Aleph Alpha, LightOn et Bloom, les alternatives européennes à ChatGPT

Description

Le traitement automatique des langues (*Natural Language Processing – NLP*) est un domaine à la croisée de la linguistique, de l'informatique et de l'intelligence artificielle portant sur la manipulation du langage naturel par les ordinateurs. Si le phénomène ChatGPT développé par l'entreprise américaine OpenAI fait grand bruit, il existe une offre européenne, avec notamment Aleph Alpha, LightOn ou encore Bloom.

Le traitement automatique des langues consiste à inventer des outils informatiques à partir de la langue naturelle pour des applications aussi variées que la traduction automatique, l'analyse syntaxique, l'analyse de discours, la synthèse vocale ou encore des applications de robots conversationnels. Parmi les différents modèles de traitement automatique des langues, les grands modèles de langage (*Large Language Models – LLM*) constituent une catégorie dont l'objet est de prédire le mot d'après dans une séquence de texte donnée en se basant sur le contexte. Ces logiciels s'entraînent à partir de grands corpus de données grâce à des « réseaux de neurones profonds » pour apprendre les modèles de langage. Un réseau de neurones profond désigne un ensemble d'algorithmes inspirés par le fonctionnement du cerveau humain, dont l'objet est de reconnaître des motifs et la transmission d'informations entre diverses couches de connexions neuronales : une couche d'entrée et une couche de sortie assorties d'au moins une couche intermédiaire. Chacune de ces couches correspond à des modèles mathématiques avancés qui effectuent différents types de tri et de catégorisation dans un processus nommé « hiérarchie de caractéristiques ». Plus le nombre de couches intermédiaires est élevé, plus le réseau est dit « profond ». Ces algorithmes d'apprentissage profond s'auto-entraînent et analysent le langage naturel à partir de quantités massives de données textuelles afin d'identifier, dans ce corpus, les relations entre entités pour générer conformément à celles-ci un nouveau texte cohérent et grammaticalement correct.

Cependant, précise Laurence Devillers, professeure en informatique à l'université Paris-Sorbonne et chercheuse au Limsi (Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur du CNRS, voir *La rem* n°46-47, p.107), *« même si la syntaxe est parfaite, il ne faut pas être leurré, ces systèmes ne raisonnent pas, ils n'ont pas de compréhension temporelle, pas de compréhension de la logique, pas d'induction, déduction, ce sont des probabilités de succession de mots basées sur l'analyse du contexte de leur utilisation »*. Cette analyse du contexte dépend donc essentiellement du corpus de textes fourni au système, qui apprend ainsi à corréler les mots entre eux, suivant un certain nombre de paramètres, dont la somme croît de manière exponentielle depuis cinq ans. Les paramètres de ces grands modèles de langage font référence aux valeurs que le système peut mettre au jour de manière indépendante au fur et à mesure qu'il « apprend », ce qui pose d'ailleurs le problème fondamental de l'explicabilité. En effet, comme le précise la Cnil (Commission nationale de l'informatique et des libertés), *« le nombre de paramètres utilisés fait qu'il est souvent impossible de comprendre d'où vient une erreur »*

. Autrement dit, plus le nombre de paramètres est élevé, moins il est possible de définir et de comprendre le résultat fourni par le système.

GPT (*Generative Pre-trained Transformer*) est un modèle de langage basé sur l'intelligence artificielle développé par l'entreprise américaine OpenAI depuis 2018, et dont la première version utilisait 117 millions de paramètres. GPT-2, lancé en 2019, utilisait 1,5 milliard de paramètres ; GPT-3 en 2020, 175 milliards et GPT-4 sorti en mars 2023, mais uniquement accessible par abonnement payant, s'appuierait sur un nombre de paramètres dépassant le trillion, sans que l'entreprise l'ait confirmé. Alors que GPT est le modèle de langage générique développé par OpenAI, ChatGPT en est l'adaptation spécifique destinée à la génération de réponses conversationnelles. Lancé en novembre 2022 dans une version gratuite et non connectée à internet, ChatGPT a atteint 180 millions d'utilisateurs en moins de six mois, faisant ainsi l'objet d'une popularité mondiale.

Pourtant, en Europe, d'autres grands modèles de langage existent. La *« guerre du contenu sémantique »*, pour reprendre la formule de Laurence Devillers, est avant tout culturelle. Le contenu sémantique fourni à ChatGPT est en anglais. Celui utilisé par Ernie Bot (*Enhanced Representation through Knowledge Integration* – Représentation linguistique améliorée au travers d'entités d'information), développé par le géant Baidu, est en chinois. Bloom, projet européen, porté notamment par la France, se veut quant à lui multilingue.

Lorsque l'on interroge ChatGPT à propos de ses « concurrents en Europe », voici ce que le système répond : *« En tant que modèle de langage générique de pointe, ChatGPT a de nombreux concurrents en Europe et dans le monde entier. Cependant, il est difficile de dresser une liste exhaustive de tous les concurrents potentiels car il y a une multitude de modèles de langage disponibles sur le marché. Néanmoins, certains des concurrents de ChatGPT en Europe peuvent inclure des modèles de langage tels que BERT (Bidirectional Encoder Representations from Transformers) développé par Google, XLNet développé par l'Université Carnegie Mellon, GPT-2 et GPT-3 développés par OpenAI, T5 (Text-to-Text Transfer Transformer) développé par Google, ainsi que des modèles de langage développés par Facebook tels que RoBERTa (Robustly Optimized BERT Pretraining Approach) et XLM-R (Cross-lingual Language Model for 100+ Languages). Cependant, il convient de noter que chaque modèle de langage a ses propres caractéristiques et domaines d'application, et que le choix d'un modèle dépendra des besoins spécifiques de l'utilisateur. »* Autrement dit, la machine américaine ne cite que des modèles américains.

Aucun mot donc sur Aleph Alpha en Allemagne ou encore LightOn et Bloom en France, ce qui démontre bien les biais de telles applications, dont les résultats ne dépendent que des contenus dont ils auront été préalablement nourris.

Aleph Alpha, installé depuis 2019 à Heidelberg en Allemagne, développe des modèles multimodaux originaux, dont Luminous, présenté à l'International Supercomputing Conference (ISC) de Hambourg en 2022. Le système combine la vision par ordinateur (capable d'analyser une image et d'en déduire des informations) avec le traitement du langage naturel (NLP) pour traiter, analyser et produire un large éventail

de textes. Ce modèle s'appuie sur la méthode MAGMA (*Multimodal Augmentation of Generative Models through Adapter-based Finetuning*). Aleph Alpha a noué un partenariat avec Graphcore, une société britannique de semi-conducteurs créée en 2016 qui développe des accélérateurs pour l'intelligence artificielle et le *machine learning*. Les deux entreprises travaillent à optimiser le rapport entre la puissance de calcul et l'empreinte énergétique nécessaire au traitement des données. Et également à faire en sorte que l'Europe avance ses pions dans le domaine de l'intelligence artificielle. Selon Jonas Andrulis, auparavant cadre chez Apple, fondateur et aujourd'hui président de l'entreprise, *« toute la diversité linguistique et culturelle de l'Europe doit se refléter dans les applications modernes de l'intelligence artificielle, car c'est le seul moyen pour chaque pays européen, grand ou petit, de bénéficier du potentiel des nouvelles technologies de l'intelligence artificielle. Cela garantit que le meilleur de l'intelligence artificielle n'est pas réservé à quelques-uns, mais est disponible pour tous de manière égale »*. Après avoir levé 23 millions d'euros en 2021, l'entreprise s'apprête à effectuer un deuxième tour de table de 100 millions d'euros, auquel l'allemand SAP, premier éditeur de logiciels en Europe et quatrième dans le monde, pourrait participer. Au côté de Bloom, Aleph Alpha est considéré comme l'un des plus grands espoirs pour une intelligence artificielle européenne indépendante.

Start-up française créée par quatre chercheurs en 2016, LightOn a lancé une offre de grands modèles de langage en 2020, appelée Paradigm, à destination des grandes entreprises européennes. À la différence de ChatGPT, dont les données sont traitées à distance dans le cloud, Paradigm est déployé sur les infrastructures du client afin de garantir la confidentialité des données. Et les clients sont au rendez-vous. *« Les demandes sont diverses, car les entreprises cherchent des gains de productivité sur des fonctions variées comme le marketing, les ressources humaines, les ventes et même la R&D. Par exemple, les larges modèles de langue sont plus subtils pour appréhender le contexte d'une interaction qu'un système marketing traditionnel. Cela fonctionne comme un chatbot avec du contexte et ces systèmes, par exemple sur un service client, sont capables de classifier les demandes des usagers en fonction des interactions précédentes avec eux et, donc, de disposer d'un contexte »*, explique Laurent Daudet, le président de l'entreprise. Paradigm serait ainsi capable de rédiger des fiches-produits en fonction de l'audience, de trouver des slogans marketing, de faire de la veille informationnelle et de la synthèse de documents ou même de retranscrire le contenu de réunions.

Lancé l'été 2021, Bloom, acronyme de *BigScience Large Open-science Open-access Multilingual Language* , est un projet de science ouverte et participative piloté par Hugging Face. Cette start-up, fondée à New York par trois Français, propose une bibliothèque de traitement automatique des langues open source permettant d'accéder à plusieurs modèles pré-entraînés. Bloom est probablement l'initiative européenne la plus ambitieuse en matière d'intelligence artificielle. Elle est l'œuvre d'un millier de chercheurs issus de 72 pays, avec le soutien du Genci (Grand Équipement national de calcul intensif, voir *La rem* n°52, p.31), du CNRS, du ministère de l'enseignement supérieur et de la recherche français ainsi que de partenaires privés tels que Airbus, Mozilla, Orange Labs ou Thales. L'objectif de Bloom est d'entraîner le plus grand modèle de langue multilingue et open source. S'appuyant sur un modèle à 70 couches de neurones et 176 milliards de paramètres, il a été entraîné durant l'équivalent de 5 millions d'heures de calcul sur un corpus bien plus

riche, comparé à ChatGPT-3, soit 46 langues, du français au basque en passant par le mandarin et 20 langues africaines, ainsi que sur 13 langages de programmation, le tout représentant 1,6 téraoctet de données.

Grâce au CNRS et au Genci, Bloom a bénéficié d'une dotation en ressources de calcul estimée à 3 millions d'euros et le modèle a été entraîné pendant cent dix-sept jours entre mars et juillet 2022 sur le supercalculateur Jean-Zay, l'un des plus puissants d'Europe, installé en région parisienne à l'Institut du développement et des ressources en informatique scientifique (Idris), le centre de calcul intensif du CNRS (voir *La rem* n°52, p.31). Le modèle Bloom est ainsi disponible sur la plateforme de l'entreprise partenaire Hugging Face. *« Il peut être téléchargé sous une licence RAIL, pour « Responsible AI License », proposée pour la première fois lors du projet BigScience. Cette licence permet aux développeurs d'empêcher que les logiciels qu'ils développent ne soient utilisés dans des applications nuisibles. Proche des licences open source existantes, elle pose certaines conditions d'utilisation pour un logiciel ou un code source, comme l'interdiction d'une utilisation pour générer de fausses nouvelles ou généralement des textes sans préciser qu'une machine en est à l'origine, pour diffuser des informations privées ou des conseils médicaux »*, explique le CNRS. Bloom est actuellement téléchargé entre 40 000 et 50 000 fois par mois pour des démonstrations, des projets de recherche, des projets d'enseignement ou encore par des entreprises qui souhaitent tester le système. Si GPT est dorénavant une boîte noire, Bloom a le mérite d'être *« « exemplaire » en matière de transparence, les bases de données utilisées pour l'entraînement étant connues et interrogeables, et les algorithmes « visibles et documentés » »*, explique Pierre-François Lavallée, chercheur au CNRS et directeur de l'Idris.

Exactement le contraire d'OpenAI, qui a reçu en échange d'une licence exclusive 1 milliard de dollars en 2019, puis 10 milliards de dollars en 2022 de la part de Microsoft, alors que l'entreprise se présentait en 2015 comme *« une société de recherche en intelligence artificielle à but non lucratif »* et que GPT était alors disponible en open source. Pour justifier ce revirement, Ilya Sutskever, l'un des cofondateurs d'OpenAI, explique dans un brillant exercice de mauvaise foi que c'est justement dans un souci de responsabilité et d'éthique que GPT est dorénavant fermé. *« Nous nous sommes trompés. Nous nous sommes carrément trompés. Si vous pensez, comme nous, qu'à un moment donné l'IA sera extrêmement, incroyablement puissante, alors il n'est tout simplement pas logique d'ouvrir le code source. C'est une mauvaise idée… Je m'attends à ce que, dans quelques années, il devienne évident pour tout le monde qu'il n'est pas judicieux de mettre l'IA en libre accès »*, justifiait-il dans les colonnes de *The Verge* en mars 2023. Pourtant, *« les algorithmes utilisés fonctionnent souvent comme des boîtes noires et les IA génératives peuvent fournir des réponses trop moyennées, erronées ou biaisées par des préjugés présents dans les données utilisées pour les entraîner »*, explique Pierre-François Lavallée. Suite au revirement faisant de GPT-4 une de ces boîtes noires qui, selon OpenAI, ne propose plus aucun détail *« sur l'architecture, la taille du modèle, le matériel, le calcul d'entraînement, la construction de l'ensemble de données, la méthode d'entraînement »*, bon nombre d'acteurs de la recherche scientifique et d'experts se montrent particulièrement inquiets et réclament même un moratoire jusqu'à la mise en place de systèmes de sécurité. Google, en retard dans le domaine, a, quant à lui, licencié en 2022 les chercheuses Timnit Gebru et Margaret Mitchell, respectivement ancienne fondatrice et coresponsable de l'équipe d'éthique en intelligence

artificielle au sein du groupe. En 2023, le chercheur marocain El Mahdi El Mhamdi, à la suite de la publication d'un article prônant la nécessité de freiner le déploiement précoce de nouvelles intelligences artificielles qui a déplu à la firme de Mountain View, a démissionné de son poste de *senior scientist*.

Bloom est, quant à lui, aligné sur la stratégie européenne telle que définie par l'Artificial Intelligence Act. Présentée en avril 2021, cette proposition de règlement, qui vise à encadrer et promouvoir l'usage d'intelligences artificielles centrées sur l'humain et dignes de confiance, devra être revue à la suite des interrogations soulevées par l'ouverture au grand public de ChatGPT. Les modèles d'intelligence artificielle qui respectent une éthique, une transparence du code et un encadrement de leurs usages inciteront peut-être les institutions et les entreprises européennes à y recourir, afin d'éviter de confier leurs données à des prestataires étrangers, notamment américains et chinois. À moins qu'elles ne renoncent tout simplement à utiliser l'intelligence artificielle dans le cadre de leur activité. Selon un audit de la Cour des comptes européenne de mars 2023, « Artificial intelligence in the EU », *« moins d'une entreprise européenne sur dix (soit 8 %) utilisait l'IA en 2021. Mais cette moyenne cache des différences notables entre États membres : alors que près d'une entreprise sur quatre au Danemark (24 %) et plus d'une sur six au Portugal et en Finlande (17 % et 16 %, respectivement) avaient recours à l'IA, elles sont moins d'une sur vingt à le faire en Tchéquie, en Grèce, en Lettonie et en Lituanie (4 %), en Bulgarie, en Estonie, à Chypre, en Hongrie et en Pologne (3 %), ainsi qu'en Roumanie (1 %) »*. La France, quant à elle, se situe à 7 %.

Sources :

- Commission nationale de l'informatique et des libertés, « Intelligence artificielle, de quoi parle-t-on ? », cnil.fr, 5 avril 2022.
- Véronique Étienne, François Yvon, Pierre-François Lavallée, « Livraison du plus grand modèle de langue multilingue « open science » jamais entraîné », cnrs.fr, 12 juillet 2022.
- Clément Bohic, « BLOOM : les choses à savoir sur ce « méga-modèle » d'IA au sang français », silicon.fr, 13 juillet 2022.
- Martin Clavey, « BLOOM : l'ambitieux modèle de langage de l'open science », nextinpact.fr, 18 octobre 2022.
- Thomas Calvi, « Graphcore et Aleph Alpha présentent un modèle d'IA clairsemé à 80 % », actuia.com, 1er décembre 2022.
- Guillaume Erner, « ChatGPT, Bard ou Ernie : quelle IA va l'emporter ? », podcast Radiofrance, intervention de Laurence Devillers, radiofrance.fr, 13 février 2023.
- Julien Lausson, « L'Europe parviendra-t-elle à contenir les IA comme ChatGPT ? », numerama.com, 3 mars 2023.
- Mélicia Poitiers, « LightOn lance Paradigm, une plateforme d'IA générative française pour les grandes entreprises », usine-digitale.fr, 9 mars 2023.
- James Vincent, « OpenAI co-founder on company's past approach to openly sharing research : « We were wrong » », theverge.com, 15 mars 2023.
- Cour des comptes européenne, « L'intelligence artificielle dans la ligne de mire de la Cour des comptes européenne », communiqué de presse, eca.europa.eu, 20 mars 2023.

- Christophe Auffray, « L'Europe, compétitive demain sur l'IA ? La Cour des comptes doute et enquête », zdnet.fr, 22 mars 2023.
- « FAZ' : SAP veut entrer dans le capital de la start-up allemande d'IA Aleph Alpha », dpa-AFX, zonebourse.com, 30 mars 2023.
- CNRS Info, « La recherche française face à ChatGPT », cnrs.fr, 25 avril 2023.
- Pablo Maillé, « Les IA génératives font diversion à celles des réseaux sociaux », usbeketrica.com, 9 mai 2023.

Categorie

1. Techniques

date créée
11 juillet 2023
Auteur
jacquesandrefines