Anesthesie - Réanimation Divers

ID: 550

Évaluation des performances de GPT-3 et -4 à un examen de Diplôme d'Université

S. Sigaut*(1), S.Rozencwajg(2), E.Kantor(2), E.Weiss(3)

(1) Département d'Anesthésie Réanimation, Hôpital Beaujon AP-HP, Clichy, France, (2) Département d'Anesthésie Réanimation, Hôpital Bichat, Paris, France, (3) Département d'Anesthésie Réanimation, Hôpital Beaujon, Clichy, France

**Auteur présenté comme orateur*

Position du problème et objectif(s) de l'étude:
Les modèles de langage tels que GPT sont en train de transformer l'enseignement médical en soulignant ses imperfections, notamment dans la prise en compte de l'incertitude [1]. Mais leur fiabilité dans le champ d'une spécialité médicale reste à définir. Les objectifs de ce travail sont d'évaluer les performances de deux modèles de langage sur un examen de formation continue en
anesthésie-réanimation et d'identifier leur potentiel pour aider les enseignants et les étudiants.

Matériel et méthodes:
Nous avons soumis à GPT-3 et -4 les 29 questions à réponses multiples (QRM) de l'examen 2022 du Diplôme Universitaire de réanimation périopératoire en chirurgie digestive de l'Université de Paris Cité. Chaque QRM est composée de 5 propositions. Le même texte d'entrée a été utilisé pour demander aux modèles 1) de lister les propositions justes et fausses, 2) de fournir un argumentaire pour justifier chaque réponse. La même méthode de notation que pour les étudiants a été appliquée. Leurs performances selon le type de question et les types d'erreurs commises ont été analysées. Nous avons classées les erreurs en «conception alternative» (réponse et argumentaire faux); «omission» (absence de réponse et d'argumentaire); «manque de contextualisation» (réponse vraie dans certains cas mais pas dans le contexte de l'énoncé); «inattention» (argumentaire correct mais liste des réponses juste et fausse erronée) et «QRM ambiguë» (autre interprétation que celle voulue par le correcteur possible).

Résultats & Discussion:
GPT-3 et -4 obtiennent la même note finale de 7,17/10, validant l'examen. Ils se classent 23ème/35 avec une note inférieure à la médiane des étudiants qui est de 7,73. GPT-3 et -4 répondent pour la majorité des questions moins bien que la médiane des étudiants, aussi bien pour les cas cliniques que pour les questions de connaissances théoriques (figure 1).
La typologie des 41 erreurs de chacun des deux modèles sont présentées dans la figure 1.

Conclusion:
Les modèles de langage GPT sont capables de réussir un examen de niveau formation continue en anesthésie-réanimation. Ils permettent d'identifier des questions ambiguës. De façon surprenante, ils peuvent commettre des fautes d'inattention. Le manque de connaissance se traduit par de nombreusesconceptions alternatives avec GPT-3, remplacées en partie par des omissions dans GPT-4. Nous souhaitons donc attirer l'attention sur ce phénomène connu aussi sous le nom d'hallucination d'IA, qui se produit lorsque les modèles sont entraînés avec des données insuffisantes ou inadéquates. Elles peuvent être prises comme des vérités et induire les étudiants en erreur s'ils utilisent ces outils pour leurs apprentissages. Ici elles sont particulièrement fréquentes avec GPT-3, l'outil le plus facilement accessible à ce jour, et si elles semblent avoir été régressées avec GPT-4, elles ne sont pas totalement exclues.

Références bibliographiques:
1 - Mbakwe AB, et.al, ChatGPT passing USMLE shines a spotlight on the flaws of medical education. Hochheiser H, editor. PLOS Digit Health. 2023;2:e0000205

	GPT-3	GPT-4
Caractéristiques du modèle		
Année de mise en service	2020	2023
Nombre de data ayant servi à l'entrainement du modèle	145 milliards	100000 milliards
Interface utilisée pour ce travail	Open AI	Bing
Performances selon le type de question		
Nombre de cas cliniques avec un score ≥ à la médiane des étudiants (n=9)	2 (22%)	4 (44%)
Nombre de questions de connaissances théoriques avec un score ≥ à la médiane des étudiants (n=20)	9 (45%)	8 (40%)
Erreurs liées au modèle de langage		
Conception alternative	25 (63 %)	10 (24 %)
Omission	3 (7 %)	16 (39 %)
Manque de contextualisation	2 (5 %)	7 (17 %)
Inattention	5 (12 %)	4 (10 %)
Erreurs liées à l'énoncé		
Ambiguïté	5 (12 %)	4 (10 %)

Les auteurs déclarent ne pas avoir toute relation financière impliquant l'auteur ou ses proches(salaires, honoraires, soutien financier éducationnel) et susceptible d'affecter l'impartialité de la présentation.