Clinical TRANSPLANTATION
The Journal of Clinical and Translational Research

ChatGPT Solving Complex Kidney Transplant Cases: A Comparative Study with Human Respondents

Journal:	*Clinical Transplantation*
Manuscript ID	CLTX-24-BCOM-0499
Wiley - Manuscript type:	Brief Communication
Date Submitted by the Author:	15-Jun-2024
Complete List of Authors:	Mankowski, Michal; New York University Department of Surgery Jaffe, Ian; New York University Department of Surgery Xu, Jingzhi; New York University Department of Surgery Bae, Sunjae ; New York University Department of Surgery; New York University Department of Population Health Oermann, Eric; New York University Department of Neurosurgery Aphinyanaphongs, Yindalon; New York University Department of Population Health; New York University Department of Medicine McAdams Demarco, Mara; New York University Department of Surgery; New York University Department of Population Health Lonze, Bonnie ; New York University Department of Surgery Orandi, Babak J.; New York University Department of Surgery; New York University Department of Medicine Stewart, Darren; New York University Department of Surgery Levan, Macey; New York University Department of Surgery; New York University Department of Population Health Massie, Allan B.; New York University Department of Surgery; New York University Department of Population Health Gentry, Sommer E.; New York University Department of Surgery; New York University Department of Population Health Segev, Dorry L.; New York University Department of Surgery; New York University Department of Population Health
Transplant Peer Review Network - Second Choice:	No referral
Transplant Peer Review Network - First Choice:	No referral
Discipline:	organ transplantation in general, kidney transplantation/nephrology
Keywords:	kidney (allograft) function / dysfunction, kidney failure / injury
Abstract:	**Introduction**: ChatGPT has shown the ability to answer clinical questions in general medicine but may be constrained by the specialized nature of kidney transplantation. Thus, it is important to explore how ChatGPT can be used in kidney transplantation and how its knowledge compares to human respondents. **Methods**: We prompted ChatGPT versions 3.5, 4, and 4 Visual (4V) with 12 multiple-choice questions related to six kidney transplant cases from the 2013-2015 American Society of Nephrology (ASN) fellowship program quizzes. We compared

	the performance of ChatGPT with US nephrology fellowship program directors, nephrology fellows, and the audience of the ASN's annual Kidney Week meeting. **Results**: Overall, ChatGPT 4V correctly answered 10 out of 12 questions, showing a performance level comparable to nephrology fellows (group majority correctly answered 9 of 12 questions) and training program directors (11 of 12). This surpassed ChatGPT 4 (7 of 12 correct) and 3.5 (5 of 12). All 3 ChatGPT versions failed to correctly answer questions where the consensus among human respondents was low. **Conclusion**: Each iterative version of ChatGPT performed better than the prior version, with version 4V achieving performance on par with nephrology fellows and training program directors. While it shows promise in understanding and answering kidney transplantation questions, Chat GPT should be seen as a complementary tool to human expertise rather than a replacement.

Brief Communication:

ChatGPT Solving Complex Kidney Transplant Cases: A Comparative Study with Human Respondents

Michal A. Mankowski, PhD[1], Ian S. Jaffe, BS[1], Jingzhi Xu, PhD[1], Sunjae Bae, MD, PhD, [1,2] Eric K. Oermann, MD,[3] Yindalon Aphinyanaphongs, MD, PhD,[2,4] Mara A. McAdams-DeMarco, PhD[1,2] Bonnie E. Lonze, MD, PhD,[1] Babak J. Orandi, MD, PhD[1,4], Darren Stewart, MS[1], Macey Levan, JD, PhD[1, 2], Allan Massie, PhD[1, 2], Sommer Gentry, PhD[1,2], Dorry L. Segev, MD, PhD[1,2]

[1] Department of Surgery, NYU Grossman School of Medicine, New York, NY
[2] Department of Population Health, NYU Grossman School of Medicine, New York, NY
[3] Department of Neurosurgery, NYU Grossman School of Medicine, New York, NY
[4] Department of Medicine, NYU Grossman School of Medicine, New York, NY

Running Title: ChatGPT Kidney Transplant Cases
Key Words: ChatGPT; kidney transplantation; artificial intelligence; generative pre-trained transformer; quiz

ORCIDs/X Handles:
Michal Mankowski - @MichalMankowski
Ian S. Jaffe - 0000-0002-7309-308X
Jingzhi Xu -
Sunjae Bae - 0000-0003-0098-8816
Eric K. Oermann - 0000-0002-1876-5963
Yindalon Aphinyanaphongs - 0000-0001-8605-5392
Mara A. McAdams-DeMarco - 0000-0003-3013-925X
Bonnie E. Lonze - 0000-0002-0973-1657
Babak Orandi - 0000-0001-6026-7135
Darren Stewart - 0000-0002-6764-4842
Macey Levan - 0000-0002-4239-1252 / @DrMaceyLevan
Allan Massie - 0000-0002-5288-5125 / @AllanBMassie
Sommer Gentry - 0000-0003-4530-8917 / @shelikesmath
Dorry L. Segev - 0000-0002-1924-4801 / @Dorry_Segev

Corresponding author:
Michal Mankowski
Center for Surgical and Transplant Applied Research
Department of Surgery
New York University Grossman School of Medicine
One Park Ave,
6th floor, Room 6-664
New York City, NY 10016
michal.mankowski@nyulangone.org

Word count: 1931/2000 (abstract 204/250)

M Mankowski, IS Jaffe, J Xu, S Bae, EK Oermann, Y Aphinyanaphongs, MA McAdams-DeMarco, BE Lonze, BJ Orandi, D Stewart, M Levan, A Massie, S Gentry, DL Segev

ChatGPT Solving Complex Kidney Transplant Cases: A Comparative Study with Human Respondents

Clin. Transpl.

Abbreviations

AI, artificial intelligence
GPT, generative pre-trained transformer
LLM, large language model

Abstract

Introduction: ChatGPT has shown the ability to answer clinical questions in general medicine but may be constrained by the specialized nature of kidney transplantation. Thus, it is important to explore how ChatGPT can be used in kidney transplantation and how its knowledge compares to human respondents.

Methods: We prompted ChatGPT versions 3.5, 4, and 4 Visual (4V) with 12 multiple-choice questions related to six kidney transplant cases from the 2013-2015 American Society of Nephrology (ASN) fellowship program quizzes. We compared the performance of ChatGPT with US nephrology fellowship program directors, nephrology fellows, and the audience of the ASN's annual Kidney Week meeting.

Results: Overall, ChatGPT 4V correctly answered 10 out of 12 questions, showing a performance level comparable to nephrology fellows (group majority correctly answered 9 of 12 questions) and training program directors (11 of 12). This surpassed ChatGPT 4 (7 of 12 correct) and 3.5 (5 of 12). All 3 ChatGPT versions failed to correctly answer questions where the consensus among human respondents was low.

Conclusion: Each iterative version of ChatGPT performed better than the prior version, with version 4V achieving performance on par with nephrology fellows and training program directors. While it shows promise in understanding and answering kidney transplantation questions, Chat GPT should be seen as a complementary tool to human expertise rather than a replacement.

Key Words: ChatGPT; kidney transplantation; artificial intelligence; generative pre-trained transformer; quiz

Off-Print Requests:
Michal Mankowski
One Park Ave,
6th floor, Room 6-664

New York City, NY 10016
michal.mankowski@nyulangone.org

1. Introduction

ChatGPT has been increasingly utilized in medicine due to its effectiveness in understanding and generating human-like text. The platform is built on the GPT[1] (Generative Pre-trained Transformer) architecture, utilizing a transformer-based[2] model that combines self-attention mechanisms with feedforward layers, enabling it to capture linguistic patterns and contexts. Its versatility has been demonstrated in various settings, including answering medical board/license examination questions[3–5] and addressing patients' queries related to medications, treatment plans, and post-treatment care.[6,7] The use of generative AI tools like ChatGPT in medicine continues to grow, with its potential to improve patient communication,[8,9] medical education,[10–12] and clinical decision-making.[13–15] Moreover, newer versions of ChatGPT have extended its applications by incorporating interpretation of visual and audio materials.[16]

However, the generalizability of ChatGPT's knowledge about general medicine to kidney transplantation may be limited due to the unique complexities of transplant medicine. AI systems like ChatGPT may lack the specificity needed for transplant-related issues like immunosuppression management and graft rejection. The dynamic nature of transplant protocols further challenges the relevance of AI tools in this field. Therefore, this study aims to explore how ChatGPT can be specifically applied to clinical interpretation in kidney transplantation using complex clinical cases.

In this study, we evaluated the performance of ChatGPT in answering multiple-choice questions from the transplant section of the American Society of Nephrology's annual fellows training quiz. We compared the answers of ChatGPT with answers provided by training program directors of US nephrology fellowship programs, nephrology fellows, and the audience of the annual Kidney Week meeting of the American Society of Nephrology. This study is the first to compare the performance of ChatGPT in solving kidney transplant cases with human respondents.

2. Methods

2.1 American Society of Nephrology Quiz and Questionnaire

The Nephrology Quiz and Questionnaire is an educational session at the annual Kidney Week meeting of the American Society of Nephrology. A panel of experts prepares various clinical cases, including kidney transplant cases, each with two multiple-choice, single best-answer questions. Before the live session, United States nephrology training program directors and nephrology fellows independently answer the questions through an internet-based questionnaire. During the session, audience members (which may include any attendee of Kidney Week, ranging from individuals with no clinical training to expert nephrologists) compare their knowledge and judgment on case-oriented questions with the training program directors and fellows.

2.2 Cases

In our study, we analyzed kidney transplantation cases from the American Society of Nephrology Quiz and Questionnaire from 2013,[17] 2014,[18] and 2015.[19] The quizzes

consisted of six cases (two per year) and 12 single-answer multiple-choice questions (two questions per case) in total. Each case has been summarized by ChatGPT 4, and these summaries alongside the related questions are presented in Table 1. Although small, this highly tailored dataset of high-quality questions is aligned with similar efforts to identify smaller, quality datasets to investigate specific aspects of LLM performance.[20,21]

2.3. Quiz Answers: ChatGPT versus Human Respondents

We used the NYU Langone Health's instance on Azure OpenAI Studio to access three ChatGPT versions: ChatGPT 3.5, ChatGPT 4, and ChatGPT 4 Vision (4V) and used default settings. Each version was prompted 10 times with each case description (including lab result tables and figures captions) along with the two associated multiple-choice questions; repeated prompting was used to mitigate random response selection given the moderate temperature settings in default ChatGPT model standard settings.[22] The last line of the prompt requested that ChatGPT select an answer to each question based on the case. ChatGPT 4V was also given access to the figures for each case (when applicable). As image interpretation is not a component of ChatGPT 3.5 or 4, we relied on the written image interpretations incorporated into the written case descriptions, which were present in all cases with images.

The responses of ChatGPT 3.5, ChatGPT 4, and ChatGPT 4V were compared with the responses of the training program directors of US nephrology fellowship programs, nephrology fellows, and the audience of the Annual Kidney Week Meeting

of the American Society of Nephrology. The number of respondents, demographic information, and training details for the human groups was not reported in their initial publication, and thus are unavailable. We evaluated whether a given group (training program directors, fellows, audience, or ChatGPT version) answered a question correctly if the most frequently chosen answer by a given group was correct. In case two or more answers were most frequently chosen and had the same frequency, we deemed these answers incorrect. We also separately evaluated the reproducibility of correct responses from the ChatGPT models; responses were considered reproducibly correct when the ChatGPT model answered the question correctly in all 10 prompting sessions.

3. Results

Figure 1 summarizes the performance of each responder group in answering quiz questions. Across the 6 cases and 12 related questions, training program directors answered 11 of 12 questions correctly. ChatGPT 4V answered 10 of 12 questions correctly and fellows answered 9 of 12 questions correctly. Both the Kidney Week audience and ChatGPT 4 answered 7 of 12 questions correctly. ChatGPT 3.5 provided 5 of 12 correct answers.

In 2015, ChatGPT 4, ChatGPT 4V, the nephrology fellows, and the training program directors correctly answered both questions of case 1 (Q1A and Q1B), while

ChatGPT 3.5 failed to answer both. None of the groups answered Q2A correctly, while only the training program directors, the Kidney Week audience, and ChatGPT 4V answered Q2B for case 2 correctly.

In 2014, ChatGPT 3.5, ChatGPT 4V, and the training program directors answered all four questions for both cases correctly. The nephrology fellows and ChatGPT 4 each failed to answer one question, Q1B and Q2A, respectively, while the Kidney Week audience answered only two questions correctly across the two cases.

In 2013, only the training program directors and fellows had answered all questions correctly. ChatGPT 4V missed just Q1A, while ChatGPT 4 missed both Q1A and Q2B and ChatGPT 3.5 only answered Q2A correctly. The audience missed only one question (Q1A).

Table S1 shows the distribution of answers to multi-choice questions by the responders' group. Responses for ChatGPT models were considered reproducibly correct when the ChatGPT model not only answered the question correctly, but did so 100% of the time. The higher correct response rates seen for ChatGPT 4V (10 of 12 questions) and Chat GPT 4 (7 of 12) compared to ChatGPT 3.5 (5 of 12), corresponded to higher reproducibly correct response rates; ChatGPT 4V was reproducibly correct on 9 of 12 questions, ChatGPT 4 on 6 of 12 and ChatGPT 3.5 on 4 of 12.

For questions where all three ChatGPT versions failed, there was less consensus among human respondents. For instance, in the 2013 quiz Q1A, only 40% of fellows, 69% of training program directors, and 27% of the audience provided correct answers, while ChatGPT versions 3.5 and 4 did not provide correct answers in any of the 10 prompt sessions and ChatGPT 4V provided the correct answer in only a single session. Similarly, in the 2015 quiz Q2A, no group most frequently selected the correct response.

4. Discussion

The utilization of Large Language Models (LLMs), such as ChatGPT, within medicine[23,24] has introduced a novel approach to assimilating medical knowledge. Our findings indicate that various versions of ChatGPT successfully answered an array of case-based kidney transplant questions from the 2013-2015 American Society of Nephrology Quiz and Questionnaire. Interestingly, in the 2015 quiz, ChatGPT 4V answered 3 of the 4 questions correctly, matching the performance of nephrology training program directors. In the 2014 quiz, both ChatGPT 3.5 and 4V answered all questions correctly, paralleling the performance of the training program directors. Yet, both ChatGPT 3.5 and 4 underperformed in the 2013 quiz, answering fewer questions correctly than all human groups, while ChatGPT 4V performed on par with the Kidney Week audience, but fell short when compared to nephrology experts.

While earlier versions of ChatGPT were unable to outperform the human groups, specifically the nephrology fellow and training program director groups, ChatGPT 4V

demonstrated comparable performance to these knowledgeable groups. This suggests an improvement in the capabilities of the GPT models to handle multiple-choice reasoning tasks in highly specialized areas like transplant nephrology. They are consistent with a recent study conducted on general nephrology test questions, where ChatGPT 4 achieved an accuracy of 74%, slightly below the average 77% accuracy of nephrology examinees.[25] Our study indicates a notable overall performance improvement in ChatGPT 4V compared to ChatGPT 4, and in ChatGPT 4 compared to ChatGPT 3.5, suggesting that LLM model performance may enhance with subsequent generations and benefit from multimodal inputs such as images. Considering that these models were iteratively improved over the span of less than two years, this improvement reflects significant change in a short period of time. The swift progress in the AI field implies that these models may eventually exceed human experts, especially in multiple-choice constrained reasoning tasks.

Furthermore, there is burgeoning evidence that LLMs can offer beneficial responses to non-constrained questions specifically related to transplantation. For instance, ChatGPT responses to questions related to the treatment of kidney transplant recipients showed considerable knowledge of kidney transplantation, albeit with some inaccuracies and a lack of depth.[26] Regarding liver disease, ChatGPT has been evaluated in answering general questions on liver transplantation, where 70.6% of clinical experts found it accurate and comparable to practicing clinicians.[27] ChatGPT was also evaluated in generating research ideas in liver transplantation.[28]

There are several reasons why AI models like ChatGPT may provide incorrect answers when compared to humans. The medical knowledge of ChatGPT is derived from its training on a diverse range of internet text. Its performance is not a result of its understanding of medicine but its ability to generate plausible-sounding text based on the patterns it learned during training.[29] ChatGPT models do not have a conceptual understanding of the topics they are trained on. Instead, they generate responses based on patterns in the data they were trained on, generating text in a probabilistic manner. ChatGPTs tend to hallucinate,[30,31] and make up fictitious information. This lack of understanding can lead to mistakes or oversights that a human expert possibly would not make. Additionally, current ChatGPT models do not have the capability to autonomously gain new knowledge or learn from experience. Their knowledge is fixed at the time of training and is not being updated or expanded upon until the next software release.

It is important to consider the limitations of our study as well. We cannot exclude the possibility that the quiz questions and answers were themselves part of the training data for either or both versions of ChatGPT since these quizzes were published in 2013-2015. This would tend to bias our conclusions toward higher performance of ChatGPT, however we observed that none of these ChatGPT versions gave entirely correct answers, suggesting the quizzes themselves may not have been part of the training corpus. Another limitation is that the assessment of ChatGPT's performance was based solely on a set of questions from the American Society of Nephrology Quiz cases, which do not represent the full spectrum of clinical scenarios in kidney

transplantation. The study's format may have introduced a limitation as it required all case information to be converted into text format to be processed by both earlier ChatGPTs (3.5 and 4) in a comparable way. This constraint could have resulted in a potential loss of context or nuance in the presented medical cases. While this was mitigated by the fact that the images were also described in the case description and in the figure captions, this could provide an explanation for some of the improvement in performance observed in ChatGPT 4V.

In conclusion, while LLM models like ChatGPT have stirred enthusiasm for potential medical applications, including kidney transplantation, it is crucial to comprehend their limitations and utilize them as a complementary tool to human expertise rather than a replacement. Our study evaluated the performance of ChatGPT in solving clinical kidney transplant scenarios and compared the results with human respondents of varying levels of expertise. The advanced ChatGPT 4V model performed comparably to nephrology fellows and nephrology training program directors on this small sample of questions, correctly answering ten out of twelve questions related to kidney transplantation cases. Such performance supports further exploration in using LLMs to assist with clinical interpretation, answering patient questions, and other tasks specific to organ transplant.

Acknowledgements and Funding:

Author Contributions: Concept/design: MM, JX, SG, DL; Data collection: MM, JX; Data analysis/interpretation: all authors; Drafting article: all authors; Critical revision of article: all authors; Approval of article: all authors; Funding secured by: DLS.

Funding: This work was supported by grant number K24AI144954 (Segev) from the National Allergy and Infectious Disease (NIAID) and R01DK132395 (Massie) from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). I.S. Jaffe was supported by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through grant award number UL1TR001445. M. McAdams-DeMarco was supported by K02AG076883 and R01AG077888 from the National Institute on Aging (NIA). The work described here is the responsibility of the authors alone and does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Conflict of Interest Disclosures:

M Mankowski: None, IS Jaffe: None, J Xu: None, None, S Bae: None, EK Oermann: None, Y Aphinyanaphongs: None, MA McAdams DeMarco: Dr. McAdams-DeMarco reports speaker honoraria from Chiesi Farmaceutici S.p.A. unrelated to this work, B Lonze: None, B Orandi: None, D Stewart: None, M Levan: None, A Massie: None, S Gentry: None, DL Segev: None

References

1. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. Published online 2018.

2. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Advances in neural information processing systems*. 2017;30.

3. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. *JMIR Medical Education*. 2023;9(1):e48002. doi:10.2196/48002

4. Joly-Chevrier M, Nguyen AXL, Lesko-Krleza M, Lefrançois P. Performance of ChatGPT on a practice dermatology board certification examination. *Journal of cutaneous medicine and surgery*. 2023;27(4):407-409.

5. Beam K, Sharma P, Kumar B, et al. Performance of a large language model on practice questions for the neonatal board examination. *JAMA pediatrics*. 2023;177(9):977-979.

6. Harris E. Large language models answer medical questions accurately, but can't match clinicians' knowledge. *JAMA*. Published online 2023.

7. Haupt CE, Marks M. AI-generated medical advice—GPT and beyond. *Jama*. 2023;329(16):1349-1350.

8. Gordon EB, Towbin AJ, Wingrove P, et al. Enhancing patient communication with Chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions. *Journal of the American College of Radiology*. 2024;21(2):353-359.

9. Nashwan AJ, Abujaber AA, Choudry H. Embracing the future of physician-patient communication: GPT-4 in gastroenterology. *Gastroenterology & Endoscopy*. 2023;1(3):132-135.

10. Lee H. The rise of ChatGPT: Exploring its potential in medical education. *Anatomical sciences education*. Published online 2023.

11. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT-Reshaping medical education and clinical management. *Pakistan Journal of Medical Sciences*. 2023;39(2):605.

12. Hswen Y, Abbasi J. AI will—and should—change medical school, says Harvard's dean for medical education. *JAMA*. Published online 2023.

13. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. *Journal of Medical Internet Research*. 2023;25:e48568.

14. Sorin V, Klang E, Sklair-Levy M, et al. Large language model (ChatGPT) as a support tool for breast tumor board. *NPJ Breast Cancer*. 2023;9(1):44.

15. Rao A, Kim J, Kamineni M, et al. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. *Journal of the American College of Radiology*. 2023;20(10):990-997.

16. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. *Advances in neural information processing systems*. 2024;36.

17. Josephson MA, Perazella MA, Choi MJ. American society of nephrology quiz and questionnaire 2013: Transplantation. *Clinical Journal of the American Society of Nephrology*. 2014;9(7):1319-1327.

18. Josephson MA, Perazella MA, Choi MJ. American society of Nephrology Quiz and Questionnaire 2014: transplantation. *Clinical Journal of the American Society of Nephrology*. 2015;10(5):903-909.

19. Josephson MA, Perazella MA, Choi MJ. American society of nephrology quiz and questionnaire 2015: Transplantation. *Clinical Journal of the American Society of Nephrology*. 2016;11(6):1114-1122.

20. Jimenez CE, Yang J, Wettig A, et al. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? Published online April 5, 2024. doi:10.48550/arXiv.2310.06770

21. Rein D, Hou BL, Stickland AC, et al. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. Published online November 20, 2023. doi:10.48550/arXiv.2311.12022

22. Davis J, Van Bulck L, Durieux BN, Lindvall C. The temperature feature of ChatGPT: Modifying creativity for clinical research. *JMIR Human Factors*. 2024;11(1):e53559.

23. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. Drazen JM, Kohane IS, Leong TY, eds. *N Engl J Med*. 2023;388(13):1233-1239. doi:10.1056/NEJMsr2214184

24. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. *Jama*. 2023;330(9):866-869.

25. Miao J, Thongprayoon C, Valencia OAG, et al. Performance of ChatGPT on nephrology test questions. *Clinical Journal of the American Society of Nephrology*. Published online 2023:10-2215.

26. Rawashdeh B, Kim J, AlRyalat SA, Prasad R, Cooper M. ChatGPT and artificial intelligence in transplantation research: is it always correct? *Cureus*. 2023;15(7).

27. Endo Y, Sasaki K, Moazzam Z, et al. Quality of ChatGPT responses to questions related to liver transplantation. *Journal of Gastrointestinal Surgery*. 2023;27(8):1716-1719.

28. Akabane M, Iwadoh K, Melcher ML, Sasaki K. Exploring the potential of ChatGPT in generating unknown clinical questions about liver transplantation: A feasibility study. *Liver Transplantation*. 2024;30(2):229-234.

29. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? □. In: *Proceedings of the 2021 ACM*

Conference on Fairness, Accountability, and Transparency. FAccT '21. Association for Computing Machinery; 2021:610-623. doi:10.1145/3442188.3445922

30. Hanna E, Levic A. Comparative Analysis of Language Models: hallucinations in ChatGPT: Prompt Study. Published online 2023.

31. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. *Critical Care*. 2023;27(1):120.

Table 1. Summaries of the 2015, 2014, and 2013 kidney transplant cases from the American Society of Nephrology Quiz and related questions. Correct answers are bolded.

Year	Case	Case Summary (by ChatGPT 4)	Lab results	Image	Question
2015	1	A 53-year-old woman with a history of ESRD, aortic stenosis, cancer, COPD, genitourinary reflux disease, and recent kidney transplant presents to the ER with shortness of breath, dyspnea, pleuritic chest pain, and cough. CT and CXR reveal dense consolidation in the left lower lobe, multiple areas of nodular opacities in both lungs, and an abnormal aortic valve, including thickened prosthesis leaflets indicating severe prosthetic aortic valve stenosis.	Laboratory results: Complete blood count	No	1A: Which of the following is the most likely cause of her chest pain and CXR findings? (A) Lung cancer, (B) Recurrent breast cancer, **(C) Pneumonia**, (D) Sarcoidosis
					1B: With the addition of voriconazole, which of the following changes would develop in her immunosuppression levels? **(A) Tacrolimus trough increased**, (B) Tacrolimus trough decreased, (C) MMF level increased, **(**D) MMF level decreased
	2	A 63-year-old woman with a history of non-alcoholic steatohepatitis, type 2 diabetes, hypertension, and end-stage renal disease treated with a kidney transplant, is admitted to the hospital for liver failure and volume overload and is being considered for a combined liver-kidney transplant. Her condition deteriorates with declining kidney function, rising serum creatinine, increased fatigue, decreased urine output, and the presence of numerous muddy brown granular casts in her urine sediment.	Laboratory results: Complete blood count	No	2A: What is the most likely underlying cause of the patient's decline in kidney function? (A) Hepatorenal syndrome, (B) Antibody-mediated rejection, (C) Acute tubular necrosis (ATN), (D) Recurrent diabetes mellitus, **(E) Impossible to predict**
					2B: Which of the following is likely contributing most to her massive volume overload? (A) Hepatorenal syndrome, (B) ATN, (C) Nephrotic syndrome, **(D) Heart failure**
2014	1	A 50-year-old man with a history of ESRD due to diabetes and hypertension, who had received a deceased donor kidney transplant, presented 9 months after operation with fever, fatigue, lightheadedness, loose stool and right upper quadrant pain. His medication list was extensive, including tacrolimus, mycophenolate mofetil, prednisone and trimethoprim sulfamethoxazole, and his lab results revealed anemia and underproduction of cells, with no evidence of malignancy, ulceration, or bleeding from prior colonoscopy and esophagogastroduodenoscopy.	(1) Laboratory findings, (2) Anemia laboratory evaluation	Post-transplant hematocrit	1A: Testing for which of the following may be most helpful in evaluating the underproduction anemia? (A) Cytomegalovirus (CMV), (B) BK virus, **(C) Parvovirus B19,** (D) Clostridium difficile, (E) JC virus
					1B: Which of the following is most effective in treating Parvovirus B19? **(A) Intravenous Ig (IVIG)**, (B) Cidofovir, (C) Valganciclovir, (D) Acyclovir, (E) Immunosuppression reduction
	2	A 52-year-old woman who received a kidney transplant was readmitted to the hospital with an altered mental status following a time period of post-transplant complications including pancreatitis, drain placement, and rising creatinine	Cerebrospinal fluid results	Brain imaging studies. (A) Head CT. (B) Brain MRI.	2A: What is the most likely diagnosis? **(A) Posterior reversible encephalopathy syndrome (PRES),** (B) CMV encephalitis, (C) Progressive multifocal leukoencephalopathy (PML), (D) Herpes simplex virus (HSV) encephalitis, (E)JC virus encephalitis,

		levels. Despite a normal diagnostic work-up, she displays signs of encephalopathy, with her most likely diagnosis being Posterior Reversible Encephalopathy Syndrome (PRES), often associated with the use of tacrolimus.			2B: Which of our patient's medications has most commonly been associated with PRES? (A) Valganciclovir, (B) Famotidine, **(C) Tacrolimus**, (D) Simvastatin
2013	1	A 66-year-old Laotian man who had a history of end-stage renal disease, hyperlipidemia, and chronic hepatitis B, and had received a kidney transplant, presented with a nonproductive cough and right flank pain, three months post-transplant. A CT scan revealed a large mass encasing a rib, and treatment for tuberculosis was initiated.		Mass encasing a rib. A CT scan image.	1A: What is the most likely diagnosis in this patient? (A) Recurrent thymoma, (B) Post-transplant lymphoproliferative disorder (PTLD), (C) Malignant BK nodule, **(D) Extrapulmonary tuberculosis (TB),** (E) Brown tumor 1B: How might TB treatment affect the tacrolimus level? **(A) Decrease the tacrolimus level, because rifampin is a CYP3A4 inducer,** (B) Increase the tacrolimus level, because INH is a CYP3A4 competitor, (C) Decrease the tacrolimus level, because ethambutol is a CYP3A4 inducer, (D) No marked change, because by using INH and ethambutol, they will cancel each other's effect, (E) Increase the tacrolimus level, because pyrazinamide is a CYP3A5 inducer
	2	A 30-year old man with ESRD and blood type AB, who had received a deceased donor kidney, developed deep vein thrombosis in his upper extremity 2.5 years post-transplant. Despite an initial delay in graft function, he developed complications such as acute T cell mediated rejection, chronic allograft arteriopathy, chronic transplant glomerulopathy, peritubular capillaritis, and moderate interstitial fibrosis and tubular atrophy.	(1) laboratory studies at the time of kidney biopsy and 19 days later (2) Selected additional laboratory studies	(1)Peritubular capillaritis consistent with AMR, (2) C4d staining, (3) Lymphocytes lifting up endothelial cells diagnostic of Banff IIA rejection.	2A: Which of the following is the most likely mechanism of anemia in this patient? (A) Iron deficiency anemia secondary to a bleed, (B) Decreased production anemia caused by bone marrow suppression, (C) Dilutional anemia caused by volume overload, **(D) Intravascular hemolysis** 2B: What is the most likely cause of the hemolytic anemia in this patient? The rejection episode (A) Solumedrol, (B) IVIG, (**C) Occult infection,** (D) Sirolimus toxicit

A

		Case 1		Case 2		
		Q1A	Q1B	Q2A	Q2B	SUM
2015 Quiz	ChatGPT 3.5	✗	✗	✗	✗	0
	ChatGPT 4	✓	✓	✗	✗	2
	ChatGPT 4V	✓	✓	✗	✓	3
	Fellow	✓	✓	✗	✗	2
	TPD	✓	✓	✗	✓	3
	Audience	✗	✓	✗	✓	2
2014 Quiz	ChatGPT 3.5	✓	✓	✓	✓	4
	ChatGPT 4	✓	✓	✗	✓	3
	ChatGPT 4V	✓	✓	✓	✓	4
	Fellow	✓	✗	✓	✓	3
	TPD	✓	✓	✓	✓	4
	Audience	✓	✗	✗	✓	2
2013 Quiz	ChatGPT 3.5	✗	✗	✓	✗	1
	ChatGPT 4	✗	✓	✓	✗	2
	ChatGPT 4V	✗	✓	✓	✓	3
	Fellow	✓	✓	✓	✓	4
	TPD	✓	✓	✓	✓	4
	Audience	✗	✓	✓	✓	3
Overall	ChatGPT 3.5					5
	ChatGPT 4					7
	ChatGPT 4V					10
	Fellow					9
	TPD					11
	Audience					7

B Total correct responses to kideny transplant cases

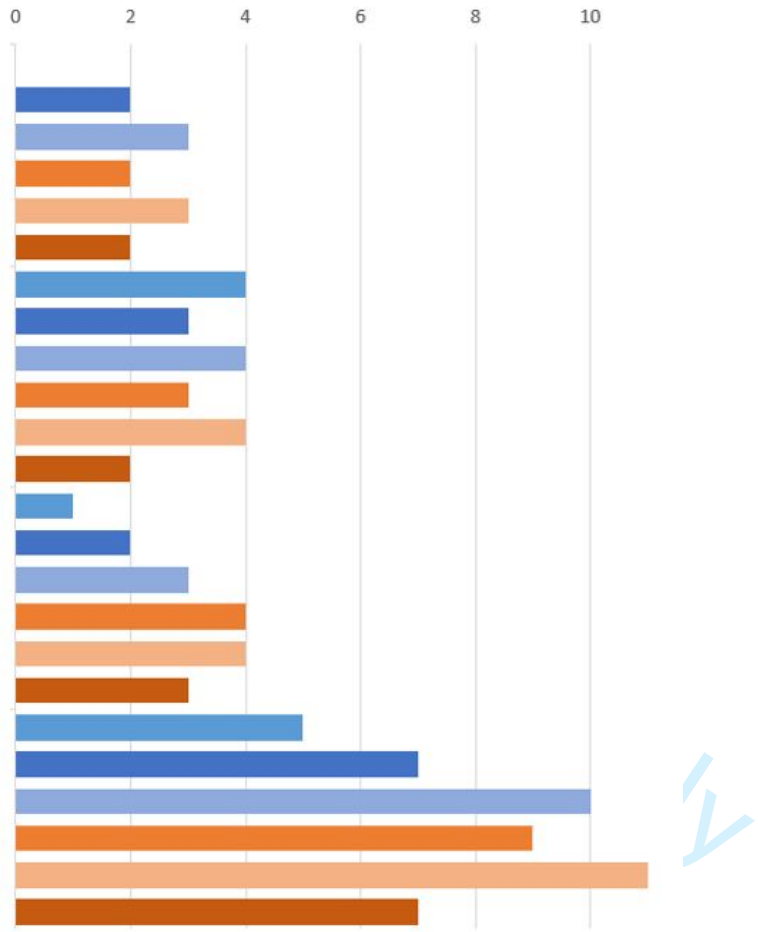

Figure 1. Human and GPT model performance on the American Society of Nephrology Quiz transplantation questions.

A) Summative performance on each question to the 2015, 2014, and 2013 kidney transplant cases from the American Society of Nephrology Quiz are shown. We assumed that a given group answered a question correctly if the most frequently chosen answer was correct. In case two or more answers were most frequently chosen, we deemed these answers incorrect. Responses for fellows, training program directors (TPD), and audience members at the Annual Kidney Week Meeting of the American Society of Nephrology were reported in Josephson et al., 2013, 2014, and 2015. Chat GPT 3.5, Chat GPT 4, and Chat GPT 4V were prompted 10 times with each case and associated questions. **B)** Total correct responses for each group by quiz year and for all three years combined.

Supplementary Materials

Table S1. Question-specific performance on American Society of Nephrology Quiz transplantation questions for human respondents and GPT models. Answers to 2015, 2014, and 2013 kidney transplant cases from the American Society of Nephrology Quiz transplantation questions are shown. Response frequencies for fellows, training program directors (TPD), and audience members at the Annual Kidney Week Meeting of the American Society of Nephrology were reported in Josephson et al., 2013, 2014, and 2015. Chat GPT 3.5, Chat GPT 4, and Chat GPT 4V were prompted 10 times with each case and associated questions. Gray columns indicate correct answers. The most frequent responses are bolded and underlined for each of the responder's groups.

		Case 1										Case 2									
		Q1A					Q1B					Q2A					Q2B				
		A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
2015 Quiz	ChatGPT 3.5	**70%**	0%	30%	0%		0%	**80%**	10%	10%		0%	0%	**100%**	0%	0%	**60%**	0%	0%	40%	
	ChatGPT 4	20%	0%	**80%**	0%		**100%**	0%	0%	0%		0%	0%	**100%**	0%	0%	**100%**	0%	0%	0%	
	ChatGPT 4V	0%	0%	**100%**	0%		**100%**	0%	0%	0%		0%	0%	**100%**	0%	0%	20%	0%	0%	**80%**	
	Fellow	13%	16%	**68%**	3%		**73%**	16%	0%	11%		6%	3%	**82%**	0%	9%	33%	**37%**	3%	27%	
	TPD	0%	19%	**81%**	0%		**94%**	0%	0%	6%		0%	0%	**73%**	0%	27%	0%	27%	27%	**46%**	
	Audience	13%	29%	17%	41%		**76%**	18%	0%	6%		19%	3%	**69%**	0%	9%	7%	21%	8%	**64%**	
2014 Quiz	ChatGPT 3.5	0%	0%	**100%**	0%	0%	**100%**	0%	0%	0%	0%	**100%**	0%	0%	0%	0%	0%	0%	**100%**	0%	
	ChatGPT 4	0%	0%	**100%**	0%	0%	**100%**	0%	0%	0%	0%	30%	0%	**70%**	0%	0%	0%	0%	**100%**	0%	
	ChatGPT 4V	0%	0%	**100%**	0%	0%	**100%**	0%	0%	0%	0%	**100%**	0%	0%	0%	0%	0%	0%	**100%**	0%	
	Fellow	11%	4%	**84%**	1%	0%	28%	5%	6%	1%	**60%**	**84%**	4%	3%	0%	9%	6%	0%	**94%**	0%	
	TPD	4%	0%	**96%**	0%	0%	**90%**	5%	0%	0%	5%	**90%**	5%	0%	0%	5%	0%	0%	**100%**	0%	
	Audience	26%	8%	**63%**	2%	1%	20%	3%	6%	1%	**70%**	**24%**	**24%**	13%	22%	17%	11%	3%	**86%**	0%	
2013 Quiz	ChatGPT 3.5	0%	**80%**	20%	0%	0%	20%	**80%**	0%	0%	0%	30%	0%	0%	**70%**		0%	**50%**	0%	50%	0%
	ChatGPT 4	0%	**100%**	0%	0%	0%	**100%**	0%	0%	0%	0%	0%	0%	0%	**100%**		0%	**100%**	0%	0%	0%
	ChatGPT 4V	0%	**90%**	0%	10%	0%	**100%**	0%	0%	0%	0%	0%	0%	0%	**100%**		0%	0%	**100%**	0%	0%
	Fellow	13%	28%	7%	**40%**	12%	**60%**	19%	7%	7%	7%	13%	13%	13%	**61%**		20%	7%	**46%**	7%	20%
	TPD	10%	21%	0%	**69%**	0%	**78%**	0%	11%	11%	0%	5%	5%	0%	**90%**		17%	0%	**44%**	0%	39%
	Audience	11%	**40%**	15%	27%	7%	**56%**	21%	5%	12%	6%	1%	20%	3%	**76%**		9%	2%	**46%**	7%	36%

Supplementary Materials

Supplement to M Mankowski, IS Jaffe, J Xu, et al. ChatGPT Solving Complex Kidney Transplant Cases: A Comparative Study with Human Respondents.

Table of Contents

Supplemental Tables

Table S1. Question-specific performance on American Society of Nephrology Quiz transplantation questions for human respondents and GPT models.
Answers to 2015, 2014, and 2013 kidney transplant cases from the American Society of Nephrology Quiz transplantation questions are shown. Response frequencies for fellows, training program directors (TPD), and audience members at the Annual Kidney Week Meeting of the American Society of Nephrology were reported in Josephson et al., 2013, 2014, and 2015. Chat GPT 3.5, Chat GPT 4, and Chat GPT 4V were prompted 10 times with each case and associated questions. Gray columns indicate correct answers. The most frequent responses are bolded and underlined for each of the responder's groups.

		Case 1										Case 2									
		Q1A					Q1B					Q2A					Q2B				
		A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
2015 Quiz	ChatGPT 3.5	**70%**	0%	30%	0%		0%	**80%**	10%	10%		0%	0%	**100%**	0%	0%	**60%**	0%	0%	40%	
	ChatGPT 4	20%	0%	**80%**	0%		**100%**	0%	0%	0%		0%	0%	**100%**	0%	0%	**100%**	0%	0%	0%	
	ChatGPT 4V	0%	0%	**100%**	0%		**100%**	0%	0%	0%		0%	0%	**100%**	0%	0%	20%	0%	0%	**80%**	
	Fellow	13%	16%	**68%**	3%		**73%**	16%	0%	11%		6%	3%	**82%**	0%	9%	33%	**37%**	3%	27%	
	TPD	0%	19%	**81%**	0%		**94%**	0%	0%	6%		0%	0%	**73%**	0%	27%	0%	27%	27%	**46%**	
	Audience	13%	29%	17%	41%		**76%**	18%	0%	6%		19%	3%	**69%**	0%	9%	7%	21%	8%	**64%**	
2014 Quiz	ChatGPT 3.5	0%	0%	**100%**	0%	0%	**100%**	0%	0%	0%	0%	**100%**	0%	0%	0%	0%	0%	0%	**100%**	0%	
	ChatGPT 4	0%	0%	**100%**	0%	0%	**100%**	0%	0%	0%	0%	30%	0%	**70%**	0%	0%	0%	0%	**100%**	0%	
	ChatGPT 4V	0%	0%	**100%**	0%	0%	**100%**	0%	0%	0%	0%	**100%**	0%	0%	0%	0%	0%	0%	**100%**	0%	
	Fellow	11%	4%	**84%**	1%	0%	28%	5%	6%	1%	**60%**	**84%**	4%	3%	0%	9%	6%	0%	**94%**	0%	
	TPD	4%	0%	**96%**	0%	0%	**90%**	5%	0%	0%	5%	**90%**	5%	0%	0%	5%	0%	0%	**100%**	0%	
	Audience	26%	8%	**63%**	2%	1%	20%	3%	6%	1%	**70%**	**24%**	**24%**	13%	22%	17%	11%	3%	**86%**	0%	
2013 Quiz	ChatGPT 3.5	0%	**80%**	20%	0%	0%	20%	**80%**	0%	0%	0%	30%	0%	0%	**70%**		0%	**50%**	0%	50%	0%
	ChatGPT 4	0%	**100%**	0%	0%	0%	**100%**	0%	0%	0%	0%	0%	0%	0%	**100%**		0%	**100%**	0%	0%	0%
	ChatGPT 4V	0%	**90%**	0%	10%	0%	**100%**	0%	0%	0%	0%	0%	0%	0%	**100%**		0%	0%	**100%**	0%	0%
	Fellow	13%	28%	7%	**40%**	12%	**60%**	19%	7%	7%	7%	13%	13%	13%	**61%**		20%	7%	**46%**	7%	20%
	TPD	10%	21%	0%	**69%**	0%	**78%**	0%	11%	11%	0%	5%	5%	0%	**90%**		17%	0%	**44%**	0%	39%
	Audience	11%	**40%**	15%	27%	7%	**56%**	21%	5%	12%	6%	1%	20%	3%	**76%**		9%	2%	**46%**	7%	36%