









RESEARCH  
ARTICLE

-  **Muhammet Fethi Saglam**<sup>1</sup>  
 **Emrah Uguz**<sup>1</sup>  
 **Kemal Esref Erdogan**<sup>1</sup>  
 **Huseyin Unsal Ercelik**<sup>2</sup>  
 **Murat Yucel**<sup>2</sup>  
 **Cevat Ahmet Sert**<sup>3</sup>  
 **Fatih Yamac**<sup>4</sup>  
 **Erol Sener**<sup>1</sup>

<sup>1</sup> 1Ankara Yıldırım Beyazıt University Faculty of Medicine, Department of Cardiovascular Surgery, Ankara, Türkiye

<sup>2</sup> Ankara Bilkent City Hospital, Ankara, Türkiye

<sup>3</sup> Ankara Etlik City Hospital, Ankara, Türkiye

<sup>4</sup> Sincan Training and Research Hospital, Ankara, Türkiye

**Corresponding Author:**  
Muhammet Fethi Sağlam  
mail:dr.m.fethisaglam@gmail.com

Received: 02.01.2025  
Acceptance: 24.03.2025  
DOI:10.18521/kt.d.1611953

**Konuralp Medical Journal**  
e-ISSN1309-3878  
konuralptipdergi@duzce.edu.tr  
konuralptipdergisi@gmail.com  
www.konuralptipdergi.duzce.edu.tr



## Artificial Intelligence in Patient Communication: Performance of GPT-3.5 and GPT-4 in Coronary Bypass Surgery

### ABSTRACT

**Objective:** This study aims to evaluate the ability of GPT-3.5 and GPT-4 to provide accurate, comprehensible, and clinically relevant responses to common patient questions about coronary bypass surgery.

**Methods:** A cross-sectional study was conducted at Ankara Yıldırım Beyazıt University Bilkent City Hospital with 80 cardiovascular surgery specialists. Participants rated the responses of GPT-3.5 and GPT-4 to 10 common patient questions about coronary bypass surgery based on four criteria: accuracy, understandability, clinical appropriateness, and overall evaluation. Statistical analysis included independent t-tests, Cronbach's Alpha reliability analysis, and Cohen's d effect size calculation.

**Results:** GPT-4 significantly outperformed GPT-3.5 across all metrics. The mean scores for GPT-4 were higher in accuracy (3.02 vs. 1.77), understandability (2.99 vs. 1.81), clinical appropriateness (2.96 vs. 1.78), and overall evaluation (2.98 vs. 1.77) ( $p < 0.05$  for all). Cronbach's Alpha values indicated good internal consistency ( $\geq 0.69$  for all metrics), and Cohen's d effect sizes demonstrated large differences (1.54 to 1.65).

**Conclusions:** GPT-4 shows superior potential compared to GPT-3.5 in answering patient questions about coronary bypass surgery. Despite its strengths, occasional inaccuracies and incomplete responses highlight the need for further refinement. Future research should integrate patient feedback and evaluate the real-world clinical impact of these models to optimize their application in healthcare.

**Keywords:** Coronary Bypass Surgery, Artificial Intelligence, GPT-3.5, GPT-4, Patient Communication.

## Hasta İletişiminde Yapay Zeka: Koroner Bypass Cerrahisinde GPT-3.5 ve GPT-4'ün Performansı

### ÖZET

**Amaç:** Bu çalışma, GPT-3.5 ve GPT-4'ün koroner bypass cerrahisiyle ilgili yaygın hasta sorularına doğru, anlaşılır ve klinik olarak uygun yanıtlar verme yeteneğini değerlendirmeyi amaçlamaktadır.

**Yöntem:** Ankara Yıldırım Beyazıt Üniversitesi Bilkent Şehir Hastanesi'nde 80 kalp ve damar cerrahisi uzmanı ile kesitsel bir çalışma yürütülmüştür. Katılımcılar GPT-3.5 ve GPT-4'ün koroner bypass cerrahisi ile ilgili 10 yaygın hasta sorusuna verdiği yanıtları dört kritere göre değerlendirmiştir: doğruluk, anlaşılabilirlik, klinik uygunluk ve genel değerlendirme. İstatistiksel analiz bağımsız t-testlerini, Cronbach Alfa güvenilirlik analizini ve Cohen's d etki büyüklüğü hesaplamasını içermektedir.

**Bulgular:** GPT-4 tüm ölçütlerde GPT-3.5'ten önemli ölçüde daha iyi performans göstermiştir. GPT-4 için ortalama puanlar doğruluk (3,02'ye karşı 1,77), anlaşılabilirlik (2,99'a karşı 1,81), klinik uygunluk (2,96'ya karşı 1,78) ve genel değerlendirme (2,98'e karşı 1,77) açısından daha yüksekti (tümü için  $p < 0,05$ ). Cronbach's Alpha değerleri iyi bir iç tutarlılık (tüm ölçütler için  $\geq 0,69$ ) ve Cohen's d etki büyüklükleri büyük farklılıklar (1,54 ila 1,65) göstermiştir.

**Sonuç:** GPT-4, koroner bypass cerrahisi ile ilgili hasta sorularını yanıtlamada GPT-3.5'e kıyasla üstün potansiyel göstermektedir. Güçlü yönlerine rağmen, zaman zaman ortaya çıkan yanlışlıklar ve eksik yanıtlar daha fazla iyileştirme ihtiyacının altını çizmektedir. Gelecekteki araştırmalar, hasta geri bildirimlerini entegre etmeli ve sağlık hizmetlerinde uygulamalarını optimize etmek için bu modellerin gerçek dünyadaki klinik etkilerini değerlendirmelidir.

**Anahtar Kelimeler:** Koroner Bypass Cerrahisi, Yapay Zeka, GPT-3.5, GPT-4, Hasta İletişimi

## INTRODUCTION

Coronary artery disease is one of the leading causes of cardiovascular morbidity and mortality worldwide, contributing to millions of deaths each year (1, 2). Coronary bypass surgery is one of the most effective surgical treatment methods to ensure adequate blood flow to the heart muscle by replacing blocked or narrowed coronary vessels with healthy vessels. However, this surgical procedure carries serious risks and can cause physical, psychological and social difficulties for the patient. Therefore, answering patient questions accurately and clearly before, during and after this surgery is critical for patient satisfaction and treatment success (3-5).

The rapid development of artificial intelligence (AI) technologies has led to significant changes in the field of healthcare. AI-based systems such as big language models have been used in a wide range of applications, from answering patient questions to clinical decision support systems. For example, GPT series models stand out as potential tools to support healthcare professionals with their text generation and natural language processing capabilities (6-8). However, the performance of these systems in terms of accuracy, comprehensibility and clinical relevance has not yet been sufficiently investigated (9-11).

In the literature, there are various studies on the use of big language models in healthcare. For example, Wang et al. (2024) (12), in their study examining the capacity of language models to provide accurate answers to general patient questions about surgical procedures, emphasized that these models are particularly effective for answers containing general health information. However, the adaptability of model performances to specific clinical domains is still unclear and studies on this topic are limited (13-15). There is no study evaluating the performance of large language models for patient questions in coronary bypass surgery. This emphasizes the originality of our study and its contribution to the literature.

The aim of this study is to evaluate the performance of GPT-3.5 and GPT-4 models in answering patient questions about coronary bypass surgery. In the study, the models were compared according to the criteria of accuracy, understandability and clinical relevance. This evaluation, based on expert opinions, provides important data to better understand the potential of large language models in the field of patient communication and clinical support and to contribute to the development of these models.

## MATERIAL AND METHODS

This is a cross-sectional study designed to evaluate the performance of GPT-3.5 and GPT-4 models in responding to patient questions about coronary bypass surgery. The study aims to examine the potential of artificial intelligence models in patient communication and clinical

decision support systems. The study was conducted at Ankara Yıldırım Beyazıt University Bilkent City Hospital Cardiovascular Surgery clinic and 80 Cardiovascular Surgery specialists with at least 5 years of professional experience and expertise in coronary bypass surgery participated in the study. Participants were academicians, clinicians, or professionals working in both positions. Incomplete or incorrectly completed forms and participants with less than 5 years of professional experience were excluded from the analysis. This study received ethical approval from the Ankara Bilkent City Hospital 1st Clinical Research Ethics Committee on 23.11.2024 with decision number TABED 1-24-679. The study was reviewed for ethical considerations and unanimously approved.

During the data collection process, the 10 most frequently asked questions about coronary bypass surgery were determined and these questions were sent to the participants via Google Form. Participants rated the answers provided by the GPT-3.5 and GPT-4 models according to four main criteria: accuracy, which refers to the scientific accuracy of the answer; understandability, which refers to how easily the answer can be understood by the patient; clinical relevance, which refers to the validity of the answer in terms of clinical practice; and overall score, which refers to the overall evaluation of the answer. Each criterion was scored using a Likert scale from 1 (inadequate) to 5 (excellent). The list of common questions asked by patients regarding coronary bypass surgery is provided in Table 1. These questions were utilized to assess the performance of the AI models in providing accurate, understandable, and clinically appropriate responses.

**Statistical Analysis:** Data for this study were analyzed using IBM SPSS Statistics 29 software. During the data preparation and cleaning phase, any records with missing, erroneous, or outlier values were removed from the analysis. Descriptive statistics were utilized to provide fundamental information about the demographics and professional experiences of the participants. This included calculating the distribution of categories and average years of professional experience among participants. Independent t-tests were conducted to evaluate the differences in mean scores between the GPT-3.5 and GPT-4 models. These tests determined whether the differences were statistically significant, with all tests maintaining a significance level of  $p < 0.05$ . Reliability analysis was performed using Cronbach's Alpha coefficient to assess the internal consistency of the measurement tools. Cronbach's Alpha values are interpreted as follows: values below 0.7 indicate acceptable reliability, values between 0.7 and 0.9 indicate good reliability, and values above 0.9 indicate excellent reliability. Effect size analysis was conducted using Cohen's d

to quantify the magnitude of differences between the models on each metric. Cohen's d values are interpreted as follows: values of 0.2 or below

indicate small effects, values around 0.5 indicate medium effects, and values of 0.8 or above indicate large effects.

**Table 1.** Common Questions Asked by Patients Regarding Coronary Bypass Surgery

Question Number	Question
1	Is it absolutely necessary for me to have coronary bypass surgery? Are there alternative treatments?
2	Is this surgery generally successful? What complications might occur during or after the procedure?
3	What are the risks of coronary bypass surgery? Is there any life-threatening danger?
4	How should I prepare before the surgery? What should I pay attention to?
5	Where will you take the veins to replace the blocked arteries? Will it cause other problems in my body?
6	How is coronary bypass surgery performed, and how long does it take?
7	Is the heart connected to a machine during the surgery? Does this procedure have any harm?
8	Will I stay in intensive care after the surgery? How long will I remain in the hospital?
9	How long will it take for me to recover after coronary bypass surgery? When can I return to my daily life?
10	Do I need to make lifestyle changes after the surgery? How should I continue my daily life?

## RESULTS

According to Table 2, 33.75% of the participants are academics, 40.00% are both clinicians and academics, and 26.25% are solely clinicians. The average years of professional experience are 13.7 for academics, 16.5 for clinicians, and 12.9 for those who are both. This diversity in professional backgrounds provides a robust foundation for the comprehensive evaluation of the AI models.

As shown in Table 3, the GPT-4 model significantly outperforms GPT-3.5 across all

primary performance metrics, including accuracy, understandability, clinical suitability, and overall score. The statistical measures, including T-statistics and P-values, indicate substantial differences, suggesting the superior efficacy of the GPT-4 model in handling clinical queries. Table 4 details the reliability of the evaluations, assessed through Cronbach's Alpha. The values obtained suggest high internal consistency across the measurements, with all metrics showing alphas above 0.70, indicating reliable assessments of the models' performances

**Table 2.** Demographics and Professional Experience of Participants by Category

Category	Count	Percentage	Average Years of Experience
Academics	27	33.75%	13.7 years
Both Clinician and Academic	32	40.00%	12.9 years
Clinicians	21	26.25%	16.5 years

**Table 3.** Comparative Performance of GPT-3.5 and GPT-4 Models on Key Performance Metrics

Metric	GPT-3.5 Mean	GPT-4 Mean	T-Statistic	P-Value
Accuracy	1.77	3.02	-9.97	<0.05
Understandability	1.81	2.99	-9.48	<0.05
Clinical Suitability	1.78	2.96	-11.14	<0.05
Overall Score	1.77	2.98	-10.57	<0.05

**Table 4.** Reliability Analysis Across Core Performance Metrics

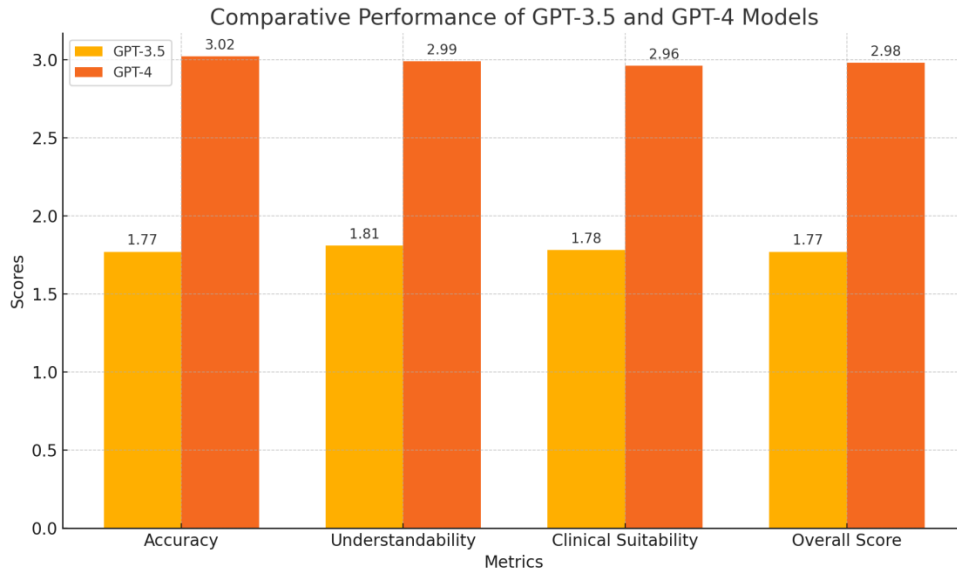
Metric	Cronbach's Alpha
Accuracy	0.72
Understandability	0.69
Clinical Suitability	0.74
Overall Score	0.71

In Table 5, the effect sizes (Cohen's d) are presented, illustrating large effect sizes for all considered metrics. These substantial effect sizes highlight the practical significance of the performance differences between the models, with GPT-4 not only statistically outperforming GPT-3.5 but also showing considerable improvements that are likely to be clinically relevant.

**Table 5.** Effect Size Analysis Between Models Across Principal Performance Metrics

Metric	Cohen's d
Accuracy	1.65
Understandability	1.56
Clinical Suitability	1.54
Overall Score	1.59

The comparative performance of GPT-3.5 and GPT-4 models on key performance metrics, including accuracy, understandability, clinical suitability, and overall score, is illustrated in Figure 1. The figure highlights the significant differences in performance between the two models, with GPT-4 consistently achieving higher scores.

**Figure 1.** Comparative performance of GPT-3.5 and GPT-4 models across key performance metrics, including accuracy, understandability, clinical suitability, and overall score.

## DISCUSSION

This study is one of the first to comparatively evaluate the performance of GPT-3.5 and GPT-4 models in responding to patient questions about coronary bypass surgery based on accuracy, understandability, clinical appropriateness, and overall evaluation criteria. The findings revealed that the GPT-4 model performed significantly better than GPT-3.5 across all metrics. These results align with existing literature on the use of large language models in healthcare and extend the understanding of their potential applications.

Accuracy is a critical metric for AI systems, especially in healthcare. Lewine et al. (2024) noted that GPT-3.5 demonstrated high accuracy in general knowledge but had limitations in clinical contexts (16). In this study, GPT-4 showed significant superiority in accuracy compared to GPT-3.5, particularly in providing specific clinical information about coronary bypass surgery. This can be attributed to GPT-4's updated knowledge base and advanced natural language processing capabilities. However, both models occasionally provided inaccurate or incomplete responses, consistent with Liu et al. (2022), who highlighted the potential for AI systems to falter in complex clinical scenarios (17).

Understandability plays a crucial role in patient communication, as it directly impacts patient engagement and comprehension of medical

procedures. Bajva et al. (2021) (18) emphasized the importance of AI systems using simple and clear language to enhance patient satisfaction. In this study, GPT-4 achieved significantly higher scores in understandability compared to GPT-3.5, likely due to its advanced language generation capabilities that produce more fluent and patient-friendly responses. However, the occasional use of technical jargon, making responses less accessible to patients, aligns with Al Kuwaiti et al. (2023) (19), who argued for further optimization of language models to better cater to patient needs.

Clinical appropriateness extends beyond accuracy, focusing on the relevance and applicability of the information in a clinical context. Maleki et al. (2024) (20) emphasized that AI systems in surgical domains must prioritize clinical appropriateness to be reliable tools for healthcare professionals. In this study, GPT-4 outperformed GPT-3.5 in this metric, demonstrating better alignment with clinical contexts. Nevertheless, some responses were either incomplete or lacked contextual depth, highlighting the need for more specific training data to enhance model performance in niche clinical areas.

The overall evaluation metric combines all individual metrics to provide a comprehensive assessment of the models' performance. GPT-4 scored significantly higher than GPT-3.5, reflecting its superior performance across the other three metrics. As noted by Liu et al. (2024) (21), AI

systems that offer user-friendly and human-like responses can play a vital role in patient communication. However, further improvements are necessary to ensure that AI models meet diverse patient needs comprehensively.

This study's strengths include being the first to evaluate AI models' responses to patient questions about coronary bypass surgery and its reliance on expert evaluations. However, several limitations must be acknowledged. First, the study is based solely on expert opinions, excluding direct feedback from patients. Second, the AI models were evaluated using a specific dataset, which limits the generalizability of the results. Furthermore, the cross-sectional design of the study does not allow for the assessment of the models' performance over time as they continue to evolve.

### CONCLUSIONS

This study demonstrates the potential of AI-based large language models as tools for patient communication in coronary bypass surgery. GPT-4 outperformed GPT-3.5 in accuracy, understandability, clinical appropriateness, and overall evaluation criteria. However, limitations such as occasional inaccuracies and incomplete responses remain evident in both models. Future research should involve larger cohorts of patients and experts, evaluate the models' impact on real-world patient outcomes, and train these systems on more specific clinical datasets. This study represents a significant step forward in exploring the effective use of AI systems in healthcare delivery.

### REFERENCES

1. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. *Journal of the American College of Cardiology*. 2020;76(25):2982-3021.
2. Vogel B, Acevedo M, Appelman Y, Bairey Merz CN, Chieffo A, Figtree GA, et al. The Lancet women and cardiovascular disease Commission: reducing the global burden by 2030. *Lancet (London, England)*. 2021;397(10292):2385-438.
3. Powell R, Scott NW, Manyande A, Bruce J, Vögele C, Byrne-Davis LM, et al. Psychological preparation and postoperative outcomes for adults undergoing surgery under general anaesthesia. *The Cochrane database of systematic reviews*. 2016;2016(5):Cd008646.
4. Açikel MET. Evaluation of Depression and Anxiety in Coronary Artery Bypass Surgery Patients: A Prospective Clinical Study. *Brazilian journal of cardiovascular surgery*. 2019;34(4):389-95.
5. Aburuz ME, Maloh H. Preoperative anxiety and depressive symptoms predicted higher incidence of delirium post coronary artery bypass graft surgery. 2024.
6. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education*. 2023;23(1):689.
7. Yuan M, Bao P, Yuan J, Shen Y, Chen Z, Xie Y, et al. Large language models illuminate a progressive pathway to artificial intelligent healthcare assistant. *Medicine Plus*. 2024:100030.
8. Guo RX, Tian X, Bazoukis G. Application of artificial intelligence in the diagnosis and treatment of cardiac arrhythmia. 2024;47(6):789-801.
9. Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, et al. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information fusion*. 2023;99:101805.
10. Coleman JJ, Owen J, Wright JH, Eells TD, Antle B, McCoy M, et al. Using Artificial Intelligence to Identify Effective Components of Computer-Assisted Cognitive Behavioural Therapy. *Clinical psychology & psychotherapy*. 2024;31(6):e70023.

### Declarations

**Funding:** Not applicable.

**Competing Interests:** The authors declare no competing interests.

**Authors Contributions:** All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Muhammet Fethi Sağlam, Emrah Uguz, Kemal Eşref Erdoğan, Hüseyin Ünsal Erçelik, Murat Yücel, Cevat Ahmet Sert, Fatih Yamaç and Erol Şener. The first draft of the manuscript was written by Muhammet Fethi Sağlam, Emrah Uguz, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

**Ethics Approval:** This study received ethical approval from the 1st Scientific and Ethical Review Committee for Medical Research (TABED) of Ankara Bilkent City Hospital on 23.10.2024, with decision number TABED 1-24-679. The study was reviewed and unanimously approved in terms of ethical considerations.

**Consent to Participate:** Informed consent was obtained from all individual participants included in the study.

**Consent to publish:** Not applicable.

**Availability of Data and Materials:** The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

**Acknowledgements:** Not applicable.

11. Ismail AMA. Chat GPT in Tailoring Individualized Lifestyle-Modification Programs in Metabolic Syndrome: Potentials and Difficulties? *Annals of biomedical engineering*. 2023;51(12):2634-5.
12. Wang D, Zhang S. Large language models in medical and healthcare fields: applications, advances, and challenges. *Artificial Intelligence Review*. 2024;57(11):299.
13. Samant S, Bakhos JJ, Wu W, Zhao S, Kassab GS, Khan B, et al. Artificial Intelligence, Computational Simulations, and Extended Reality in Cardiovascular Interventions. *JACC Cardiovascular interventions*. 2023;16(20):2479-97.
14. Krajcer Z. Artificial Intelligence for Education, Proctoring, and Credentialing in Cardiovascular Medicine. *Texas Heart Institute journal*. 2022;49(2).
15. Biswas SS. Role of Chat GPT in Public Health. *Annals of biomedical engineering*. 2023;51(5):868-9.
16. Levine DM, Tuwani R, Kompa B, Varma A, Finlayson SG, Mehrotra A, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. *The Lancet Digital Health*. 2024;6(8):e555-e61.
17. Liu J, Wang C. Utility of ChatGPT in Clinical Practice. 2023;25:e48568.
18. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. *Future healthcare journal*. 2021;8(2):e188-e94.
19. Al Kuwaiti A, Nazer K, Al-Reedy A, Al-Shehri S, Al-Muhanna A. A Review of the Role of Artificial Intelligence in Healthcare. 2023;13(6).
20. Maleki Varnosfaderani S, Forouzanfar M. The Role of AI in Hospitals and Clinics: Transforming Healthcare in the 21st Century. 2024;11(4).
21. Liu C-L, Ho C-T, Wu T-C, editors. Custom GPTs enhancing performance and evidence compared with GPT-3.5, GPT-4, and GPT-4o? A study on the emergency medicine specialist examination. *Healthcare*; 2024:MDPI.