AFP Infos Françaises
mardi 4 juin 2024 - 23:01:02 UTC 733 mots

GPT-4, Bard etc: les grands modèles de langage ne sont pas raisonnables

Paris, 4 juin 2024 (AFP) - - Les grands modèles de langage (LLM), comme GPT-4, produits phare de l'intelligence artificielle, ont beaucoup de mal à raisonner face à des tests de logique: ils se trompent de façon inconstante et s'appuient sur des raisonnements souvent absurdes.

Les LLM (de l'anglais Large language model) sont réputés refléter les biais de genre, d'éthiques et moraux des humains dans les textes dont ils se nourrissent, rappelle une étude parue mercredi dans la revue Open Science de la Royal Society britannique.

Reflètent-ils aussi les biais cognitifs des humains dans des tests de raisonnement? s'est interrogé Olivia Macmillan-Scott, doctorante au département des sciences informatiques de University College of London (UCL).

Résultat des travaux: ces LLM font preuve "souvent de raisonnements irrationnels, mais d'une façon différente de celle des humains", résume-t-elle auprès de l'AFP

Sous la direction de Mirco Musolesi, professeur et directeur du Machine Intelligence Lab à l'UCL, elle a soumis sept LLM --deux GPT (3.5 et 4) d'OpenAI, Bard de Google, Claude 2 d'Anthropic et trois versions du Llama de Meta-- à une batterie de tests psychologiques administrés aux humains.

Comment se débrouilleraient-ils par exemple avec le biais de "négligence du dénominateur", qui désigne le fait de privilégier la solution avec le plus grand nombre d'éléments, plutôt que celle avec la juste proportion ?

Une urne contient neuf billes blanches et une rouge. L'autre urne en contient respectivement 92 et 8. Quelle urne choisir pour avoir le plus de chances de piocher une bille rouge? Contre-intuitive, la bonne réponse est la première urne, car elle fournit 10% de chance contre seulement 8% pour la deuxième.

- Refus de répondre -

Les réponses des LLM ont été classées selon qu'elles étaient correctes ou pas, et sur la validité du raisonnement logique accompagnant ces réponses.

Premier résultat de l'expérience, répétée dix fois pour chaque test, l'inconstance des réponses. Certains ont répondu correctement par exemple six fois sur dix à un même test ou seulement deux fois.

"On obtient à chaque fois une réponse différente", selon la chercheuse. Or la fiabilité sera "particulièrement importante si on doit les utiliser dans le monde réel", fait-elle valoir. Les LLM "peuvent se révéler très bons pour résoudre une équation mathématique compliquée... avant de sortir que 7 plus 3 font 12", décrit-elle.

Le plus surprenant? Les réponses étaient souvent décorrélées du raisonnement les soutenant, basé sur la logique et les probabilités. Dans le test des urnes par exemple, Claude 2 a fourni la bonne réponse une fois sur deux, mais à chaque fois avec un raisonnement en apparence logique, similaire à celui d'un humain.

Plus étonnant, des LLM ont refusé de répondre à un test, comme ce fut le cas de "Llama 2 70b", au motif que l'énoncé contenait des "stéréotypes de genre nocifs".

- "Je ne suis pas sûr" -

Les "modèles n'échouent pas dans ces tâches de la même façon qu'un humain échoue", remarque l'étude. Ce que le Pr. Mosulesi résume en parlant d'"erreurs machines: il y a une forme de raisonnement logique qui est potentiellement correct si on le prend par étapes, mais qui est faux pris dans son ensemble".

La machine fonctionne avec "une sorte de pensée linéaire". Comme dans cet exercice où Bard effectue correctement une certaine tâche à une étape, puis une autre à la suivante, avant de de ne retenir que le résultat de cette dernière étape - bref, sans vue d'ensemble.

Interrogé sur le sujet, Maxime Amblard, professeur en informatique à l'Université de Lorraine, rappelle que "les LLM, comme toutes les intelligences artificielles génératives, ne fonctionnent pas comme des humains". Ces derniers sont "des machines à faire du sens", qui échappe aux machines, a-t-il dit à l'AFP.

Comme les humains aussi, les machines du test ne sont pas toutes égales. Dans l'ensemble GPT-4, sans être infaillible, a mieux réussi que les autres la batterie de tests.

Olivia Macmillan-Scott soupçonne que de tels modèles dits "fermés", c'est-à-dire dont le code de fonctionnement reste secret, "incorporent d'autres mécanismes en tâche de fond" pour répondre à des questions mathématiques.

Impensable pour autant de confier à ce stade une décision à un LLM. Mais pourquoi ne pas utiliser leur drôle de façon de penser comme un outil d'aide à la réflexion ? Autre piste, selon le Pr. Mosulesi, les entraîner à répondre, quand c'est le cas: "je ne suis pas trop sûr".