Ce texte a-t-il servi à préentraîner une intelligence arti?cielle ?

Description

Deux équipes de chercheurs, l'une des universités de Washington et de Princeton et l'autre de l'Imperial College de Londres, se sont intéressées à la question de la transparence des données avec lesquelles les grands modèles de langage ont été préentraînés.

Microsoft, Facebook, Google proposent dorénavant tous de grands modèles de langage (Large Language Models, LLM) à destination du grand public, et ChatGPT, Llama 2 ou encore Gemini ont définitivement trouvé leur audience. Mais les modalités selon lesquelles ces logiciels américains sont conçus restent opaques, notamment les corpus de textes à partir desquels ils ont été préentraînés, parfois protégés par le droit d'auteur ou intégrant des données et informations personnelles.

Pour atteindre 100 millions d'utilisateurs, il aura fallu deux ans et demi à Instagram, neuf mois à TikTok et seulement deux mois à ChatGPT. Au point que certaines écoles et universités s'inquiètent des contenus générés à partir de ces programmes informatiques. Aux États-Unis, Jenna Lyle, porte-parole du département de l'Éducation de la ville de New York, a récemment expliqué qu'en raison « *des inquiétudes concernant l'impact négatif sur l'apprentissage des élèves, ainsi que la sécurité et l'exactitude du contenu, l'accès à ChatGPT sera restreint sur les réseaux et les appareils des écoles publiques de la ville* ». Mais à la question de savoir si un texte a été écrit ou non par ChatGPT s'en ajoute une autre, tout aussi pertinente : ce texte a-t-il servi à préentraîner ChatGPT ? Autrement dit, telles œuvres ou tels documents protégés par le droit d'auteur ou contenant des informations personnelles ont-ils été utilisés par les géants du web pour entraîner leur grand modèle de langage ?

GPT – pour *generative pre-trained transformer*, transformeur génératif préentraîné – est un grand modèle de langage servant à générer un texte dont on ne saurait dire s'il a été produit par une machine ou écrit par un humain. Tous les grands modèles de langage de ce type sont « préentraînés » selon différentes méthodes d'apprentis- sage automatiques, à partir d'immenses corpus de données textuelles, contenant plusieurs milliards de mots. On peut citer, parmi les plus grands ensembles de données textuelles disponibles aujourd'hui, The Common Crawl, The Pile, MassiveText mais aussi Wikipedia ou encore GitHub. Ces grandes « machines » de langage, qui ne connaissent pas les mots, ne font que manipuler des tokens. En informatique appliquée à l'intelligence artificielle (IA), on parle de *tokenization* pour désigner la conversion d'une séquence de caractères, d'un mot ou de la ponctuation en une liste de symboles, en anglais *tokens*, qui seront manipulables par le programme informatique. Ce « préentraînement » vise à calculer statistiquement la probabilité du « token » suivant dans une phrase. Le corpus de données textuelles sur lequel ces intelligences artificielles sont « formées » est donc, littéralement, essentiel. La première version de ChatGPT, d'OpenAI, a été entraînée sur BookCorpus, composé de 985 millions de mots. Dirigé par

Hugging Face et déployé en juillet 2022 (voir *La rem* n°65-66, p.27), le programme scienti- fique BLOOM, qui fournit le détail des données d'apprentissage de son modèle, s'est appuyé sur 1,6 téra- octets de textes prétraités, convertis en 350 milliards de tokens. Son apprentissage a duré onze semaines, totalisé cinq millions d'heures de calcul, et a été effectué sur le supercalculateur Jean Zay, installé en région parisienne à l'Institut du développement et des ressources en informa- tique scientifique (Idris), l'un des centres de calcul intensif du CNRS (voir *La rem* n°52, p.31). En 2022, des chercheurs de l'université d'Aberdeen en Écosse, de l'université de Tübingen en Allemagne et du MIT (Massachusetts Institute of Technology) aux États-Unis se sont d'ailleurs déjà posé la question de savoir quand le « *stock de données de qualité* » allait être épuisé. Leur réponse est « *avant 2026* ». Selon leur définition, les données de qualité sont « *des données ayant passé des filtres d'utilité ou de qualité* » et issues de livres, d'articles de presse, de publications scientifiques, de projets de code open-source et de Wikipédia. L'importance d'un « *stock de données de qualité* » devient en effet cruciale pour garantir le perfectionnement de ces intelligences artificielles, et certaines entreprises n'ont pas hésité à préentraîner leurs modèles de langage sur des contenus protégés, ce qui pourrait expliquer en partie leur réticence à divulguer les corpus employés.

Une équipe de chercheurs des universités de Washington et de Princeton et une autre de l'Imperial College de Londres ont chacune publié en prépublication récemment un article sur arXiv, une archive ouverte de plus de 2,4 millions de prépublications électroniques d'articles scientifiques, en s'attaquant à cette question de transparence. Selon Gemini, l'IA générative de Google, « *si le premier* [de ces deux articles] *se concentre sur l'inférence d'appartenance au niveau du document, le second a une portée plus large et s'intéresse à la détection de données de pré- apprentissage* ». Pour reprendre une définition de la Cnil, une attaque par inférence d'appartenance (*membership inference attack*) « *vise à permettre à un attaquant d'acquérir des connaissances sur les données utilisées pour la production du modèle d'IA*». Mais, plus largement, une attaque par inférence survient lorsqu'un attaquant est capable de déduire, à partir d'informations triviales, des informations plus solides sur une base de données, sans y accéder directement. Ces attaques sont considérées comme une réelle atteinte à la vie privée, notamment dans les domaines des capteurs mobiles, des objets connectés et de l'internet des objets. Par exemple, les données des accéléromètres d'applications mobiles permettent de déduire des informations à propos d'une personne, en fonction de ses mouvements, comme un comportement au volant, un niveau d'ivresse, une situation statique, etc. Ces attaques par inférence d'appartenance sont aujourd'hui utilisées dans le domaine de l'intelligence artificielle parce que les concepteurs des modèles de langage gardent confidentielles les données à partir desquelles ces IA génératives ont été préentraînées.

Les chercheurs anglais, dans un article intitulé « Les neurones ont-ils lu votre livre ? Inférence d'appartenance au niveau du document pour les grands modèles de langage », démontrent la faisabilité de l'inférence d'appartenance au niveau du document pour les grands modèles de langage et introduisent une nouvelle métrique pour quantifier la probabilité d'appartenance à un document. La méthode consiste à distinguer deux groupes de documents dont le premier a probablement été vu par le modèle de langage pendant l'apprentissage, comme Common Crawl, des pétaoctets de données collectées sur le web depuis 2008, et dont la moitié est en anglais, ou RedPajama, un ensemble de données publiques – 20 000 milliards

de mots, publié spécifiquement pour la formation de grands modèles de langage par Together, une startup américaine créée en 2022 et installée en Californie, ou encore arXiv, ou le projet Gutenberg, une collection de 70 000 livres, en majorité de langue anglaise, et dont les droits d'auteur américains ont expiré, ces corpus de textes étant systématiquement utilisés par ces grands modèles de langage. Le second groupe de documents comporte des textes que le modèle n'est pas susceptible d'avoir vus. À partir de ces documents connus, dits « positifs », et inconnus, dits « négatifs », il est calculé un score d'inférence utilisant cette méthodologie : « *demander au modèle des prédictions au niveau du token, normaliser les prédictions en fonction de la fréquence du token, agréger les prédictions au niveau du document et, enfin, construire un méta-classificateur* ». Selon que le score d'inférence est élevé ou faible, il est alors possible de déduire si un document a été utilisé ou non pour l'entraînement du modèle. Les chercheurs américains, dans un article intitulé « Détection des données de préentraînement des grands modèles de langage » se sont, quant à eux, intéressés aux manières d'identifier les sources de données utilisées pour entraîner ces modèles. Leur approche repose, entre autres, sur un système d'horodatage des données de l'encyclopédie en ligne Wikipédia, une source de contenu systématiquement utilisée pour entraîner les grands modèles de langage. Un benchmark, appelé WIKIMIA (pour Wikipedia Membership Inference Attack), repose sur la distinction entre des données anciennes, qualifiées de « *données vues pendant le préentraînement* » et des données récentes qualifiées de « *données non vues* ». Il s'agit de mesurer la probabilité que le modèle génère le token suivant dans une phrase de manière inattendue, compte tenu de son contexte. Les chercheurs utilisent MIN-K% PROB, qui veut dire « minimum k pour cent probabilités », le « k » représentant le pourcentage de probabilités que l'événement se produise et qui désigne donc une métrique pour évaluer la performance d'un modèle de langage lors de la génération de texte, c'est-à-dire la probabilité minimale que le modèle génère un token donné. Ainsi, « *parce que MIN-K% PROB calcule la moyenne des probabilités des tokens atypiques, un texte non vu a tendance à contenir quelques mots atypiques avec des probabilités faibles, alors qu'un texte vu est moins susceptible de contenir de tels mots* » expliquent-ils. L'originalité de la méthode tient à ce qu'elle peut être appliquée, selon les chercheurs, « *sans aucune connaissance du corpus de préentraînement ni aucun entraînement supplémentaire, contrairement aux méthodes MIA existantes* ».

Books3 est une collection de 195 000 livres électroniques piratés, représentant 37 gigaoctets de données textuelles, créée par le chercheur en IA Shawn Presser en 2020, et incluse dans un projet plus vaste appelé The Pile, dont l'objet était justement de fournir des données open source pour les modèles de langage et que tous les grands modèles de langage ont utilisés. Or, aujourd'hui, bon nombre d'auteurs de ces livres piratés ont manifesté leur refus, *a posteriori*, que leurs œuvres soient utilisées pour préentraîner ces IA génératives. En décembre 2023, Meta a même admis que la collection Books3 avait servi à entraîner Llama 1 et Llama 2. En expérimentant leur méthode de détection de contenus, les chercheurs américains arrivent à la même conclusion pour l'IA d'OpenAI : « *D'après nos expériences de détection de livres protégés par le droit d'auteur, nous observons des preuves solides suggérant que GPT-3 1 est préentraîné sur des livres protégés par le droit d'auteur provenant de la collection Books3* ».

Même si des méthodes sont développées pour se protéger de ces détections, il semble de plus en plus complexe pour ces grandes entreprises de ne pas partager les données sur lesquelles leurs grands modèles de

langage respectifs sont préentraînés. Lorsque l'on interroge les IA Gemini de Google et GPT-3 d'OpenAI sur les enjeux de cette transparence, elles décrivent, sans en omettre, les risques liés à « *leur opacité dont notamment l'amplification des biais, discriminations et stéréotypes présents dans la société, aux questions de propriété intellectuelle et de protection des droits d'auteur, aux questions liées à l'utilisation de données personnelles sensibles, aux problèmes de la reproductibilité et de l'explicabilité des résultats, essentielles pour garantir la fiabilité et l'équité des systèmes d'IA générative* ». Mais elles utilisent exactement les mêmes arguments pour justifier pourquoi les données de préentraînement ne sont pas publiques. Ainsi pour GPT-3, « *la non-divulgation des données de préentraînement est principalement due à des considérations de protection des droits d'auteur, de confidentialité, de propriété intellectuelle, de responsabilité et de sécurité* ».

Sources :

- Villalobos Pablo, Sevilla Jaime, Heim Lennart, Besiroglu Tamay, Hobbhahn Marius, Ho Anson, « Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning », ArXiv, abs/2211.04325, October 26, 2022.
- Elsen-Rooney Michael, « NYC education department blocks ChatGPT on school devices, networks », chalkbeat.org, January 4, 2023.
- Gayte Aurore, « Ce texte a-t-il été écrit par ChatGPT ? Cette IA vous le dit », numerama.com, 1er février 2023.
- Chow Andrew R., « How ChatGPT Managed to Grow Faster Than TikTok or Instagram », time.com, February 8, 2023.
- Reisner Alex, « Revealed: the authors whose pirated books are powering generative AI », theatlantic.com, August 19, 2023.
- Reisner Alex, « These 183,000 books are fueling the biggest fight in publishing and tech », theatlantic.com, September 25, 2023.
- Miller Katharine, « Introducing The Foundation Model Transparency Index », hai.stanford.edu, October 18, 2023.
- Meeus Matthieu, Jain Shubham, Rei Marek, Montjoye Yves-Alexandre de, « Did the Neurons Read your Book? Document-level Membership Inference for Large Language Models », ArXiv, abs/2310.15007, October 23, 2023.
- Shi Weijia, Ajith Anirudh, Xia Mengzhou, Huang Yangsibo, Liu Daogao, Blevins Terra, Chen Danqi, Zettlemoyer Luke, « Detecting Pretraining Data from Large Language Models », ArXiv, abs/2310.16789, October 25, 2023.
- Larousserie David, « Comment savoir si un contenu a été utilisé par une intelligence artificielle ? », lemonde.fr, 16 novembre 2023.
- Barrabi Thomas, « New York Times sues OpenAI, Microsoft for seeking to "free-ride" on its articles to train chatbots », nypost.com, December 27, 2023.
- Van der Sar Ernesto, « Meta Admits Use of "Pirated" Book Dataset to Train AI », torrentfreak.com, January 11, 2024.

Categorie

1. Techniques

date créée
21 mars 2024
Auteur
jacquesandrefines