Biais Humain, Biais Machine: les LLMs génératifs sont-ils WEIRD?

Étienne Ollion

L'irruption de chatGPT fin 2022 a donné une visibilité inédite aux Large Language Models (LLMs) génératifs. Depuis, le recours à ces modèles a suscité nombre d'interrogations. L'une d'entre elles a trait aux opinions implicites et aux biais que pourrait avoir ces modèles. Entraînés principalement sur des corpus de textes issus de certains pays et de certains groupes sociaux, ils risqueraient de n'être représentatifs que d'une partie de la population. Plus précisément encore, ils auraient de forte chance d'être WEIRD (Western, Educated, Industrialized, Rich and Democratic). L'interrogation, importante, est devenue essentielle à mesure que ces modèles étaient employés pour remplacer des humains – dans des enquêtes d'opinions, dans des expériences, pour de la connaissance client. Cette présentation présente les résultats d'une enquête où on a sondé des LLMs sur une batterie de questions, avant de les comparer à celles de différentes populations. Elle explore les biais humains, et ce qu'on appelle les biais machine des LLMs, et on propose de réfléchir aux implications de ce résultat.