Journal of Clinical Question, 2024, Vol. 1, No. 3, 82–88
https://doi.org/10.69854/jcq.2024.0011
Advance access publication date 02 November 2024

Journal of Clinical Question
Gleam Publishing

Original Research

Assessing the Ethical and Professional Capabilities of AI: A Study of ChatGPT and Google Gemini versus PREview (Situational Judgement Test) for Medical Student Applicant

Hillary Chu[#], Emily Noelle Pasion[#], Stephanie Yeh, Gary Chu[*]

California Northstate University College of Medicine, CA, USA.

[*]***Corresponding Author:*** e-mail: Gary.Chu@CNSU.edu
[#]Co-first authors

Submitted: October 26, 2024 **Accepted:** November 01, 2024

Clinical Question Box

As artificial intelligence (AI) continues to advance in supporting various aspects of human life, can AI systems like ChatGPT and Google Gemini effectively demonstrate ethical reasoning and professionalism in medicine to assist healthcare practitioners in delivering quality care?

Based on their responses to the PREview exam, developed by the Association of American Medical Colleges (AAMC), ChatGPT and Google Gemini demonstrated competence in ethical reasoning and professionalism relevant to the medical field. Both systems scored above average on the tests, indicating a level of proficiency in these areas.

Abstract

Introduction: Artificial intelligence (AI) is increasingly integrated into healthcare, supporting tasks ranging from administrative functions to clinical decision-making. This study evaluates the ethical reasoning and professionalism of two AI models, ChatGPT and Google Gemini, by examining their responses to the Association of American Medical Colleges' (AAMC) PREview exam—a situational judgment test assessing ethical and professional competencies in healthcare. **Methods:** ChatGPT 4.0 and Google Gemini 1.5 were evaluated using two sets of AAMC PREview scenarios. Each response was rated on a three-point effectiveness scale: "not correct," "partially correct," and "correct." Full points were awarded for responses closely matching the AAMC's ideal answers, while partial points were given for responses that were similar but varied. Statistical significance in performance differences was analyzed using a one-way ANOVA test. **Results:** ChatGPT and Google Gemini were conducted separately on both tests. ChatGPT achieved an accuracy rate of 79.3% on the first test and 77% on the second, while Google Gemini scored 68% and 70%, respectively. A statistically significant difference was observed in Test 1 ($p = 0.002$) but not in Test 2 ($p = 0.101$). Overall, ChatGPT demonstrated a stronger alignment with AAMC's ethical standards than Google Gemini. **Conclusion:** Both AI models exhibited competence in ethical reasoning, with ChatGPT achieving a higher degree of alignment with medical ethics standards. While these models show promise as tools for ethical decision-making, they should complement, not replace, human judgment in complex healthcare contexts.

Keywords: Artificial intelligence, ChatGPT, Google Gemini, medical ethics, PREview exam.

Introduction

Artificial Intelligence (AI) is a user-friendly interface that performs a variety of tasks, provides insights, and generates content. Generative AI has become widely accessible and beneficial in boosting productivity for those across a wide range of disciplines.[1] Fundamentally, scientific research in science, technology, engineering, and mathematics fields can be enhanced as AI is promising in categorizing, predicting, and digesting large datasets.[2] It effectively increases digital accessibility for all, including individuals with disabilities, and supports tailored learning experiences.[3] In healthcare, AI is currently employed for administrative task completion, personalized medicine, preventative care, and diagnostic capabilities.[4] Medical institutions are starting to recognize this paradigm shift in the medical education revolution. In the Fall of 2024, Harvard Medical School was the first to launch an introductory course of generative AI in the healthcare curriculum to equip future doctors with data and machine learning skills to tackle real problems in healthcare. Chang, the dean of medical education advocating for this course, emphasizes the importance of assessing the current limitations of AI in clinical decision-making.[5] Effectively, there is a lack of studies examining AI's responses to situational moral dilemmas within the medical field.

Chat GPT 4.0 and Google Gemini are both commonly utilized AI-powered language models.[6] OpenAI's ChatGPT 4.0 specializes in logical text creation, while Google Gemini under Google DeepMind was designed to be multimodal and versatile in text and image/visual commands.[7] Chat GPT 4.0 is trained through reinforcement learning from human feedback, where the large language model of Chat GPT 3.5 was tuned and taught the different kinds of responses human users prefer instead of its initial tendency to regurgitate information taken from the internet.[8] Recent studies have shown that the recent version of ChatGPT, ChatGPT 4.0, achieves an 81% accuracy rate on medical examinations, which is a significant improvement upon its earlier counterpart ChatGPT 3.5.[9] Google Gemini is trained with a changing dataset as it has the ability to use information from Google Search to respond to the users with more updated information and process real-world information.[10] Due to the slight differences in how they were programmed and their intended purposes, it is essential to assess the quality of their responses in healthcare decision-making and determine how effectively AI can enhance valuable medical decisions.

The Association of American Medical Colleges (AAMC) is a nonprofit organization involved in advancing medical education and advocating for healthcare policy in the United States. It is well known for being an association that administers the Medical College Admission Test (MCAT).[11] However, a new exam has emerged, called the Professional Readiness Exam or the PREview exam, and was piloted in 2020.[12] In contrast, the MCAT covers mainly science, technology, engineering, and mathematics-related questions; the PREview exam is a Situational Judgment Test (SJT). Various professions have utilized SJT since the 1920s to simulate a dilemma and analyze an applicant's problem-solving or leadership aptitudes.[13] Currently, nine medical schools require PREview scores as a component of a candidate's admission to medical school, and over 30 medical schools have elected to view PREview scores to conduct further research.[14] Findings from the 2020–2021 PREview exam reveal a notable degree of overlap in scores amongst candidates of many racial and ethnic backgrounds, supporting the AAMC's goal to assess test takers across all ethnic groups equitably.[15] For incoming medical students, the PREview exam's format as a curated SJT will account for one's familiarity with approaching difficult medical situations effectively.

The PREview multiple-choice exam presents examinees with hypothetical scenarios commonly encountered in healthcare settings. The AAMC provides responses that span scales of effective to ineffective behavior. The scoring key is established by a diverse group of medical professionals, including educators, faculty, and admissions officers. Points are awarded when the test taker's rating aligns with that of a medical educator, with partial credit given for similar but not identical ratings.[16] Our study examines whether AI systems can uphold solid moral and professional values in patient care. Specifically, we aim to compare two standard AI models, ChatGPT and Google Gemini, to evaluate their ability to assess doctor-patient interactions involving critical medical ethics, using a standardized premedical SJT as a benchmark.

Methods

Resources

Two AI models, ChatGPT 4.0 (April 2023) and Google Gemini 1.5 Flash (September 2024), were evaluated using situational judgment questions from the AAMC PREview® preparation resources. These questions are designed to assess essential competencies for healthcare professionals, such as ethical reasoning, professionalism, empathy, and clinical decision-making. The PREview assessment included two tests, each containing 186 questions.

Scoring Criteria

The responses from each AI model were compared with the AAMC answer key, and scores were assigned based on alignment: 1 Full Point was awarded if the AI's rating matched the ideal AAMC rating; 0.5 Points (Partial Credit) were given if the response aligned with the general effectiveness (effective vs. ineffective) but differed on the specific scale; 0 Points were assigned if the response did not match either the effectiveness or the specific rating scale. This scoring approach quantified each model's adherence to medical ethics standards, generating an overall performance score of 372 points per model.

Procedure

ChatGPT 4.0 and Google Gemini 1.5 Flash were evaluated in two distinct PREview questionnaire tests. For each AI system, scores were calculated independently and then compared to assess their performance on key competencies.

Statistical Analysis

A one-way ANOVA was used to compare the average accuracy rates of ChatGPT 4.0 and Google Gemini 1.5 Flash across the scenarios. This statistical test was chosen to evaluate the mean differences between two independent groups. A significance level of 0.05 was applied, with p-values reported for each comparison to determine if performance differences between the AI models were statistically significant. Both models were tested under controlled and consistent conditions to ensure reliable findings, with identical question inputs and scoring criteria.

Results

Descriptive statistics for ChatGPT and Google Gemini scores on the two tests are presented in Table 1. In PREview Test 1, ChatGPT achieved an accuracy rate of 79.3%, while Google Gemini scored 68.0%. In Test 2, ChatGPT's accuracy slightly decreased to 77.2% while Google Gemini's improved to 71.2%. A one-way ANOVA revealed a statistically significant difference in performance between ChatGPT and Google Gemini in Test 1 ($p = 0.002$), indicating that ChatGPT's responses are more closely aligned with AAMC standards. However, the difference in performance in Test 2 was not statistically significant ($p = 0.101$), suggesting comparable efficacy in that test.

As shown in Fig. 1, the average performance across both tests indicates that ChatGPT had an overall accuracy rate of 78.2% with a 95% confidence interval (CI) of 74.4–82.1, while Google Gemini had an overall accuracy rate of 70.0% (95% CI: 63.8–75.5). An ANOVA test confirmed that the overall difference in accuracy rates across both tests was statistically significant ($p < 0.001$). However, some overlap in the 95% confidence intervals—specifically between 74.3% and 75.5%—suggests that while ChatGPT generally outperformed Google Gemini, certain scenarios might yield similar results from both AI models.

In Table 2, ChatGPT had a higher percentage of correct answers than Google Gemini, with ChatGPT achieving full points on 66.4% (247/372) of questions across both exams, compared to Google Gemini's 53.5% (199/372). ChatGPT scored 23.7% (88/372) partially correct answers and 9.9% (37/372) incorrect answers, which were lower than Google Gemini's 32.2% (120/372) partially correct and 14.2% (53/372) incorrect answers.

Table 1. Descriptive statistics of ChatGPT and Google Gemini scores on Test 1 and Test 2

Tests	Test 1		Test 2	
Items	ChatGPT	Gemini	ChatGPT	Gemini
Total score	147.5	126.5	143.5	132.5
Max score possible	186	186	186	186
% of max score	79.3%	68.0%	77.2%	71.2%
Total number of questions	186	186	186	186
p-value[a]	0.002		0.101	
Combined p-value	<0.001			

Note: a: The ANOVA test was used to compare the scores of the two AI systems.

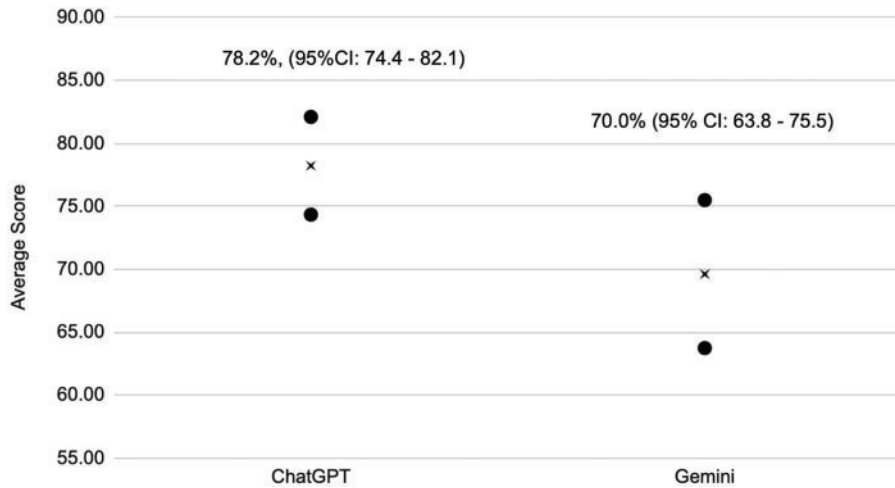

Figure 1. Overall performance of ChatGPT and Google Gemini in combined Test 1 and Test 2.

Table 2. Tallies for each question on the PREview exam that were either correct, partially correct, or incorrect

Items	ChatGPT	Gemini
Combined correct (%)	247/372 (66.4%)	199/372 (53.5%)
Combined partially correct (%)	88/372 (23.7%)	120/372 (32.2%)
Combined incorrect (%)	37/372 (9.9%)	53/372 (14.2%)

Discussion

Our data indicate that both ChatGPT and Google Gemini demonstrate sufficient adherence to ethical principles in healthcare. The lowest score among both AIs on two practice PREview exams was 68% accuracy compared to the AAMC Answer Key, equating to a score range of roughly 6–8 out of 9 on the PREview exam. This suggests that ChatGPT and Google Gemini are adept at assessing complex ethical issues. The PREview exam is graded on a scale of 1–9. Participants earn a single number between one (lowest) and nine (highest), as well as their percentile ranking amongst applicants who took the exam from the two previous testing years.[17,18] The scoring algorithm is not made public, so the percent of the correct answers may not directly translate to the number score received but should align accordingly when standardized. The purpose of this examination is to evaluate the ethical and

professional attributes of incoming medical students, including cultural awareness, humility, empathy, compassion, ethical responsibility, resilience, and adaptability.[19]

While this study provides insights into the ethical alignment of AI responses in medical contexts, limitations include inherent differences in model training and the lack of real-time contextual adaptability in AI responses. These findings provide preliminary evidence of AI models' ethical capacities but emphasize the need for professional oversight in critical clinical decision-making. Traditional methods of reviewing medical school applicants tend to prioritize academic excellence while overlooking cognitive awareness of biases and moral fairness.[20] Factors like grade point average and MCAT scores are weighted heavily in the application process, with less weight given to understanding of proper situational conduct. The AAMC aimed to balance academic and ethical merits in the medical school application process by creating the PREviewexam. This judgment-based examination is incorporated into a student's portfolio for medical school to encourage a holistic review process that includes measures of preprofessional competency.[15]

The PREview exam reflects American ideals of professionalism in medicine. Using a standardized SJT designed for prospective medical students, ChatGPT outperformed Google Gemini. Given that the AAMC collaborates with practicing medical educators to set scoring rubrics for the PREview exam, there is some evidence that AI's decision-making aligns with standards established in the medical field. Future research could involve other AI models developed in different countries and cultures, such as Baidu ERNIE Bot from China, YandexGPT from Russia, and Naver HyperCLOVA from South Korea, as these AI systems are trained differently and may offer diverse responses based on varying cultural values.[21–23]

There are several limitations in this study. One limitation of this study is the absence of a conversion table to translate test-taker percentiles into corresponding accuracy percentages accurately. Since PREview exams are standardized, scores are relative to examinees' performance in recent years. Additionally, there is no conclusive evidence linking strong PREview exam performance to the development of competent physicians, as the first cohort of medical students who participated in the exam pilot has not yet graduated. Further research would be beneficial to explore how well the AAMC's PREview exam predicts the development of well-rounded physicians.

Conclusion

AI appears capable of grasping ethical values that align with principles upheld by medical professionals. ChatGPT and Google Gemini can serve as resources to assist with questions of moral ethics. When interpreted correctly, AI can be a powerful tool for certain decision-making processes in healthcare; however, medical professionals should avoid complete reliance on programs like ChatGPT and Google Gemini until these systems are further refined and made free from bias. While not infallible, ChatGPT and Google Gemini demonstrate a basic understanding of behaviors that may contribute to effective practices in the medical field.

Acknowledgments

None.

Funding Source

No financial support was provided.

Author Contributions

H.C. and E.N.P. are acknowledged as co-first authors for their contributions to drafting the manuscript, which they primarily authored and which was substantially revised by G.C. The concept and design were developed by G.C., while data acquisition, analysis, and interpretation were performed by H.C., E.N.P., and S.Y. Overall supervision of the project was provided by G.C.

All authors have reviewed the final version of this research article and agree to take responsibility for all aspects of the work.

Data Availability Statement

The datasets used in the current study are available from the corresponding author upon reasonable request.

Ethical Statement

The article does not involve the participation of any animals. The Institutional Review Board approval was waived because this study did not include human data.

Conflicts of Interest

The authors report no conflicts of interest in this work.

References

[1] Dwivedi YK, Kshetri N, Hughes L, et al. Opinion paper: so what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int J Inf Manag*. August 1, 2023;71(2):102642. doi:10.1016/j.ijinfomgt.2023.102642.

[2] Xu Y, Liu X, Cao X, et al. Artificial intelligence: a powerful paradigm for scientific research. *Innovation (Camb)*. November 28, 2021;2(4):100179. doi:10.1016/j.xinn.2021.100179.

[3] Chemnad K, Othman A. Digital accessibility in the era of artificial intelligence-Bibliometric analysis and systematic review. *Front Artif Intell*. 2024;7:1349668. doi:10.3389/frai.2024.1349668.

[4] Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ*. September 22, 2023;23(1):689. doi:10.1186/s12909-023-04698-z.

[5] Gehrman E. How generative AI is transforming medical education. 2024. Accessed October 28, 2024. https://magazine.hms.harvard.edu/articles/how-generative-ai-transforming-medical-education

[6] Carlà MM, Giannuzzi F, Boselli F, et al. Testing the power of Google DeepMind: gemini versus ChatGPT 4 facing a European ophthalmology examination. *AJO Int*. October 3, 2024;1(3):100063. doi:10.1016/j.ajoint.2024.100063.

[7] Google DeepMind. Gemini models. 2024. Accessed October 28, 2024. https://deepmind.google/technologies/gemini/

[8] Open AI. How ChatGPT and our foundation models are developed. 2024. Accessed October 28, 2024. https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed

[9] Liu M, Okuhara T, Chang X, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. *J Med Internet Res*. July 25, 2024;26:e60807. doi:10.2196/60807.

[10] Imran M, Almusharraf N. Google Gemini as a next generation AI educational tool: a review of emerging educational technology. *Smart Learn Environ*. May 23, 2024;11(1):22. doi:10.1186/s40561-024-00310-z.

[11] Association of American Medical Colleges. Changing the MCAT® exam. 2024. Accessed October 28, 2024. https://students-residents.aamc.org/about-mcat-exam/changing-mcat-exam

[12] Association of American Medical Colleges. AAMC PREVIEW® professional readiness exam research. 2024. Accessed October 28, 2024. https://www.aamc.org/services/admissions-lifecycle/aamc-preview-professional-readiness-exam-research

[13] Lievens F, Motowidlo SJ. Situational judgment tests: from measures of situational judgment to measures of general domain knowledge. *Indus Organizat Psychol*. 2016;9(1):3–22. doi:10.1017/iop.2015.71.

[14] Association of American Medical Colleges. Medical schools participating in the AAMC PREVIEW® exam. 2024. Accessed October 28, 2024. https://students-residents.aamc.org/aamc-preview/participating-medical-schools

[15] Ellison HB, Grabowski CJ, Schmude M, et al. Evaluating a situational judgment test for use in medical school admissions: two years of AAMC PREview exam administration data. *Acad Med*. February 1, 2024;1(2):183–191. doi:10.1097/ACM.0000000000005548.

[16] Association of American Medical Colleges. AAMC PREview scores. 2024. Accessed October 28, 2024. https://students-residents.aamc.org/aamc-preview/aamc-preview-scores

[17] Association of American Medical Colleges. Summary of AAMC PREviewprofessional readiness exam scores. 2024. Accessed October 28, 2024. https://students-residents.aamc.org/media/15831/download

[18] Association of American Medical Colleges. Summary of AAMC PREviewprofessional readiness exam scores (formerly situational judgment test). 2023. Accessed October 28, 2024. https://students-residents.aamc.org/media/14531/download?attachment

[19] Association of American Medical Colleges. The premed competencies for entering medical students. 2024. Accessed October 28, 2024. https://students-residents.aamc.org/real-stories-demonstrating-premed-competencies/premed-competencies-entering-medical-students

[20] Conrad SS, Addams AN, Young GH. Holistic review in medical school admissions and selection: a strategic, mission-driven response to shifting societal needs. *Acad Med*. November 2016;91(11):1472–1474. doi:10.1097/ACM.0000000000001403.

[21] He W, Zhang W, Jin Y, et al. Physician versus large language model chatbot responses to web-based questions from autistic patients in chinese: cross-sectional comparative analysis. *J Med Internet Res*. April 30, 2024;26:e54706. doi:10.2196/54706.

[22] Wikimedia. YandexGPT. 2024. Accessed October 28, 2024. https://en.wikipedia.org/wiki/YandexGPT

[23] CLOVA. HyperCOLVAX. 2024. Accessed October 28, 2024. https://clova.ai/en/hyperclova