Canadian Medical Education Journal

You Should Try This!

AI that teaches: an evidence-based GPT model to improve medical student understanding of pulmonary function tests

Une IA pédagogique : un modèle GPT fondé sur des données probantes pour améliorer la compréhension des tests de fonction pulmonaire par les étudiants en médecine

Anusha Aiyar,[1] Henry Moon[1]

[1]Medical College of Georgia, Augusta University, Georgia, USA

Correspondence to: Anusha Aiyar, BS; email: aaiyar@augusta.edu

Published ahead of issue: Dec 9, 2025. CMEJ 2025 Available at https://doi.org/10.36834/cmej.80873

Implication Statement

This study explores the integration of an augmented Generative Pre-trained Transformer (GPT) tool with curated scientific sources to enhance the learning of pulmonary function test (PFT) interpretation in pre-clerkship medical education. Our findings suggest that this approach offers notable improvements in accuracy, reliability, and the quality of explanations compared to existing tools, such as Out-of-Box GPT and USMLE Q-Banks. The PFT learning assistant can support medical students in navigating common learning barriers, provide a personalized and scalable approach to evidence-based medical education

Énoncé des implications de la recherche

Cette étude explore l'intégration d'un outil de transformation générative pré-entraînée (GPT) enrichi de ressources scientifiques sélectionnées afin d'améliorer l'apprentissage de l'interprétation des épreuves fonctionnelles respiratoires (EFR) dans la formation médicale préclinique. Nos résultats suggèrent que cette approche offre des améliorations notables en termes de précision, de fiabilité et de qualité des explications par rapport aux outils existants, tels que le GPT standard et les banques de questions USMLE. Cet assistant d'apprentissage des EFR peut aider les étudiants en médecine à surmonter les obstacles d'apprentissage courants et propose une approche personnalisée et adaptable de la formation médicale fondée sur les preuves.

Introduction

Pulmonary Function Tests (PFTs) are integral to the diagnosis and treatment of respiratory diseases.[1] However, medical students often struggle with this topic in their pre-clinical years because its complex anatomical and physiological pathways require more active learning than traditional large-group instruction typically supports.[2,3] Moreover, the self-study resources students rely on, such as the United States Medical Licensing Examination (USMLE) question bank materials, provide surface-level explanations that lack adaptive reasoning and have limitations in the quality of explanation provided.[4]

To assist medical students in better understanding complex PLFT interpretation in a pre-clerkship medical curriculum, we developed and evaluated a customized GPT model that integrates the Retrieval Augmented Generation (RAG) framework within OpenAI GPT[5] and PFT-related scientific articles found in PubMed, Medline, EMBASE, and WILEY. Generic out-of-box GPTs generate responses based solely on their pretrained knowledge and no curated resources, whereas our RAG-enabled GPT enhances general-purpose models by integrating relevant external information to provide evidence-based and up-to-date responses.[6]

Description of the innovation

To address the persistent challenges medical students face in learning complex PFTs, we developed an AI-augmented educational tool. This tool was specifically designed to function as a learning assistant in interpreting PFTs by drawing on curated content from peer-reviewed scientific evidence. This innovation stems from the limitations of conventional self-study resources and fails to adapt to personalized feedback.[2,3] Moreover, out-of-box GPT models rely on pretrained knowledge and lack the ability to access and integrate domain-specific evidence. Our PFT learning assistant bridges this gap by accessing relevant scientific literature to provide evidence-based and up-to-date responses.[6] To evaluate its effectiveness, the PFT learning assistant model was examined on 13 USMLE-level PFT-related questions and compared to both the answers and feedback from an out-of-box GPT model and the USMLE question bank. Two researchers reviewed and rated the models' outputs across four domains: accuracy, reliability, similarity, and quality of feedback. Accuracy was determined by comparing the first output from AI-generated responses to the USMLE answer keys. Reliability was tested using each model to generate 100 outputs without learning effects. We also evaluated the quality of feedback across the three tools using a 7-point scale rubric with seven criteria: specificity, constructiveness, relevance, clarity, tone, empathy, and diversity. To analyze the results, we conducted paired t-tests for accuracy, ICC for reliability, Cohen's Kappa for similarity, and one-way ANOVA for feedback quality. This study was considered exempt from review by the Augusta University IRB.

Outcomes

The results showed that the PFT learning assistant significantly outperformed the out-of-box GPT in terms of accuracy ($t(12) = 2.45$, $p < 0.03$) with a mean score (M = 0.92 vs. 0.69) and reliability (ICC = 0.89 vs. 0.74). While both models often arrived at the same answers, the PFT learning assistant was more consistently correct, leading to only a fair similarity in response patterns (k = 0.32). Also, in terms of the depth of explanation provided, the two models diverged in some instances, possibly due to their evidence-based augmentation and variations in contextual interpretation. The one-way ANOVA also revealed significant differences in the quality of feedback ($F(2,36) = 9.22$, $p < 0.01$). Post-hoc Tukey's tests indicated that the PFT learning assistant produced significantly higher-quality explanations (M = 4.7, SD = 0.3) compared to both the out-of-box GPT (M = 4.1, SD = 0.5) and USMLE Q-Banks (M = 3.8, SD = 0.6), particularly outperforming in specificity, relevance, and clarity (See Table 1). The results support that the PFT learning assistant was significantly more accurate and consistent than the out-of-the-box GPT in interpreting complex PFT questions and contribute to a richer, personalized learning experience.

Suggestions for next steps

The findings suggest that incorporating evidence-based content has the potential to enhance the ability of generic GPT tools to support preclinical students in understanding complex medical concepts. Our augmented GPT tool can complement existing educational resources and provide scalable learning experiences for large cohorts. By providing tailored explanations, the PFT learning assistant can support individual learning challenges. However, it is crucial to be mindful of the model's dependency on external sources and potential challenges related to technological accessibility. In addition, this study is also limited by the small number of questions tested, leading to potential biases in the evaluation results. Despite these limitations, the PFT learning assistant provides significant advantages in terms of accuracy, reliability, and quality for medical education. Future research should include outcome evaluations to examine the effectiveness of this tool in supporting medical students' learning. Additionally, studies should consider learner perceptions and contextual factors that may influence their educational value and potential for integration into other education settings.

Conflicts of Interest: Henry Moon is an editor for the CMEJ. He has adhered to the CMEJ policy on editors as authors.
Funding: The authors declare that they have no conflict of interest and no external funding to support this study.
Authorship: All authors have each signed their own attestation statement that they meet the requirements of authors. The lead author, in addition to their own attestation as an author, has signed an attestation that all authors listed on this paper meet the requirements for authors.
Edited by: Anita Acai (senior section editor); Marcel D'Eon (editor-in-chief)

Table 1. Comparison of the explanation quality provided by the PFT learning assistant and out-of-box GPT models

	Out-of-Box GPT	**PFT Learning Assistant**
Specificity	"Diffusion of oxygen across the respiratory membrane directly impacts the DLCO. Conditions that damage the alveolar-capillary membrane, such as emphysema (common in long-term smokers), would impair the diffusion of gases, leading to a decreased DLCO."	"DLCO measures how well gases like oxygen transfer from the lungs into the blood, and it depends on factors such as the surface area of the alveolar-capillary membrane and the thickness of this membrane. In chronic smokers, conditions like emphysema can significantly reduce the surface area available for gas exchange due to the destruction of alveoli, leading to a lower DLCO. This is primarily because the structural damage and reduced alveolar surface area directly impair the diffusion process. Other factors listed, such as the ability of hemoglobin to bind oxygen or hemoglobin content in the blood, do affect oxygen transport but are less directly related to the changes typically seen in DLCO for a patient with a heavy smoking history and potential emphysematous changes"
Relevance	"...This patient's symptoms and clinical findings are suggestive of chronic obstructive pulmonary disease (COPD), particularly emphysema. Emphysema is characterized by the destruction of the alveolar walls and the enlargement of airspaces, leading to decreased elastic recoil of the lungs and air trapping. The FEV1:FVC ratio of 45% confirms the presence of an obstructive lung disease, as this ratio is significantly lower than the normal value. The primary cells responsible for the development of emphysema are neutrophils which release proteases, such as elastase, which break down elastin and other components of the extracellular matrix in the alveolar walls. This destruction leads to the loss of alveolar structure and the characteristic changes seen in emphysema..."	"...This patient's presentation is suggestive of chronic obstructive pulmonary disease (COPD), most likely emphysema, given the long smoking history, progressive dyspnea, and the described spirometry and chest X-ray findings of widened intercostal spaces and hyperlucent lung fields. Emphysema is characterized by destruction of the alveolar walls, leading to enlarged air spaces and loss of elastic recoil. This process results in airflow limitation that is not fully reversible. The FEV1/FVC ratio of 45% confirms an obstructive pattern, which is a hallmark of COPD. The primary cell type involved in the pathogenesis of emphysema in smokers is the neutrophil. Neutrophils release proteases such as elastase, which break down the elastic fibers in the alveolar walls. In smokers, the balance between protease activity and antiprotease defense (e.g., alpha-1 antitrypsin) is disrupted, leading to alveolar destruction..."
Clarity	"The key points in this case are the boy's intermittent shortness of breath and dry cough, worse at night, following a recent upper respiratory tract infection. The lungs are clear on auscultation, and the spirometry is initially normal. The significant finding is that a drug administration results in a reduced peak expiratory flow rate. This suggests the drug induced bronchoconstriction or increased airway hyperresponsiveness, which points towards a provocative agent."	"Methacholine is a cholinergic drug used in the methacholine challenge test to diagnose bronchial hyperreactivity, which is characteristic of asthma. Methacholine causes bronchoconstriction, leading to a decrease in peak expiratory flow rate (PEFR). In individuals with asthma or airway hyperreactivity, administration of methacholine results in a significant reduction in airflow, evidenced by a drop in PEFR on spirometry. Here is the reasoning for the other options: -Atenolol: This is a beta-blocker that could theoretically cause bronchoconstriction, especially in asthmatic patients, but it is not typically used in diagnostic testing for asthma. -Ipratropium bromide: This is an anticholinergic bronchodilator, which would likely increase PEFR by reducing bronchoconstriction, not decrease it. -Albuterol: This is a beta-agonist bronchodilator used to relieve bronchospasm, and it would increase PEFR. -Methoxyflurane: This is an anesthetic agent and not used in the context of asthma diagnosis. - Epinephrine: This is a sympathomimetic drug that would cause bronchodilation and increase PEFR. "

References

1. Smith JA, Doe RL. The impact of AI on personalized learning. *J Educ Technol.* 2020;35(4):123-145. https://doi.org/10.1016/j.edutech.2020.05.012
2. Ranu H, Wilde M, Madden B. Pulmonary function tests. *Ulster Med J*. 2011;80(2):84-90. Available from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3229853/ [Accessed on Jan 23, 2025].
3. West JB. Challenges in teaching the mechanics of breathing to medical and graduate students. *Adv Physiol Educ.* 2008;32(3):177-184. https://doi.org/10.1152/advan.90146.2008
4. Kann MR, Huang GW, Pugazenthi S, et al. Unlocking medical student success: a systematic review and meta-analysis of third-party resources used for medical education and USMLE board preparation. *Med Sci Educ.* 2024. https://doi.org/10.1007/s40670-024-02116-7
5. OpenAI. *ChatGPT.* OpenAI; 2024. Available from: https://openai.com [Accessed on Jan 23, 2025].
6. Yang R, Ning Y, Keppo E, et al. Retrieval-augmented generation for generative artificial intelligence in health care. *NPJ Health Syst*. 2025;2(2). https://doi.org/10.1038/s44401-024-00004-1