PRESS RELEASE – PARIS – 12 JULY 2022

Release of largest trained open-science multilingual language model ever

Though they routinely yield fascinating results, the big AI models are generally black boxes: we do not know exactly how they arrive at their responses, and many of their details are not made public. With BLOOM, the BigScience project—which adopts an approach of open, participatory science involving a thousand researchers—is changing all of this. BLOOM is the largest multilingual language model to be trained 100% openly and transparently. AI models of its kind simultaneously learn how to generate and represent text by repeating the same basic task: prediction of the next word in a text whose beginning is known, much like predictive virtual keyboards. In addition to handling 46 human languages, from English to Basque, BLOOM—by virtue of its open-science ethos—lets scientists from all horizons freely explore how language models work, in order to improve them. BigScience, launched by the company Hugging Face, has received support from the CNRS, GENCI,[1] and the French Ministry of Higher Education and Research, allowing BLOOM to be trained on the Jean Zay supercomputer, one of Europe's most powerful.

Language models are AI systems whose primary applications concern natural (i.e., human) language. They might answer questions; generate sentences; detect emotions; or summarize, simplify, or translate text. Usually designed by giant tech firms, most existing models have solely been trained with English text and apply principles and methods that are difficult to fully replicate. For example, when one of these models replies to a question, it is impossible to know whether the answer given was arrived at through some algorithm or already in its training database.

BigScience was launched in the spring of 2021 by the Franco-American AI start-up Hugging Face, with the aim of addressing these issues by training a new model: BLOOM. BLOOM learns using large collections of texts, or corpora, by applying a simple principle: it first predicts what the next word in a sentence will be, and its prediction is then checked against the actual word. The model's parameters can then be adjusted on the basis of its performance. BLOOM has evaluated trillions of words, resulting in a model with 176 billion parameters. Equal to 5 million hours of computer processing, its training lasted several months and required hundreds of graphics processing units (GPUs) running side by side. Such computing power is only possible with supercomputers like the Jean Zay.

Unlike other language models, BLOOM has been trained simultaneously on 46 human languages—many rarely considered, including twenty African tongues—with textual sources as diverse as literature, scientific articles, and sports news, in addition to thirteen programming languages! The total volume of text digested is equivalent to the content of several million books. As the diversity of the model's approach and sources increases, the range of tasks BLOOM can accomplish widens. The input was not sorted by language because, surprisingly, BLOOM learns better this way. Combining content in various languages makes it possible to train powerful and robust models for all of those languages, and often yields better results than monolingual models. Another element that sets BLOOM apart from the pack is its open-science credo: its architecture, catalogue of data used, and training log are all publicly available, to

facilitate research into language models. Last but not least, BLOOM is freely distributed under its Responsible AI Licence explicitly prohibiting use for malicious purposes.

'The creation of the BLOOM model and the success of the BigScience research collaboration demonstrate that another way of creating, studying, and sharing AI innovations is possible, bringing together industry, academia, and non-profits around an international, multidisciplinary, open-access project,' said Thomas Wolf—co-founder and chief science officer of Hugging Face—adding that 'I am thrilled that Hugging Face was able to find the support it needed in France to pursue a novel approach of global scale.'

'BigScience has introduced a world first, clearing the way for other scientific breakthroughs,' declared GENCI CEO Philippe Lavocat. 'It has benefited from access to the resources of the Jean Zay converged supercomputer, one of the most powerful in Europe, commissioned in 2019 shortly after adoption of the AI for Humanity plan. Today over a thousand research projects are mobilizing its resources. A crucial factor in this success is the extension of the Jean Zay deployed at the start of the year, which is the fruit of the joint efforts of the French Ministry of Higher Education and Research; the CNRS, via the Institute for Development and Resources in Intensive Scientific Computing (IDRIS); and GENCI.'

Antoine Petit, Chairman and CEO of the CNRS added: 'We are delighted with this unique public-private partnership that illustrates how the complementarity of skills and means, such as the power of the Jean Zay supercomputer, is essential to meeting a challenge as important and current as AI research. Behind this scientific advance lies the commitment, which we salute, of the IDRIS staff members that permitted training on the supercomputer. And we celebrate the central role played by the CNRS through mobilization of the entire natural language processing community.'

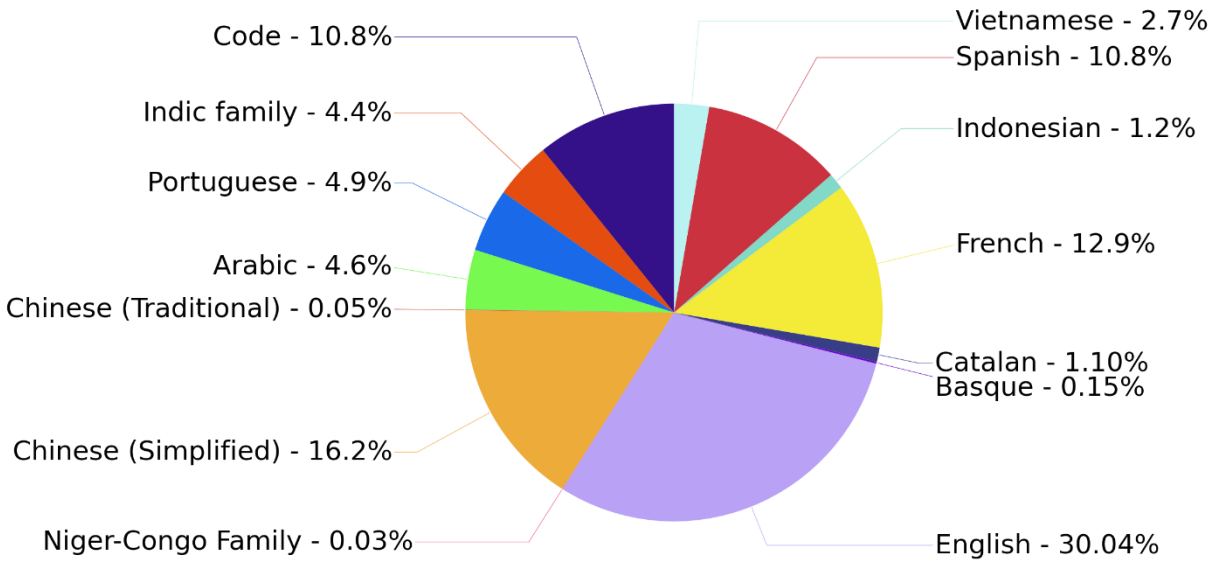
'I am pleased that this international project lying at one of the technological frontiers of AI was supported through the National Strategy for Artificial Intelligence [SNIA] and that the BLOOM model will soon be open-access,' stated Jean-Noël Barrot, French Minister Delegate for the Digital Transition and Telecommunications. 'That will allow all players in innovation to develop new use cases and applications.'

'The BigScience consortium represents a public-private collaboration of worldwide scope involving over a thousand contributors', explained Sylvie Retailleau, French Minister of Higher Education and Research. 'Although these models still demand much more scientific investigation and their energy consumption requires a thorough assessment before any scale-up, I am proud that the French AI ecosystem is hosting an international project such as this.'

To find out more: huggingface.co/bigscience/bloom
news.cnrs.fr/articles/bigscience-a-new-language-model-for-all

Notes

[1] The CNRS was involved in particular through its Institute for Development and Resources in Intensive Scientific Computing (IDRIS). GENCI (*Grand équipement national de calcul intensif*) is responsible for promoting high-performance computing in France.



Languages used to train BLOOM

BLOOM has been trained on thirteen languages in the Indic family from the Indian subcontinent (e.g., Hindi, Tamil, and Urdu) and twenty sub-Saharan languages in the Niger-Congo family (e.g., Swahili, Yoruba, and Wolof). Code in thirteen different programming languages accounted for 10.8% of its input.

Source: Hugging Face

Contacts

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