COMMUNIQUÉ DE PRESSE NATIONAL – PARIS – 12 JUILLET 2022

Livraison du plus grand modèle de langue multilingue « open science » jamais entraîné

S'ils fournissent régulièrement des résultats fascinants, les grands modèles d'intelligence artificielle sont généralement des boîtes noires : on ne sait pas exactement comment ils calculent leurs réponses et de nombreux éléments ne sont pas rendus publics. Le projet BigScience, impliquant un millier de chercheurs et chercheuses dans une démarche de science participative et ouverte, change la donne avec « Bloom ». Il s'agit du plus gros modèle de langue multilingue entrainé de manière complètement ouverte et transparente. Ce type d'intelligence artificielle apprend simultanément un modèle de génération de textes et un modèle de représentation de textes en effectuant de manière répétitive une tâche élémentaire : prédire le prochain mot d'un texte dont on connait le début, à la manière de ce que font les claviers « intelligents ». En plus de gérer 46 langues, allant de l'anglais au basque, son caractère *open science* aidera les scientifiques de tous horizons à explorer le fonctionnement des modèles de langue pour les améliorer. Le projet BigScience, initié par l'entreprise Hugging Face, a été soutenu par le CNRS, GENCI[1] et le ministère de l'Enseignement supérieur et de la Recherche, ce qui a permis d'entraîner Bloom sur la machine « Jean Zay », un des plus puissants supercalculateurs d'Europe.

Les modèles de langue sont des intelligences artificielles dont les premières applications concernent les textes en langue naturelle : réponses à des questions, génération automatique de phrases, détection de « sentiments », résumé et simplification automatiques ou encore traduction automatique. Généralement conçus par des géants des nouvelles technologies, la plupart des modèles existants ont été entrainés seulement avec des textes écrits en anglais et selon des principes et méthodes difficiles à reproduire dans tous leurs détails. Il n'est par exemple pas possible de savoir, lorsqu'un modèle répond à une question, si la réponse est le fruit d'un calcul ou si la réponse figurait déjà dans ses bases de données d'apprentissage.

Le projet BigScience a été initié au printemps 2021 par la start-up franco-américaine en intelligence artificielle Hugging Face, pour remédier à ces problèmes en entraînant un nouveau modèle : Bloom. Il apprend à partir de grands corpus de textes, en utilisant un principe simple, qui consiste à prédire à compléter des phrases, mot après mot. Chaque prédiction du modèle est comparée avec le mot correct, ce qui permet d'ajuster les paramètres internes du modèle. Dans le cas de Bloom, l'apprentissage est réalisé en évaluant des milliers de milliards de mots, conduisant à un modèle qui contient 176 milliards de paramètres. Cet apprentissage a duré plusieurs mois, nécessitant des centaines de processeurs graphiques (GPU) tournant en parallèle, soit l'équivalent de 5 millions d'heures de calcul. Une telle puissance de calcul ne peut être obtenue que sur des supercalculateurs comme la machine Jean Zay.

Bloom se distingue des autres modèles de langue par le fait qu'il est entraîné simultanément en 46 langues, réparties sur des sources aussi variées que de la littérature, des articles scientifiques ou des dépêches sportives et incluant de nombreuses langues rarement prises en compte, en particulier une vingtaine de langues d'Afrique. Le corpus d'apprentissage contient même du code informatique ! L'ensemble équivaut à plusieurs millions de livres. Or, plus l'approche et les sources sont diverses, plus le modèle est capable de remplir des tâches différentes. Les données n'ont de plus pas été triées en

fonction de leur langue car, paradoxalement, Bloom apprend mieux ainsi. Agglomérer des contenus en des langues variées permet d'apprendre des modèles robustes et performants pour toutes les langues considérées, et conduit même souvent à des résultats meilleurs que des modèles monolingues. Autre particularité : l'architecture de Bloom, la liste des données utilisées et son journal d'apprentissage seront entièrement disponibles en *open science*, afin de faciliter la recherche sur les modèles de langue. Bloom est enfin librement diffusée avec une licence responsable, qui prohibe explicitement les usages malveillants du modèle.

« *La création du modèle Bloom et le succès de la collaboration de recherche BigScience montrent qu'une autre manière de créer, étudier et partager les innovations en IA est possible, rassemblant industriels, académiques et associations autour d'un projet international, multidisciplinaire et d'accès ouvert. Je suis ravi que Hugging Face ait pu trouver en France les soutiens nécessaires pour cette démarche inédite à l'échelle mondiale* », indique Thomas Wolf, co-fondateur et directeur scientifique de la start-up Hugging Face.

« *BigScience initie une première mondiale et ouvre la voie à d'autres percées scientifiques. Il a bénéficié des ressources du supercalculateur convergé Jean Zay, l'un des plus puissants d'Europe, mis en service en 2019 dans le sillage du plan AI for Humanity. Aujourd'hui, plus de 1000 projets de recherche mobilisent ses ressources. Déterminante dans ce succès, l'extension de Jean Zay déployée en début d'année est issue d'un travail conjoint entre le ministère de l'Enseignement supérieur et de la Recherche, le CNRS à travers l'Institut du développement et des ressources en informatique scientifique (Idris), et GENCI* », déclare Philippe Lavocat, président-directeur général de GENCI.

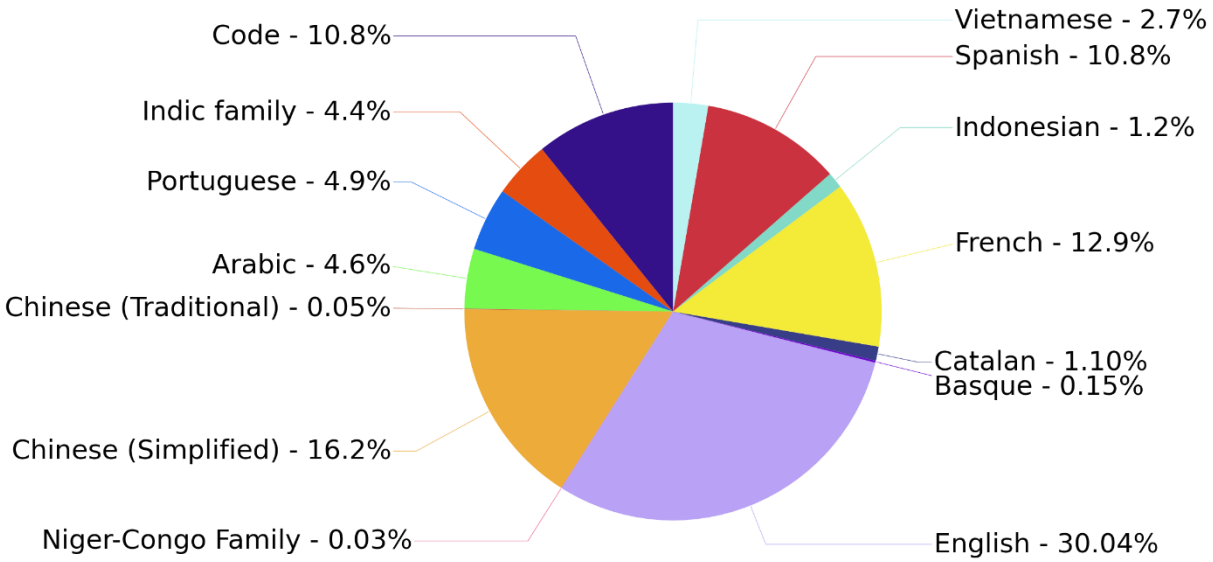
« *Nous nous réjouissons de ce partenariat public-privé original qui montre à quel point la complémentarité de compétences et de moyens—comme la puissance du supercalculateur Jean Zay—est essentielle pour relever un défi aussi important et actuel que la recherche en intelligence artificielle. Derrière l'avancée scientifique, nous saluons l'implication des personnels de l'Idris qui ont permis cet entraînement sur le supercalculateur, Et nous nous félicitons du rôle essentiel joué par le CNRS à travers la mobilisation de toute la communauté de traitement automatique des langues* », ajoute Antoine Petit, président-directeur général du CNRS.

« *Je suis heureux que ce projet international se situant sur l'une des frontières technologiques actuelles de l'IA ait été soutenu par la Stratégie nationale pour l'IA, et que le modèle Bloom soit prochainement accessible dans un cadre ouvert. Cela permettra à l'ensemble des acteurs innovants de développer de nouveaux cas d'usages et applications* », souligne Jean-Noël Barrot, ministre délégué au Numérique et aux Télécommunications.

« *Le consortium BigScience traduit une collaboration public-privé d'envergure mondiale dépassant le millier de contributeurs. Même si ces modèles nécessitent encore beaucoup d'investigations scientifiques et si leur impact énergétique nécessite une évaluation approfondie avant tout déploiement d'échelle, je suis fière que l'écosystème français en IA accueille un tel projet d'envergure internationale* », déclare Sylvie Retailleau, ministre de l'Enseignement supérieur et de la Recherche.

En savoir plus sur Bloom : huggingface.co/bigscience/bloom

A lire sur les sites du CNRS :
lejournal.cnrs.fr/articles/bigscience-voit-grand-pour-les-modeles-de-langue
www.cnrs.fr/fr/cnrsinfo/la-recherche-francaise-moteur-dun-nouveau-modele-dia



Langues utilisées pour l'entraînement de Bloom.

"Indic family" recouvre une quinzaine de langues du sous-continent indien (hindi, tamoul, ourdou, …) et "Niger-Congo family" une vingtaine de langues d'Afrique sub-saharienne (swahili, yoruba, wolof, …). 10,8 % des données étaient constituées de code informatique, avec 13 langages différents.

Source : Hugging Face

Notes

[1] Le CNRS a été impliqué en particulier via son Institut du développement et des ressources en informatique scientifique (Idris). GENCI : Grand équipement national de calcul intensif.

Contacts

Chercheur CNRS I François Yvon I francois.yvon@cnrs.fr
Chercheur CNRS I Pierre-François Lavallée I pierre-francois.lavallee@idris.fr
Presse CNRS I Véronique Etienne I **T +33 1 44 96 51 37** I veronique.etienne@cnrs.fr