ordonnées et respectant les consignes de l’utilisateur. C’est ici que le paramétrage, ou alignement, du modèle entre en jeu.

Woupidou, quand ChatGPT voudrait parler comme vous

Les concepteurs de ChatGPT ont décidé de favoriser les attentes des utilisateurs plutôt que des performances théoriques lors d’une phase de paramétrage, ou fine-tuning, du modèle par un apprentissage partiellement supervisé.

D’après les auteurs de **Ouyang et al., 2022**, la publication d’OpenAI où cette étape est détaillée, l’objectif de cette démarche était de mieux correspondre aux attentes explicites (le suivi d’instructions données par exemple), et implicites (l’absence de biais ou de contenus toxiques, la vraisemblance des réponses, etc.). Pour cela deux techniques ont été utilisées :

> Le paramétrage par apprentissage supervisé : des réponses types sont rédigées par des annotateurs pour servir de données d’entraînement au modèle de langage. Le paramétrage est réalisé tout d’abord selon une métrique d’utilité des réponses (*helpfulness*) lors de l’entraînement puis selon des métriques de véracité (*truthfulness*) et d’innocuité (*harmlessness*) lors de la validation, puisque ces métriques peuvent entrer en conflit (si un utilisateur demande de générer du contenu toxique par exemple, l’innocuité est préférable à l’utilité de la réponse).
>
> Le paramétrage par apprentissage par renforcement : lors de cette phase, un modèle de récompense est entraîné de manière supervisée à reconnaître ce qui est une « bonne » réponse. Ce modèle est ensuite utilisé pour guider le paramétrage du chat : le chat génère une réponse, le modèle de récompense détermine si celle-ci est conforme aux attentes et fournit un haut score le cas échéant, finalement utilisé pour guider le paramétrage du modèle de langage.

Les auteurs ont jugé qu’une collecte spécifique de données pour cette phase de paramétrage était nécessaire car les jeux de données publics reflètent mal les préférences réelles des utilisateurs. Il est notamment indiqué que l’API proposée par OpenAI est utilisée à 46% environ pour des demandes de « génération » (c’est-à-dire pour des demandes telles

que « raconte-moi une histoire … ») alors que les jeux de données publics contiennent majoritairement des exemples de classification, de réponse à des questions précises, de traduction ou de synthèse de texte.

Paroles (annotées), paroles (non-annotées), paroles (partiellement annotées)…

Comme vu précédemment, le développement et l'amélioration de ChatGPT repose sur plusieurs phases d'apprentissage, dont certaines sont supervisées (c'est-à-dire sur des données annotées), non-supervisées (c'est-à-dire sur des données non-annotées), et d'autres partiellement supervisées (où on tente d'utiliser certaines données non annotées ou partiellement annotées pour l'apprentissage). Pour ces étapes, des catégories et volumes de données différents sont traités.

Les données non annotées : le tout-venant

Tout d'abord, pour l'entraînement du modèle de langage sur lequel repose ChatGPT, un apprentissage non supervisé a lieu.

Cette technique permet de se libérer de l'annotation des données, étape la plus lourde et la plus coûteuse pour l'entraînement d'un modèle, mais nécessite un très grand volume de données en contrepartie. D'après **Brown et al., 2020**, plusieurs jeux de données textuels ont été utilisés **dont le CommonCrawl** (également utilisé par BigScience pour **le projet Bloom qui a fait l'objet d'une interview du LINC**), le jeu **WebText**, le contenu de Wikipédia et des œuvres littéraires issues du domaine public. Le CommonCrawl et WebText sont tous deux issus de moissonnage, ou scraping, du Web et contiennent ainsi des sources de données grand public telles que Reddit ou BlogSpot, mais également des sites institutionnels (Europarl.eu, nasa.gov) ou académiques (mit.edu, cornell.edu, berkeley.edu, cnrs.fr, etc.) et des sources de presse (euronews.fr, lefigaro.fr, ouest-france.fr, etc.).

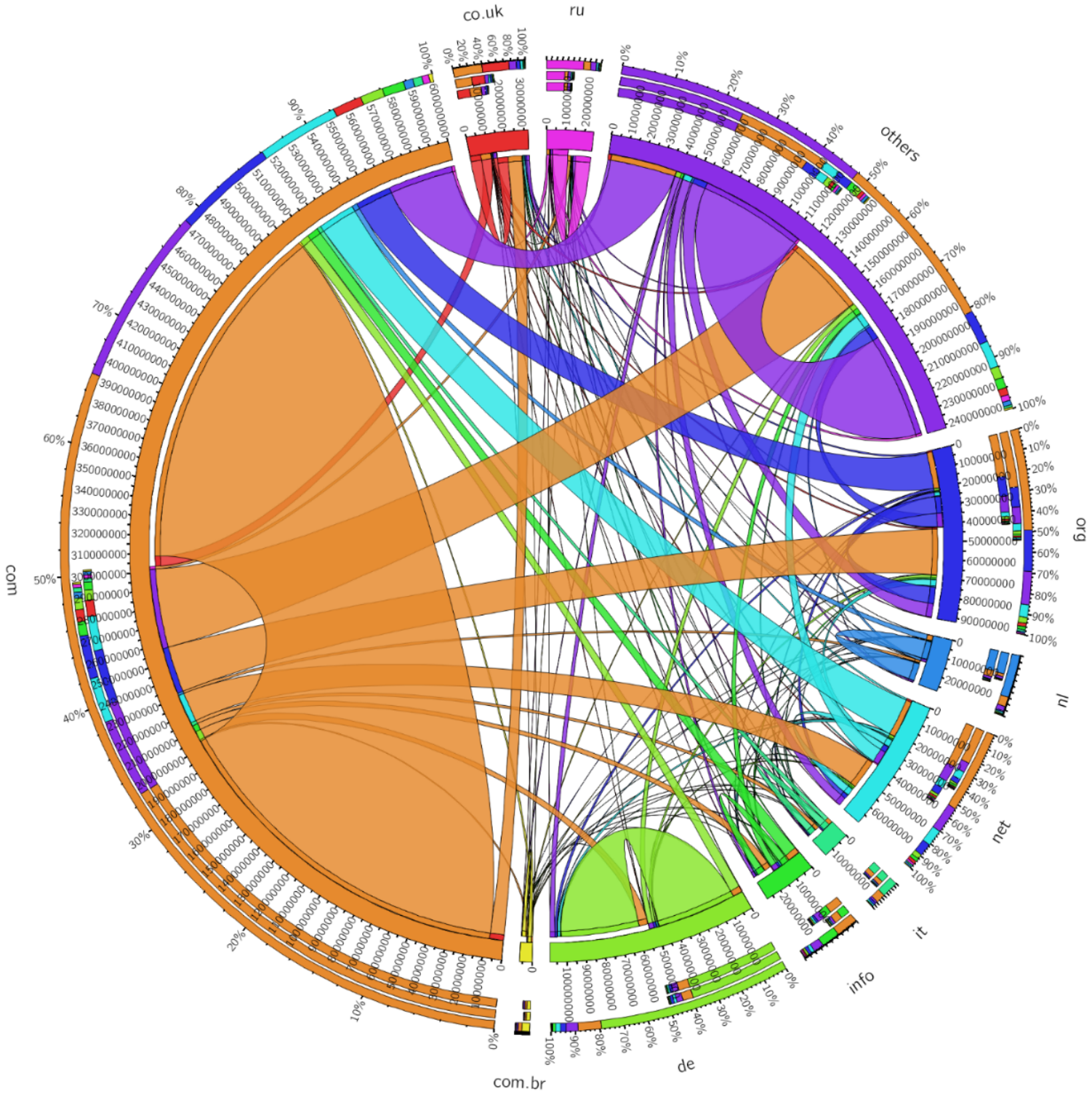


*Figure 5 : Représentation de la provenance des url moissonnées par le CommonCrawl (source : **WebDataCommons dans KDNuggets en 2015**)*

Comme on le voit, la représentativité des domaines visés par le CommonCrawl – qui représente la majorité des données pour l'entraînement non supervisé et dont une documentation détaille **le « Top-500 » des domaines visités lors du projet** – pose question. Si pour les sites anglophones, de nombreuses universités sont trouvées dans ce jeu de données (79 sites web appartenant au domaine « .edu » réservé aux établissements d'enseignement accrédités aux États-Unis), on ne compte qu'une seule université française (ens-lyon.fr). Des disparités semblent ainsi exister entre les pays, non seulement en ce qui concerne le volume de données, mais également en ce qui concerne les types de sources.

De plus, pour une même catégorie de source de données (comme les sources de presse), les domaines moissonnés pourraient manquer de représentativité. En France par exemple, les seuls domaines figurant dans le Top-500 mentionné plus haut et correspondant à des

sources de presses sont ouest-france.fr et lefigaro.fr. Avec seulement deux domaines, si le modèle venait à reproduire un point de vue suite à son apprentissage, celui-ci risquerait fortement de ne pas être représentatif de la société dans son ensemble.

Enfin, avec une collecte d'une telle ampleur, il semble difficile de juger de la fiabilité des sources utilisées sans une analyse dédiée pour chacune des langues. Des critères arbitraires peuvent parfois être utilisés, tel qu'une notation (le corpus WebText par exemple exclut les publications sur Reddit ayant reçu moins de 3 « karmas »[4], la notation utilisée sur le site). Il reste néanmoins à prouver que cette notation est garante d'une information de qualité puisque celle-ci pourrait être répétée par le chat.

Les données annotées : le véritable nouvel or noir

Une fois le modèle de langage entraîné, il est paramétré par un apprentissage supervisé, sur des données annotées.

Comme décrit dans **Ouyang et al., 2022**, une équipe de 40 annotateurs a été recrutée afin de rédiger des réponses qui serviront de modèle lors du paramétrage du modèle de langage. OpenAI souligne qu'une attention particulière a été portée au recrutement de ces personnes. Les annotateurs sélectionnés auraient ainsi démontré une sensibilité « aux préférences de divers groupes démographiques » et une grande capacité à identifier « des réponses potentiellement nocives ». Ainsi, environ 11000 exemples ont été utilisés lors de cette phase.

Les données partiellement annotées : le meilleur de deux mondes ?

Après le premier paramétrage par apprentissage supervisé, un second paramétrage a lieu par apprentissage par renforcement. Cette phase requiert l'entraînement (supervisé) d'un modèle de récompense qui servira de guide pour orienter l'apprentissage (non supervisé) du modèle utilisé par le chat.

Pour l'annotation des données servant à entraîner le modèle de récompense, l'API d'OpenAI a été modifiée pour permettre à ses utilisateurs de sélectionner la meilleure réponse du modèle à une question parmi plusieurs choix. Environ 30 mille réponses notées ont ainsi été obtenues et sont utilisées pour entraîner un modèle de récompense capable de comparer les générations du chat aux réponses les mieux notées et de le « récompenser » ou non lors de l'apprentissage.

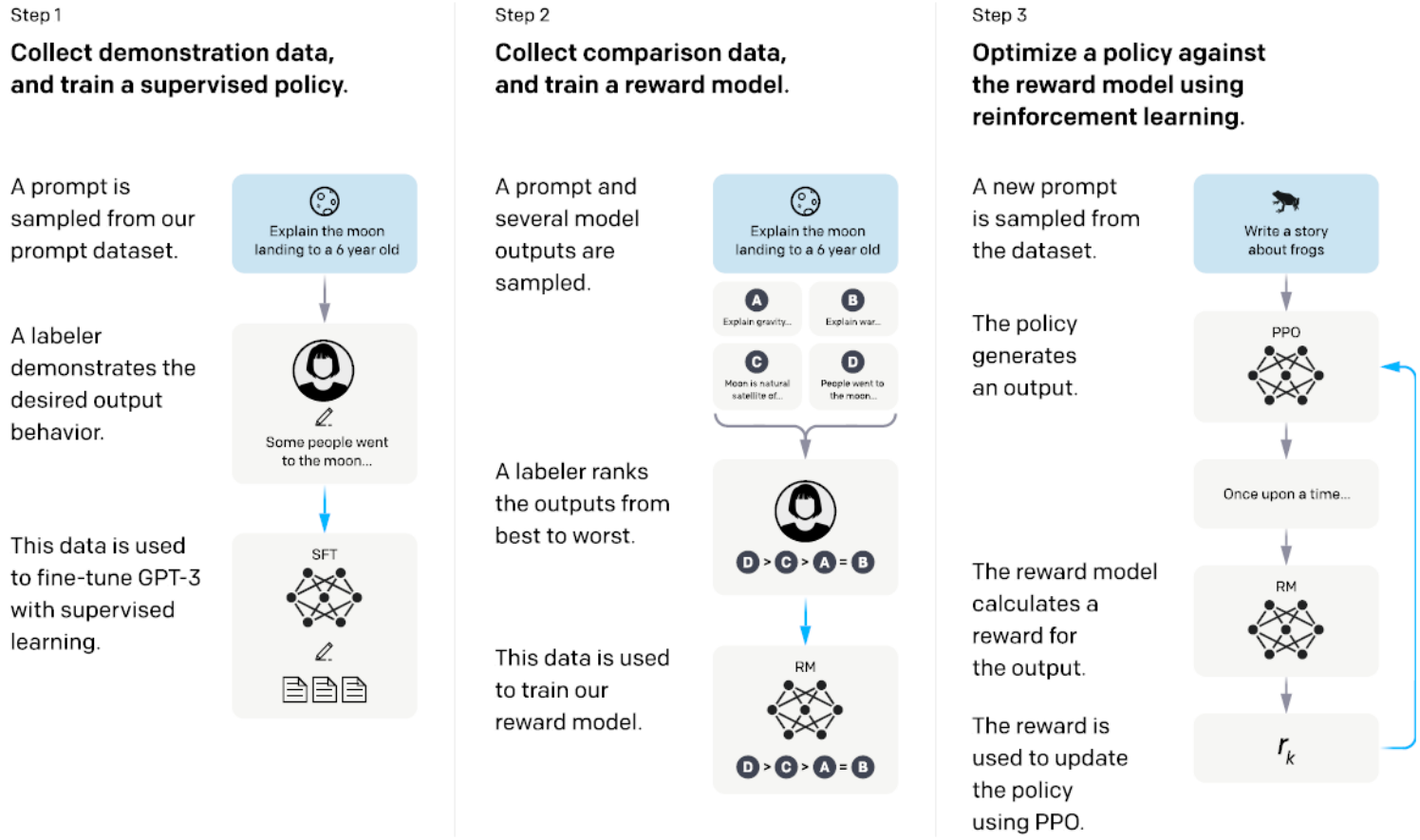


*Figure 5 : description des trois phases du paramétrage de ChatGPT. Ce schéma ne décrit pas la première phase de la conception, c'est-à-dire l'entrainement du modèle de langage (source : **Ouyang et al., 2022**)*

Pour finir, **comme indiqué dans l'aide en ligne de ChatGPT**, les conversations des usagers sont réutilisées pour améliorer les performances de l'outil, sans plus de précisions, le seul élément directement visible étant la possibilité de donner un « pouce » négatif ou positif à chaque réponse (👍👎). La **FAQ** du site indique que les conversations peuvent être revues par des annotateurs, il semble donc vraisemblable que ces données soient réutilisées pour le paramétrage de l'outil (qu'il s'agisse de l'apprentissage supervisé ou par renforcement).

A noter qu'un droit d'opposition à cette réutilisation est prévu.

ChatGPT2.0 ?

Dans **un billet de blog datant du 16 février 2023**, OpenAI revient sur les critiques les plus courantes concernant les performances du chat et les défauts observés. L'entreprise détaille également un plan d'action pour prendre en compte ces retours et améliorer ses systèmes. Certaines de ces modifications ont d'ailleurs été intégrée dans le modèle GPT4 d'après **une publication du 5 avril 2023**.

Premièrement, afin de corriger les biais du chat, reflet notamment des biais potentiel des annotateurs sur certains sujets politiques ou contreversés, OpenAI indique que la première étape consistera en l'amélioration des procédés de revue des réponses générées par les examinateurs. L'entreprise souhaite pour cela faire évoluer les lignes directrices suivies par les examinateurs, mais également améliorer la représentativité démographique de ces derniers. Dans une optique de transparence, elle indique par ailleurs envisager de publier davantage d'informations concernant leur répartition démographique.

De plus, l'entreprise cherche des pistes techniques permettant de rendre l'étape de paramétrage du modèle plus contrôlable et transparente. A ce sujet, **Glaese et al., 2022** et **Bai et al., 2022** proposent deux approches différentes. La première vise à fournir une grille d'analyse plus détaillée aux annotateurs afin de leur permettre de détailler pourquoi une réponse ne correspond pas aux attentes. La seconde approche repose davantage sur la technique : un modèle est entraîné à améliorer le texte généré par le chat selon un ensemble de règles, appelée « constitution ». Ainsi, dans cette seconde approche, un second modèle vient corriger les sorties du premier, qui apprend de ces erreurs dans une logique d'apprentissage par renforcement. Les réponses améliorées sont ensuite utilisées dans la phase d'apprentissage par renforcement.

Enfin, OpenAI donne des pistes de développement général pour le chat. Tout d'abord, l'entreprise indique que les retours des utilisateurs seront utilisés pour corriger les biais, les « hallucinations » et la capacité de l'algorithme à identifier les requêtes toxiques. Des outils permettant aux utilisateurs de paramétrer ChatGPT à leurs cas d'usage spécifiques seront également développés. Enfin, OpenAI indique souhaiter inclure davantage les utilisateurs dans le développement de l'outil pour corriger les défauts dus à une trop grande « concentration de pouvoir ».

Ces évolutions à venir témoignent de l'entrée du produit dans une phase de mise sur le marché où il s'agit non seulement de développer un outil performant, mais qui correspondrait également aux attentes des clients potentiels. Les usages de l'outil, qui sont encore émergents posent toutefois des questions d'ordre juridique et éthique, dont une analyse est à lire **dans l'article suivant du dossier**.

---

[1] Les réseaux de neurones récurrents – recurrent neural network ou RNN – sont une catégorie de modèles d'apprentissage automatique, dont les réseaux Gated Recurrent Unit ou GRU et Long Short-Term Memory ou LSTM sont les exemples les plus fréquents, capables d'utiliser le contexte d'une information comme la phrase dans laquelle est contenue un mot.

*[2] En pratique, l'opération de calcul de l'attention avec le contexte est réalisée plusieurs fois en parallèle puisqu'il peut y avoir plusieurs manières d'interpréter un même contexte : on parle d'attention multi-head. Pour chacune des « têtes », des paramètres sont entraînés et permettent de tirer des informations différentes du contexte. La somme des têtes donne ainsi une interprétation plus complète du contexte. Les différentes versions de GPT3 possèdent entre 12 et 96 de ces têtes d'après **Brown et al., 2020**.*

[3] Generative Pretrained Transformer – Transformeur génératif préentraîné.

*[4] Les karmas sont des récompenses attribuées par les utilisateurs de Reddit entre eux. Plus d'informations sur **le site de Reddit**.*

Illustration - [Chimpanzee seated at a typewriter] (Picryl - Library of Congress)

---

*Article rédigé par **Alexis Léautier**, Ingénieur Expert*

VOIR PLUS D'ARTICLES DE L'AUTEUR