Lâ€™IA : une course eï¬€rÃ©nÃ©e Ã la donnÃ©e

Description

Ã€ la fin de lâ€™annÃ©e 2021, OpenAI, Ã lâ€™origine de ChatGPT, a Ã©tÃ© confrontÃ© Ã un problÃ¨me de taille pour entraÃ®ner la nouvelle version de son intelligence artificielle (IA) gÃ©nÃ©rative, GPT-4 : lâ€™Ã©puisement des contenus en anglais de qualitÃ©, lâ€™intÃ©gralitÃ© ayant dÃ©jÃ Ã©tÃ© utilisÃ©e.

La conclusion dâ€™une enquÃªte menÃ©e par cinq journalistes duÂ *New York TimesÂ* parue en avril 2024 est sans appel : Â«Â *OpenAI, Google et Meta ont ignorÃ© les politiques dâ€™entreprise, altÃ©rÃ© leurs propres rÃ¨gles et ont discutÃ© de la possibilitÃ© de contourner la loi sur le droit dâ€™auteur alors quâ€™ils cherchaient des informations en ligne pour entraÃ®ner leurs derniers systÃ¨mes dâ€™intelligence artificielle.Â* Â»

En janvier 2020, Jared Kaplan, physicien thÃ©oricien Ã lâ€™universitÃ© Johns-Hopkins, a publiÃ© un article majeur dans le domaine de lâ€™IA qui dÃ©montre quâ€™un modÃ¨le neuronal est caractÃ©risÃ© par quatre paramÃ¨tres â€“ la taille du modÃ¨le, la taille du jeu de donnÃ©es dâ€™entraÃ®nement, le coÃ»t de lâ€™entraÃ®nement et la performance aprÃ¨s lâ€™entraÃ®nement â€“, et que ces paramÃ¨tres sont liÃ©s par des lois statistiques dont lâ€™une stipule que plus lâ€™on dispose de donnÃ©es pour entraÃ®ner un grand modÃ¨le de langage, mieux il fonctionnera. Â«Â *Tout le monde a Ã©tÃ© trÃ¨s surpris que ces tendances â€“ ces lois dâ€™Ã©chelle comme nous les appelons â€“ soient essentiellement aussi prÃ©cises que ce que lâ€™on observe en astronomie ou en physiqueÂ* Â», a-t-il dÃ©clarÃ© aprÃ¨s la publication de cet article, coÃ©crit avec neuf chercheurs dâ€™OpenAI.

Depuis, les gÃ©ants de la tech impliquÃ©s dans lâ€™IA sont engagÃ©s dans une course effrÃ©nÃ©e Ã la donnÃ©e. Les grands modÃ¨les de langage sont en effet prÃ©entraÃ®nÃ©s Ã partir dâ€™immenses corpus de donnÃ©es textuelles, contenant plusieurs milliards de mots, reprÃ©sentÃ©s sous la forme de tokens, afin dâ€™Ãªtre manipulables par le programme informatique (voirÂ *La remÂ* nÂ°65-66, p.27). Selon Pablo Villalobos, de lâ€™Institut de recherche Epoch, citÃ© par leÂ *Wall Street Journal*, ChatGPT-4 a Ã©tÃ© entraÃ®nÃ© Ã partir de 12 000 milliards de tokens et, si ChatGPT-5 suit la mÃªme Ã©volution que le passage de ChatGPT-3 Ã ChatGPT-4, 60 000 Ã 100 000 milliards de tokens seront nÃ©cessaires.Â Or, il nâ€™existe plus assez de nouvelles donnÃ©es de qualitÃ© pour entraÃ®ner ces IA,Â câ€™est-Ã -dire des donnÃ©es issues de livres, dâ€™articles de presse, de publications scientifiques, de code informatique open source, de contenus provenant de WikipÃ©dia ainsi que de tous les corpus de textes spÃ©cialement dÃ©diÃ©s Ã lâ€™entraÃ®nement de ces programmes.

En 2022, des chercheurs de lâ€™universitÃ© dâ€™Aberdeen en Ã‰cosse, de lâ€™universitÃ© de

TÃ¼bingen en Allemagne et du MIT (Massachusetts Institute of Technology) aux Ã‰tats-Unis avaient dâ€™ailleurs prÃ©dit ce phÃ©nomÃ¨ne de rarÃ©faction des contenus Ã la fois de qualitÃ© et non protÃ©gÃ©s par le droit dâ€™auteur ou par les rÃ©glementations en vigueur sur les donnÃ©es personnelles (voirÂ *La remÂ* nÂ°65-66, p.27).

Plusieurs solutions sâ€™oï¬€rent alors aux dÃ©veloppeurs de ces IA

Ils peuvent passer des contrats avec des fournisseurs de contenus, en particulier les mÃ©dias ; entraÃ®ner les IA sur des contenus crÃ©Ã©s par des IA ou bien utiliser sans autorisation des contenus protÃ©gÃ©s par le droit dâ€™auteur. Alors que, depuis dÃ©cembre 2023, lâ€™entreprise fait toujours lâ€™objet de poursuites pour violation des droits dâ€™auteur de la part duÂ *New York Times*, OpenAI a conclu un premier accord pluriannuel avec lâ€™agence Associated Press la mÃªme annÃ©e.Â En 2024, des accords similaires ont Ã©tÃ© signÃ©s avecÂ *Le Monde*Â en France, Prisa pourÂ *El Pais*Â en Espagne, Springer pour leÂ *Bild*,Â *Die Welt*,Â *Politico*Â en Allemagne, et rÃ©cemment avec leÂ *Financial Times*Â en Angleterre.Â Si peu dâ€™informations filtrent sur ces accords ; leÂ *Financial Times*Â avait justement rÃ©vÃ©lÃ©, concernant le groupe allemand, la signature dâ€™un contrat se chiffrant en dizaines de millions dâ€™euros par an.

Une autre solution pour pallier la pÃ©nurie de contenus serait dâ€™entraÃ®ner des IA sur des donnÃ©es synthÃ©tiques, câ€™est-Ã -dire des donnÃ©es elles-mÃªmes produites par une IA. Â«Â *Nous avons remarquÃ© que si le modÃ¨le est suï¬€samment bon et que la quantitÃ© de donnÃ©es gÃ©nÃ©rÃ©es nâ€™est pas trop importante par rapport aux rÃ©elles, alors le modÃ¨le ne dÃ©gÃ©nÃ¨re pasÂ* Â», explique Quentin Bertrand, chercheur au Mila et Ã lâ€™UniversitÃ© de MontrÃ©al, coauteur dâ€™une Ã©tude publiÃ©e en janvier 2024 et intitulÃ©e Â« StabilitÃ© du recyclage itÃ©ratif des modÃ¨les gÃ©nÃ©ratifs sur leurs propres donnÃ©es Â».Â MÃªme si le risque dâ€™un effondrement du modÃ¨le,Â *model collaps*, est rÃ©el, et que certains chercheurs considÃ¨rent les donnÃ©es synthÃ©tiques comme lâ€™Ã©quivalent informatique de la consanguinitÃ©,Â toutes les grandes entreprises du secteur travaillent activement sur le sujet.

Autre option, enfin : se servir, sans autorisation, de contenus protÃ©gÃ©s par le droit dâ€™auteur. En 2021, les chercheurs dâ€™OpenAI ont crÃ©Ã© un outil de reconnaissance vocale, appelÃ© Whisper, capable de transcrire lâ€™audio des vidÃ©os hÃ©bergÃ©es sur YouTube sous la forme de textes, afin de mettre la main sur des contenus inÃ©dits. Bien Ã©videmment, Google interdit tout autant lâ€™utilisation par des tiers des vidÃ©os publiÃ©es sur la plateforme que lâ€™usage dâ€™outils permettant un accÃ¨s automatisÃ© Ã ces mÃªmes vidÃ©os. Or, selon lâ€™enquÃªte publiÃ©e par leÂ *New York Times*, Â«Â *certains employÃ©s de Google savaient quâ€™OpenAI avait exploitÃ© des vidÃ©os de YouTube pour obtenir des donnÃ©es*Â [â€¦]. *Mais ils nâ€™ont pas arrÃªtÃ© OpenAI parce que Google avait Ã©galement utilisÃ© des transcriptions de vidÃ©os YouTube pour entraÃ®ner ses modÃ¨les dâ€™intelligence artificielle*Â Â». Alors que lui-mÃªme violait les droits dâ€™auteur des utilisateurs de sa propre plateforme, Google ne pouvait prendre le risque dâ€™accuser OpenAI de faire la mÃªme chose.

Open AI, Meta ou encore Google sont donc confrontÃ©s en mÃªme temps au mÃªme problÃ¨me : la pÃ©nurie de donnÃ©es de qualitÃ© pour entraÃ®ner leur IA. Â«Â *La seule chose qui nous empÃªche dâ€™Ãªtre aussi bons que ChatGPT, câ€™est littÃ©ralement le volume de donnÃ©esÂ* Â», dÃ©clarait Nick Grudin, vice-prÃ©sident des partenariats et du contenu mondial de Meta, dans des propos rapportÃ©s par leÂ *New York Times*. Des enregistrements de rÃ©unions internes Ã Meta datant de 2023, impliquant des responsables, des avocats et des ingÃ©nieurs obtenus parÂ *The Times*, indiquent Ã©galementÂ des discussions au sujet de la collecte de donnÃ©es protÃ©gÃ©es par le droit dâ€™auteur sur lâ€™ensemble du web en assumant le risque de poursuites judiciaires,Â la nÃ©gociation de licences avec les Ã©diteurs, les artistes, les musiciens et lâ€™industrie de lâ€™information Ã©tant jugÃ©es trop chronophage.

OpenAI reconnaÃ®t dâ€™ailleurs ouvertement lâ€™utilisation de documents protÃ©gÃ©s par le droit dâ€™auteur. Dans un tÃ©moignage Ã©crit, fourni dans le cadre dâ€™une enquÃªte sur les grands modÃ¨les de langage menÃ©e par la Commission des communications et du numÃ©rique de la Chambre des Lords du Parlement anglais, lâ€™entreprise tient ces propos : Â«Â *Ã‰tant donnÃ© que le droit dâ€™auteur couvre aujourdâ€™hui pratiquement toutes les formes dâ€™expression humaine â€“ y compris les articles de blog, les photographies, les messages de forum, les bouts de code de logiciel et les documents gouvernementaux â€“, il serait impossible dâ€™entraÃ®ner les principaux modÃ¨les dâ€™IA actuels sans utiliser des documents protÃ©gÃ©s par le droit dâ€™auteur.Â* Â»

Le principal argument juridique, systÃ©matiquement invoquÃ© par OpenAI pour justifier ces violations du droit dâ€™auteur, repose sur leÂ *fair use*.Â En dÃ©cembre 2023, en rÃ©ponse Ã la plainte pour violation du copyright dÃ©posÃ©e par leÂ *New York Times*, OpenAI considÃ¨re quâ€™Â«Â *il est clair depuis longtemps que lâ€™utilisation non- consommatrice de matÃ©riel protÃ©gÃ© par le copyright (comme la formation de grands modÃ¨les de langage) est protÃ©gÃ©e par leÂ* fair use Â». Ce principe juridique de propriÃ©tÃ© intellectuelle propre au droit amÃ©ricain dâ€™un usage loyal ou Ã©quitable autorise, dans des circonstances prÃ©cises, Ã utiliser des Å“uvres protÃ©gÃ©es par le droit dâ€™auteur sans lâ€™autorisation prÃ©alable du titulaire des droits, Ã condition que cet usage soit considÃ©rÃ© comme raisonnable et Ã©quitable. Au-delÃ du raisonnement juridique, qui semble tout Ã fait inappropriÃ© si lâ€™on considÃ¨re les millions de livres ou dâ€™articles et les millions dâ€™heures de vidÃ©os YouTube protÃ©gÃ©s par le droit dâ€™auteur pris sans autorisation par ces IA et lâ€™activitÃ© dorÃ©navant purement commerciale de lâ€™entreprise,Â lâ€™argumentation dâ€™OpenAI consiste Ã©galement Ã minimiser lâ€™importance et le volume des contenus collectÃ©s sans autorisationÂ auprÃ¨s dâ€™une entreprise en les rapportant Ã lâ€™intÃ©gralitÃ© des donnÃ©es sur lesquelles le modÃ¨le dâ€™IA est entraÃ®nÃ©.

Autrement dit, puisque la ligne droite est plus que la somme des points qui la composent, pour reprendre une idÃ©e dâ€™Aristote, le contenu duÂ *New York Times*, comme toute source unique, ne Â«Â *contribue pas de maniÃ¨re significativeÂ* Â» Ã lâ€™entraÃ®nement de ChatGPT, aurait dÃ©clarÃ© OpenAI au mÃ©dia amÃ©ricain. Et une fois quâ€™un contenu a servi Ã prÃ©entraÃ®ner une IA, il est extrÃªmement complexe de le Â« retirer Â» ou de le faire Â« dÃ©sapprendre Â» ou mÃªme de montrer ou

prouver quoi que ce soit puisque ces programmes sont des boÃ®tes noires, y compris pour leurs concepteurs (voirÂ *La remÂ *nÂ°68, p.41). On se demande de quelle loyautÃ© ou Ã©quitÃ© se prÃ©valent les dirigeants de ces gÃ©ants de la tech qui, sans scrupule, foulent aux pieds les lÃ©gislations nationales et internationales, au motif que celui qui aura fait avaler le plus de donnÃ©es Ã son IA remportera le marchÃ©.

Sources :

- UK Parliament, Â« OpenAI â€“ written evidence (LLM0113). House of Lords Communications and Digital Select Committee inquiry: Large language models Â», December 5, 2023.Â
- Thomas Daniel, Murgia Madhumita, Â« Axel Springer strikes landmark deal with OpenAI over access to news titles Â», ft.com, December 13, 2023.Â
- Marin JÃ©rÃ´me, Â« Pour Ã©viter de nouveaux procÃ¨s, OpenAI nÃ©gocie avec des Ã©diteurs de presse Â», usine-digitale.fr, 5 janvier 2024.Â
- SÃ©ramour CÃ©lia, Â« Les Å“uvres soumises au droit dâ€™auteur indispensables pour entraÃ®ner ChatGPT, admet OpenAI Â», usine-digitale.fr, 9 janvier 2024.Â
- Clavey Martin, Â« OpenAI contre-attaque et accuse leÂ *New York Times*Â dâ€™avoir Â«Â hackÃ©Â Â» ses produits Â», next.ink, 29 fÃ©vrier 2024.Â
- Metz Cade, Kang Cecilia, Frenkel Sheera, Thompson Stuart A., Grant Nico, Â« How Tech Giants Cut Corners to Harvest Data for A.I. Â»,Â *The New York Times*, nytimes.com, April 6, 2024.Â
- Seetharaman Deepa, Â« For Data-Guzzling AI Companies, the Internet Is Too Small Â», wsj.com, April 1, 2024.
- SÃ©ramour CÃ©lia, Â« AprÃ¨s Axel Springer etÂ *Le Monde*, OpenAI sâ€™achÃ¨te les faveurs duÂ *Financial Times*Â Â», usine-digitale.fr, 30 avril 2024.

Categorie

1. Techniques

date crÃ©Ã©e
4 juillet 2024
Auteur
jacquesandrefines