Anesthesie - Réanimation Divers

ID: 523

Utilisation de GPT-4 pour l'amélioration des examens des DESAR

S. Rozencwajg*(1), E.Kantor(1), S.Sigaut(2), E.Weiss(2), A.Bouglé(3), M.Raux(3)

(1) Département d'Anesthésie-Réanimation, CHU Bichat Claude-Bernard, Paris, France , (2) Service d'Anesthésie-Réanimation, CHU Beaujon, Clichy, France , (3) Département d'Anesthésie-Réanimation, CHU Pitié-Salpêtrière, Paris, France

**Auteur présenté comme orateur*

Position du problème et objectif(s) de l'étude:
Les questions à réponses multiples (QRM) sont couramment utilisées pour les examens des internes en anesthésie-réanimation (DESAR), mais leur pertinence pour comprendre les erreurs des étudiants est limitée. Nous avons fait l'hypothèse que GPT-4 (Generative Pre-trained Transformer 4), un modèle de langage développé par openAI, pourrait aider à mieux identifier et classifier les erreurs des étudiants, afin d'améliorer l'évaluation et l'enseignement.

Matériel et méthodes:
GPT-4 a été soumis aux 50 QRM de l'examen 2022 du DESAR phase socle. Nous avons demandé à GPT-4 1) de répondre aux QRM et 2) de donner un niveau de confiance pour chacune de ses réponses. Une foisles réponses obtenues, la même méthode de notation que pour les étudiants a été appliquée et GPT-4 aété classé parmi eux. Ensuite, nous avons comparé les erreurs de GPT-4 et des étudiants puis analysé l'argumentaire de GPT-4, permettant de classer les erreurs en quatre catégories : «manque de connaissances» (absence de réponse et d'argumentaire), «erreur de raisonnement» (réponse fausse et argumentaire faux ou réponse vraie dans certains cas mais pas dans ce contexte), «inattention» (réponse fausse malgré un argumentaire vrai) et «ambiguité dans l'énoncé» (autre interprétation que celle du correcteur). Nous avons classé les erreurs des étudiants dans ces mêmes catégories selon la méthode DELPHI. Enfin, les erreurs de GPT-4 ont été classées selon les thématiques principales des questions.

Résultats & Discussion:
GPT-4 obtient la note finale de 13,44/20, validant l'examen et se classant 9ème parmi les 99 étudiants. GPT-4 a répondu aux QRM en 1 minute contre 72 [60-78] minutes pour les étudiants, avec un niveau de confiance médian de 80 [80-90] %. Il a commis 26 erreurs dans les domaines de pharmacologie (n=10), physiologie (n=4), neurologie (n=4), hémostase (n=3) et réanimation (n=1) avec une estimation correcte de ses erreurs dans 46% des cas. La répartition des erreurs était la suivante : erreur de raisonnement (n=17), manque de connaissances (n=16) et inattention (n=2). GPT-4 a pu faire suspecter une ambiguïté dans l'énoncé pour deux questions. Pour les étudiants, leurs erreurs ont été classées comme suit : manque de connaissances (n=26), erreur de raisonnement (n=2), inattention (n=2). Nous avons suspecté une ambiguïté dans l'énoncé pour 7 propositions, dont les deux que GPT-4 classait comme "ambiguës". Ainsi, cinq "ambiguïtés" pour les auteurs n'ont pas été classées comme telles par GPT-4.

Conclusion:
L'utilisation de GPT-4 lors d'un examen de 3ème cycle de DESAR a confirmé ses bonnes performances et révélé qu'il n'est pas un modèle exempt d'erreurs et ne devrait donc être utilisé seul pour la conception de QCM. Ces erreurs sont principalement dues à des défauts de raisonnement (notamment en matière de contextualisation) plutôt qu'à un manque de connaissances comme celles des étudiants. GPT-4 est utile pour confirmer qu'un étudiant commet une erreur par manque de connaissance. De manière intéressante, il pourrait être un d'outil de validation en cas de suspicion d'une ambiguïté d'énoncé, contribuant ainsi à améliorer la qualité des examens dans un souci d'impartialité et d'égalité des chances.

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.