ORIGINAL

How does ChatGPT perform on the European Board of Pediatric Surgery examination? A randomized comparative study

¿Cuál es el rendimiento de ChatGPT en el examen del Consejo Europeo de Cirugía Pediátrica? Un estudio comparativo aleatorizado

Mustafa Azizoğlu[1], Bahattin Aydoğdu[2]

1. Dicle University Medical School Department of Pediatric Surgery, Diyarbakır, Turkey.
2. Balıkesir University Department of Pediatric Surgery, Turkey.

Corresponding author
Mustafa Azizoğlu
E-mail: mdmazizoglu@gmail.com

Received: 25 - VIII - 2023
Accepted: 23 - IX - 2023

doi: 10.3306/AJHS.2024.39.01.23

Abstract

Purpose: The purpose of this study was to conduct a detailed comparison of the accuracy and responsiveness of GPT-3.5 and GPT-4 in the realm of pediatric surgery. Specifically, we sought to assess their ability to correctly answer a series of sample questions of European Board of Pediatric Surgery (EBPS) exam.

Methods: This study was conducted between 20 May 2023 and 30 May 2023. This study undertook a comparative analysis of two AI language models, GPT-3.5 and GPT-4, in the field of pediatric surgery, particularly in the context of EBPS exam sample questions. Two sets of 105 (total 210) sample questions each, derived from the EBPS sample questions, were collated.

Results: In General Pediatric Surgery, GPT-3.5 provided correct answers for 7 questions (46.7%), and GPT-4 had a higher accuracy with 13 correct responses (86.7%) (p=0.020). For Newborn Surgery and Pediatric Urology, GPT-3.5 correctly answered 6 questions (40.0%), and GPT-4, however, correctly answered 12 questions (80.0%) (p= 0.025). In total, GPT-3.5 correctly answered 46 questions out of 105 (43.8%), and GPT-4 showed significantly better performance, correctly answering 80 questions (76.2%) (p<0.001). Given the total responses, when GPT-4 was compared with GPT-3.5, the Odds Ratio was found to be 4.1. This suggests that GPT-4 was 4.1 times more likely to provide a correct answer to the pediatric surgery questions compared to GPT-3.5.

Conclusion: This comparative study concludes that GPT-4 significantly outperforms GPT-3.5 in responding to EBPS exam questions.

Key words: ChatGPT, Pediatric Surgery, exam, questions, artificial intelligence.

Resumen

Introducción: El propósito de este estudio fue realizar una comparación detallada de la precisión y capacidad de respuesta de GPT-3.5 y GPT-4 en el ámbito de la cirugía pediátrica. En concreto, pretendíamos evaluar su capacidad para responder correctamente a una serie de preguntas de muestra del examen del European Board of Pediatric Surgery (EBPS).

Métodos: Este estudio se llevó a cabo entre el 20 de mayo de 2023 y el 30 de mayo de 2023. Este estudio llevó a cabo un análisis comparativo de dos modelos de lenguaje de IA, GPT-3.5 y GPT-4, en el campo de la cirugía pediátrica, particularmente en el contexto de las preguntas de muestra del examen EBPS. Se cotejaron dos conjuntos de 105 (210 en total) preguntas de muestra cada uno, derivadas de las preguntas de muestra del EBPS.

Resultados: En Cirugía Pediátrica General, la GPT-3.5 proporcionó respuestas correctas para 7 preguntas (46,7%), y la GPT-4 tuvo una mayor precisión con 13 respuestas correctas (86,7%) (p=0,020). Para Cirugía neonatal y Urología pediátrica, la GPT-3.5 respondió correctamente a 6 preguntas (40,0%), y la GPT-4, sin embargo, respondió correctamente a 12 preguntas (80,0%) (p= 0,025). En total, la GPT-3.5 respondió correctamente a 46 preguntas de 105 (43,8%), y la GPT-4 mostró un rendimiento significativamente mejor, respondiendo correctamente a 80 preguntas (76,2%) (p<0,001). Teniendo en cuenta el total de respuestas, cuando se comparó la GPT-4 con la GPT-3.5, se observó que la Odds Ratio era de 4,1. Esto sugiere que la GPT-4 era 4,2 veces más eficaz que la GPT-3.5. Esto sugiere que GPT-4 tenía 4,1 veces más probabilidades de proporcionar una respuesta correcta a las preguntas de cirugía pediátrica en comparación con GPT-3.5.

Conclusiones: Este estudio comparativo concluye que GPT-4 supera significativamente a GPT-3.5 a la hora de responder a las preguntas del examen EBPS.

Palabras clave: ChatGPT, Cirugía Pediátrica, examen, preguntas, inteligencia artificial.

Cite as: U

Introduction

The Chat Generative Pre-trained Transformer (ChatGPT) is a natural language processing tool that was trained on massive amounts of data and is driven by artificial intelligence (1). Due to its extraordinary capacity to generate human-like responses in response to text input within a conversation, ChatGPT has been gaining significant attention ever since it was first made available to the public in November 2022. GPT-3.5 was the foundational large language model that was used to support ChatGPT when it first launched. In March of 2023, an improved version known as GPT-4 was made available to the public with the promise of increased precision. Although there is a lot of interest in the use of ChatGPT, there is some debate about whether or not it should be used in medical practice[1-3].

The advent and progression of Artificial Intelligence (AI) in various fields have redefined the way we understand and implement knowledge[4]. AI's integration into medicine, particularly the field of pediatric surgery, offers an innovative lens to examine, decode, and provide solutions to complex surgical problems. The AI language models GPT-3.5 and GPT-4, developed by OpenAI, have demonstrated promising applications in diverse fields, including medicine. Yet, a comprehensive understanding of their capacity to accurately respond to professional, field-specific queries remains to be thoroughly explored[5,6].

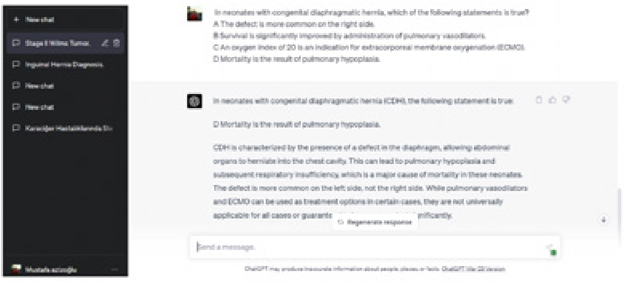
The purpose of this study was to conduct a detailed comparison of the accuracy and responsiveness of GPT-3.5 and GPT-4 in the realm of pediatric surgery. Specifically, we sought to assess their ability to correctly answer a series of sample questions of EBPS exams.

Materials and methods

This study was conducted between 20 May 2023 and 30 May 2023. This study undertook a comparative analysis of two AI language models, GPT-3.5 and GPT-4, in the field of pediatric surgery, particularly in the context of EBPS exam sample questions. Two sets of 105 (total 210) sample questions each, derived from the EBPS sample questions, were collated. These questions spanned a broad range of pediatric surgical knowledge and were structured in a variety of formats to best assess the capabilities of the AI models.

Both GPT-3.5 and GPT-4 models were set up for the study. The testing was conducted on a computer with an internet connection, and responses from the models were recorded on a digital platform for further analysis. Each model was independently presented with a set of 105 questions in the same sequence which are selected randomly (General Pediatric Surgery, Newborn Surgery, Thoracic Surgery, Pediatric Urology, Traumatology, Hepatopancreatobiliary Surgery, Pediatric Oncological Surgery; each one 15 questions per group). Each question was input individually, and the generated response was recorded (Sample; **Figure 1, 2, 3**). The responses from both AI models were carefully reviewed by experts in pediatric surgery. The responses were scored based on their correctness and relevance to the question.

Figure 1: Sample question in Newborn Surgery section.



In neonates with congenital diaphragmatic hernia, which of the following statements is true?
A The defect is more common on the right side.
B Survival is significantly improved by administration of pulmonary vasodilators.
C An oxygen index of 20 is an indication for extracorporeal membrane oxygenation (ECMO).
D Mortality is the result of pulmonary hypoplasia.

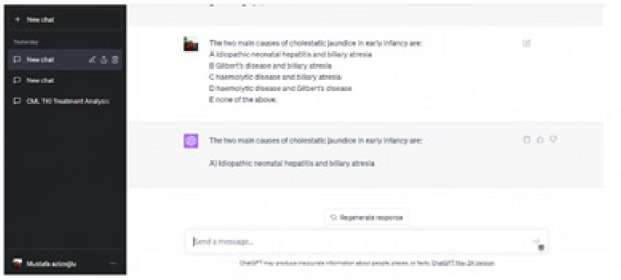
In neonates with congenital diaphragmatic hernia (CDH), the following statement is true:

D Mortality is the result of pulmonary hypoplasia.

CDH is characterized by the presence of a defect in the diaphragm, allowing abdominal organs to herniate into the chest cavity. This can lead to pulmonary hypoplasia and subsequent respiratory insufficiency, which is a major cause of mortality in these neonates. The defect is more common on the left side, not the right side. While pulmonary vasodilators and ECMO can be used as treatment options in certain cases, they are not universally applicable for all cases or guarante nificantly.

ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT May 24 Version

Figure 2: Sample question in Hepatopancreatobiliary Surgery section.



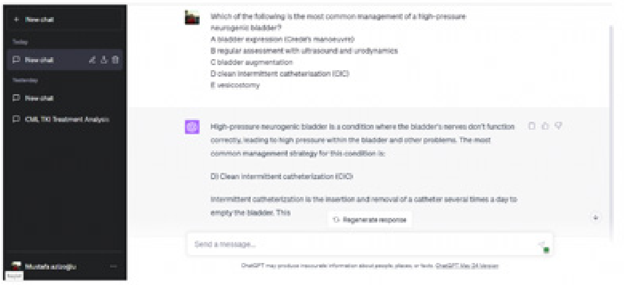
The two main causes of cholestatic jaundice in early infancy are:
A idiopathic neonatal hepatitis and biliary atresia
B Gilbert's disease and biliary atresia
C haemolytic disease and biliary atresia
D haemolytic disease and Gilbert's disease
E none of the above.

The two main causes of cholestatic jaundice in early infancy are:

A) Idiopathic neonatal hepatitis and biliary atresia

ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT May 24 Version

Figure 3: Sample question in Pediatric Urology section.



Which of the following is the most common management of a high-pressure neurogenic bladder?
A bladder expression (Credé's manoeuvre)
B regular assessment with ultrasound and urodynamics
C bladder augmentation
D clean intermittent catheterisation (CIC)
E vesicostomy

High-pressure neurogenic bladder is a condition where the bladder's nerves don't function correctly, leading to high pressure within the bladder and other problems. The most common management strategy for this condition is:

D) Clean intermittent catheterization (CIC)

Intermittent catheterization is the insertion and removal of a catheter several times a day to empty the bladder. This

ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT May 24 Version

Following the review and scoring process, the number of correct answers provided by each model was summed up. The total correct answers were then expressed as a percentage of the total questions, giving an accuracy score for each model. The accuracy scores of GPT-3.5 and GPT-4 were compared to determine which model demonstrated superior performance in answering the EBPS sample questions.

Inclusion criteria

Randomly selected questions from the EBPS study questions have been included in the study.

Exclusion criteria

Questions containing images were excluded from the study as the GPT program does not accept images.

Statistical analysis

For all items, descriptive statistics, frequency, and other characteristics were used in the statistical analysis of the data. The Chi-square test was used to evaluate the categorical variables, and when necessary, Fisher exact test was used on some of the data. SPSS Statistics for Windows, Version 21.0 (IBM Corp., Armonk, NY, USA) was used to conduct the analyses. P values were all two-sided, and p values below 0.05 were regarded as statistically significant.

Results

In this comparative study, the GPT-3.5 and GPT-4 language models were tested on a range of topics within pediatric surgery. Each topic was assessed for accuracy, with results indicating significant differences in performance between the two models.

A total of 210 questions were included in the study, 105 questions from each group. In General Pediatric Surgery, GPT-3.5 provided correct answers for 7 questions (46.7%) and incorrect answers for 8 questions (53.3%). In contrast, GPT-4 had a higher accuracy with 13 correct responses (86.7%) and only 2 incorrect responses (13.3%). This difference was found to be statistically significant (p= 0.020). For Newborn Surgery and Pediatric Urology, GPT-3.5 correctly answered 6 questions (40.0%) and incorrectly answered 9 questions (60.0%) for both topics. GPT-4, however, correctly answered 12 questions (80.0%) and incorrectly answered 3 questions (20.0%) for both categories, with the difference also being statistically significant (p= 0.025). In Thoracic Surgery, GPT-3.5 had a slightly higher accuracy, correctly answering 8 questions (53.3%) and incorrectly answering 7 questions (46.7%). GPT-4 had better performance in this category, with 12 correct responses (80.0%) and 3 incorrect responses (20.0%). However, the difference was not statistically significant (p= 0.121). The Traumatology category had GPT-3.5 correctly answering 7 questions (46.7%) and incorrectly answering 8 questions (53.3%). GPT-4 managed to correctly answer 11 questions (73.3%) while incorrectly answering 4 questions (26.7%). This difference was not statistically significant (p= 0.136). For Hepatopancreatobiliary Surgery and Pediatric Oncological Surgery, GPT-3.5 correctly answered 6 questions (40.0%) and incorrectly answered 9 questions (60.0%) in each category. On the other hand, GPT-4 correctly answered 10 questions (66.7%) and incorrectly answered 5 questions (33.3%) in both categories, with no statistically significant difference (p= 0.143). In total, GPT-3.5 correctly answered 46 questions out of 105 (43.8%), and incorrectly answered 59 questions (56.2%). GPT-4 showed a significantly improved performance, correctly answering 80 questions (76.2%) and incorrectly answering 25 questions (23.8%). The overall difference in performance was found to be statistically significant (p<0.001) (**Table I** and **figure 4**).

Given the total responses, when GPT-4 was compared with GPT-3.5, the Odds Ratio was found to be 4.1. This suggests that GPT-4 was 4.1 times more likely to provide a correct answer to the pediatric surgery questions compared to GPT-3.5.

Figure 4: Comparison of the frequency of the answers.

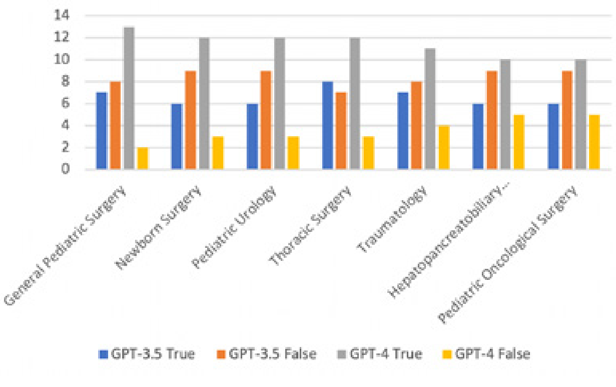


Table I: Comparison of GPT-3.5 and GPT-5 in terms of correct answers number.

	GPT-3.5				GPT-4				
	True		False		True		False		
	N	%	N	%	N	%	N	%	p-value
General Pediatric Surgery	7	46,7	8	53,3	13	86,7	2	13,3	0.020
Newborn Surgery	6	40,0	9	60,0	12	80,0	3	20,0	0.025
Pediatric Urology	6	40,0	9	60,0	12	80,0	3	20,0	0.025
Thoracic Surgery	8	53,3	7	46,7	12	80,0	3	20,0	0.121
Traumatology	7	46,7	8	53,3	11	73,3	4	26,7	0.136
Hepatopancreatobiliary Surgery	6	40,0	9	60,0	10	66,7	5	33,3	0.143
Pediatric Oncological Surgery	6	40,0	9	60,0	10	66,7	5	33,3	0.143
Total	46	43,8	59	56,2	80	76,2	25	23,8	<0.001

Discussion

To our knowledge, this represents the inaugural comparative analysis of GPT-3.5 and GPT-4 in the context of responding to EBPS examination questions. The results from this study demonstrate the significant differences in the accuracy and responsiveness of GPT-3.5 and GPT-4 to field-specific queries within the realm of pediatric surgery. In particular, GPT-4 showed a statistically significant improvement in answering questions related to pediatric surgery as compared to its predecessor, GPT-3.5.

Previous work in the field of medical question-answering research has frequently concentrated on more specific tasks with the intention of improving model performance at the expense of their generalizability[7,8]. For instance, Jin et al.[9] were able to achieve an accuracy of 68.1% with their model that responds to yes-or-no questions, the answers to which can be found in the corpus of abstracts that are available through PubMed. The pursuit of more generalizable models has been met with an increasing number of obstacles. On a data set consisting of 12,723 questions taken from Chinese medical licensing exams, a different Jin et al[10] achieved an accuracy of 36.7%. In a similar, Ha et al.[11] reported only a 29% accuracy on 454 USMLE Step 1 and Step 2 questions in the year 2019. Gilson and colleagues found that when posing questions from the United States Medical Licensing Examination Step 1 and Step 2 exams to ChatGPT, the correct response rate was determined to be 58%[12-14]. However in our study, In total, GPT-3.5 correctly answered 46 questions out of 105 (43.8%), and incorrectly answered 59 questions (56.2%). GPT-4 showed a significantly improved performance, correctly answering 80 questions (76.2%) and incorrectly answering 25 questions (23.8%). The overall difference in performance was found to be statistically significant ($p<0.001$). In this study, the highest accuracy rate was observed in the General Pediatric Surgery section (86.7%).

This study has several limitations. To begin, the ChatGPT algorithm was initially trained on a corpus that was constructed using data that was produced on or before the year 2021. Because of this restriction, the model's prompts can only contain information that was discovered before that date. Second, because this model is closed and does not have a public application programming interface (API), we are unable to fine-tune it using data that is specific to a task and investigate the extent of the inherent stochasticity that it possesses. The fact that this work investigates ChatGPT's performance in context on the EBPS exam, however, means that these limitations did not hinder our analysis in any way. Third, updates to ChatGPT are being released on a regular basis. It is believed that these updates are the result of training on inputs provided by users as they are received. The version of ChatGPT that was used in this research was an older model than the one that was published at the time of the study's completion. When everything is taken into consideration, it is reasonable to hypothesize that the performance of the model will not suffer a significant decline with each new iteration of the model when it is applied to the task that we have outlined, and that the performance may even improve.

In conclusion, this comparative study concludes that GPT-4 significantly outperforms GPT-3.5 in responding to EBPS exam questions, showing a 76.2% accuracy rate compared to 43.8%. Thus, newer iterations of ChatGPT models may offer promising applications for professional, field-specific inquiries in medicine and pediatric surgery. Further studies are needed to evaluate the effectiveness of the GPT for EBPS exam preparation.

Conflict of Interest

The authors declare that there is no conflict of interest.

References

1. OpenAI. Introducing ChatGPT. Accessed from: https://openai.com/blog/chatgpt, Accessed May 15, 2023.

2. Biswas S. ChatGPT and the Future of Medical Writing. Radiology. 2023;307(2):e223312.

3. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023;11(6):887.

4. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. J Med Syst. 2023;47(1):33.

5. Eysenbach G. The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers. JMIR Med Educ. 2023;9:e46885.

6. Fatani B. ChatGPT for Future Medical and Dental Research. Cureus. 2023;15(4):e37285.

7. Huang J, Tan M. The role of ChatGPT in scientific communication: writing better scientific review articles. Am J Cancer Res. 2023 Apr 15;13(4):1148-54.

8. Liebrenz M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. Lancet Digit Health. 2023 Mar;5(3):e105-e106.

9. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: a dataset for biomedical research question answering. arXiv doi: 10.48550/arXiv.1909.06146. Preprint posted online on September 13, 2019

10. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences. 2021;11(14):6421.

11. Ha LA, Yaneva V. Automatic question answering for medical MCQs: can it go further than information retrieval?. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019); RANLP 2019; September 2-4, 2019; Varna, Bulgaria. 2019. pp. 418-22.

12. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023;9:e45312.

13. Ulus SA. How does ChatGPT perform on the European Board of Orthopedics and Traumatology examination? A comparative study. Academic Journal of Health Sciences 2023; 38 (6):43-6.

14. Eroğlu Çakmakoğlu E. The place of ChatGPT in the future of dental education. J Clin Trials Exp Investig. 2023;2(3):121-9.