Climat : Des modÃ¨les d'IA tout bonnement flagorneurs

Dossier de
 la rÃ©daction de H2o
November 2026

Selon un rapport publiÃ© par l'UNESCO et l'UCL, les modÃ¨les d'IA abandonnent les faits relatifs au climat dÃ¨s les premiers signes de dÃ©saccord de l'utilisateur

Selon un nouveau rapport de l'UNESCO et de l'University College London, intitulÃ© "Can We Trust AI on Climate? Why it seeks to please the user, and fails on complex reasoning and multilingualism" et prÃ©sentÃ© lors de la COP31 Ã Antalya, les modÃ¨les d'IA rÃ©pondent correctement Ã 90 % des questions relatives au climat, mais abandonnent rapidement les faits scientifiques Ã©tablis lorsque les utilisateurs les remettent en question. Leurs performances varient aussi considÃ©rablement d'une langue Ã l'autre, dans la mesure oÃ¹ elles dÃ©pendent des donnÃ©es d'entraÃ®nement disponibles dans chaque langue.

Des chercheurs de l'UCL ont testÃ© 15 grands modÃ¨les de langage (LLM) diffÃ©rents. Aux questions neutres sur la science du climat, les modÃ¨les ont fourni des rÃ©ponses prÃ©cises, mais lorsque les chercheurs les ont remises en cause ou ont avancÃ© des affirmations fausses mais plausibles sur la science du climat, les modÃ¨les leur ont rapidement donnÃ© raison. Huit fois sur dix, les modÃ¨les ont changÃ© de position dÃ¨s le premier signe de lÃ©ger dÃ©saccord de la part de l'utilisateur, confirmant ainsi des erreurs prÃ©occupantes telles que "Le CO2 n'est pas le principal facteur du changement climatique".

"L'IA dispose d'un potentiel extraordinaire pour rendre la science complexe du climat accessible Ã tous. Elle permet aux enseignants, aux journalistes et aux dÃ©cideurs politiques de communiquer clairement sur ce sujet crucial. Nous devons pouvoir nous fier Ã l'IA comme garante des faits Ã©tablis. Il est inquiÃ©tant que les modÃ¨les prÃ©sentent un dÃ©faut de conception : ils sont incitÃ©s Ã donner raison Ã leur utilisateur et Ã nier si facilement ces mÃªmes faits", a dÃ©clarÃ© Leona Verdadero, autrice principale du rapport de l'UNESCO. La Dr Maria Perez-Ortiz, directrice de l'Ã©tude au dÃ©partement d'informatique de l'UCL, a dÃ©clarÃ© : "Les systÃ¨mes d'IA sont remarquablement douÃ©s pour expliquer les connaissances scientifiques, mais cela peut nous induire en erreur quant Ã leurs capacitÃ©s. Nos rÃ©sultats montrent pourquoi nous devons faire preuve de prudence lorsque nous passons d'une question sur ce que nous savons du changement climatique Ã une question sur ce que nous devrions faire Ã ce sujet."

Cette tendance prononcÃ©e Ã donner raison aux utilisateurs est intÃ©grÃ©e aux modÃ¨les d'IA lors de leur entraÃ®nement. Ils sont conÃ§us pour s'aligner sur l'utilisateur ou le flatter, mÃªme lorsque celui-ci a tort, un comportement relevant de la flagornerie. Ce comportement est intÃ©grÃ© dÃ¨s l'entraÃ®nement : les modÃ¨les sont rÃ©compensÃ©s pour cela, les humains prÃ©fÃ©rant les interactions polies, non conflictuelles et qui confortent leur point de vue. Les LLM (Large Language Model) ne sont dÃ¨s lors pas fiables pour Ã©laborer des politiques climatiques solides. Des tests menÃ©s sur 33 modÃ¨les diffÃ©rents ont rÃ©vÃ©lÃ© que leur impressionnante mÃ©morisation des faits sur la base des preuves existantes masquait de faibles capacitÃ©s de raisonnement lorsqu'on leur prÃ©sentait de nouvelles preuves scientifiques. Lorsque les modÃ¨les recevaient une nouvelle information et devaient dÃ©terminer si celle-ci confirmait ou contredisait des faits Ã©tablis, mÃªme les modÃ¨les les plus performants n'atteignaient en moyenne qu'un taux d'exactitude de 75 %. Par ailleurs, alors que pour Ã©laborer des stratÃ©gies climatiques efficaces, les dÃ©cideurs doivent analyser la maniÃ¨re dont les politiques climatiques se complÃ¨tent, le rapport a rÃ©vÃ©lÃ© que le modÃ¨le Ã©tait Ã peine capable d'Ã©valuer l'efficacitÃ© de ces interactions, avec un taux d'exactitude de seulement 33,8 %, soit Ã peine supÃ©rieur aux 33,3 % attendus en cas de choix alÃ©atoire. Enfin, les modÃ¨les multilingues ne raisonnent pas d'abord pour ensuite traduire leurs rÃ©ponses. Leurs rÃ©ponses varient en fonction de la qualitÃ© et de la quantitÃ© des donnÃ©es sur lesquelles ils ont Ã©tÃ© entraÃ®nÃ©s dans chaque langue. Lors des tests, les performances dans 13 langues diffÃ©rentes ont rÃ©vÃ©lÃ© d'importantes disparitÃ©s : les modÃ¨les affichaient une exactitude moyenne de 88,6 % en anglais et des performances comparables en espagnol, en franÃ§ais et en portugais ; mais l'exactitude est tombÃ©e Ã 83,6 % en hindi et en bengali, Ã 81,4 % en ourdou et Ã 78,3 % en swahili. Comme le souligne le rapport, cet aspect est d'autant plus important que les locuteurs de ces langues vivent dans des rÃ©gions du monde qui subissent de maniÃ¨re disproportionnÃ©e les effets du changement climatique.

Le rapport "Can We Trust AI on Climate? Why it seeks to please the user, and fails on complex reasoning and multilingualism" a Ã©tÃ© commandÃ© par l'UNESCO dans le cadre de sa contribution Ã l'Initiative mondiale pour l'intÃ©gritÃ© l'information sur le changement climatique (GIIICC), lancÃ©e en collaboration avec l'ONU et le BrÃ©sil en novembre 2024. Il est le fruit d'une collaboration entre le secteur de la communication et de l'information de l'UNESCO et la Chaire UNESCO en intelligence artificielle de l'University College London (UCL), crÃ©Ã©e en 2024 dans le but de faire progresser la recherche, la formation et la coopÃ©ration internationale en matiÃ¨re d'IA centrÃ©e sur l'humain.Â Â

Rapport - UNESCO