Partagez vos données avec OpenAI pour nourrir l'intelligence artificielle du futur

OpenAI Data Partnerships

OpenAI Data Partnerships

Open AI a besoin de données plus nombreuses et plus qualitatives pour améliorer ses modèles de langages (LLMs) et son populaire chatbot, qui aurait atteint plus de 100 millions d'utilisateurs actifs par semaine. Le chouchou de Microsoft, conscient que les entraîner sur ce qui est accessible à tous sur Internet est insuffisant, en appelle aux organisations pour constituer et partager avec lui des grands ensembles de données.

Pas de rémunération évoquée, OpenAI se contente d'arguer que l'amélioration de ses modèles d'IA "*profiteront à toute l'humanité*". C'est l'initiative Open AI Data Partnerships, dévoilée le 9 novembre 2023, par laquelle OpenAI compte progressivement supprimer les lacunes et les biais que comprennent ses LLMs, améliorer la prise en charge des requêtes vocales de son chatbot ainsi que ses capacités conversationnelles.

Combler ses lacunes sur les sujets et dans les langues sous-représentées

OpenAI recherche un type de donnée en particulier : des informations privées et publiques “*à grande échelle*” “*sur n’importe quel sujet et dans n’importe quelle langue*” (en particulier ceux les moins représentés), “*qui ne sont pas encore facilement accessibles au public en ligne*“.

“*Nous aimerions que les modèles d’IA comprennent en profondeur tous les sujets, industries, cultures et langues, ce qui nécessite un ensemble de données de formation aussi large que possible*“, a expliqué la société.

Améliorer les capacités conversationnelles de son chatbot

Elle dit par ailleurs préférer “*de longs écrits et des conversations plutôt que des extraits déconnectés*” qui donneront à ses modèles d’IA davantage d’indications sur la façon dont les humains communiquent. Cela leur permettra de produire des réponses plus complexes et nuancées et d’adopter un style plus conversationnel.

La société précise que les données ne doivent pas nécessairement être quantitatives et qu’elle accepte les formats textuels mais aussi les images, les audios et les vidéos.

Un ensemble de données open source, un autre privé

OpenAI dit avoir déjà collaboré sur ce principe avec le gouvernement islandais pour améliorer la capacité de GPT-4 à

parler islandais, ainsi qu'avec Free Law Project, une organisation à but non lucratif qui vise à démocratiser l'accès à la compréhension du droit en fournissant un accès gratuit et ouvert à des données juridiques (lois, jurisprudences…).

La manière dont OpenAI gèrera cette nouvelle masse de données sera certainement scrutée. Le programme prévoit deux types d'ensembles de données : un en open source qui sera accessible au public pour l'entraînement des modèles d'IA, et un privé et confidentiel destiné aux données sensibles et à l'entraînement de modèles d'IA propriétaires, éventuellement affinés pour des domaines spécifiques.