Original Paper

Large Language Models for Structured Critical Appraisal of Randomized Controlled Trials in Otolaryngology: Randomized Crossover Study Comparing GPT-4o and Medical Students

Sofia Olsson[1], BS, MD; Nicholas Del Mundo[2], BS, MD; Raika Bourmand[3], BSc, MSc; Fiza Baloch[3], BS; Jonathan Liang[2], MD; Nancy Jiang[4], MD

[1]Department of Otolaryngology - Head and Neck Surgery, University of California Davis Medical Center, Sacramento, CA, United States
[2]Department of Otolaryngology - Head and Neck Surgery, Kaiser Permanente Oakland Medical Center, Oakland, CA, United States
[3]Anne Burnett Marion School of Medicine at Texas Christian University, Fort Worth, TX, United States
[4]Division of Laryngology, Department of Otolaryngology - Head and Neck Surgery, Stanford Medicine, Palo Alto, CA, United States

Corresponding Author:

Sofia Olsson, BS, MD
Department of Otolaryngology - Head and Neck Surgery
University of California Davis Medical Center
2521 Stockton Blvd, Ste 7200
Sacramento, CA 95817
United States
Phone: 1 6509193727
Email: seolsson@health.ucdavis.edu

Abstract

Background: Critical appraisal of medical literature is an essential skill for trainees but can be time-intensive and difficult to master. Large language models (LLMs) such as ChatGPT have emerged as potential tools to support learners; however, their accuracy and reliability in structured appraisal tasks remain unclear.

Objective: The objective of this study was to compare the performance of GPT-4o with that of medical students in structured critical appraisal of randomized controlled trials (RCTs), evaluating its potential role in medical education.

Methods: In this prospective randomized crossover study, 12 RCTs published in *The Laryngoscope* in 2024 were identified and appraised using the Critical Appraisal Skills Programme (CASP) checklist. Two second-year medical students independently appraised each article with and without the assistance of GPT-4o, separated by a 1-month washout period. The GPT-4o model was also used to independently appraise each study. Two otolaryngologists evaluated the randomized and anonymized appraisals using a standardized rubric assessing binary correctness, free-text accuracy, completeness, and overall quality. Medical students and evaluators also completed poststudy questionnaires regarding AI use and perceptions.

Results: A statistically significant overall effect of appraisal condition was observed for free-text accuracy ($F_{2,22}$=3.94; P=.04; η^2p=0.263); however, no individual pairwise comparison remained statistically significant following Holm adjustment for multiple comparisons. No statistically significant effect of appraisal condition was detected for binary correctness ($F_{2,22}$=0.07; P=.93, η^2p=.006), completeness ($F_{2,22}$=0.18; P=.84; η^2p=0.016), or overall quality ($F_{2,22}$=0.52; P=.60; η^2p=0.045). Physician evaluators identified the presence of AI involvement in approximately 56% of student appraisals completed with AI assistance and 58% of AI-only appraisals. Medical students reported that GPT-4o reduced the time required to complete appraisals and assisted with understanding study topics, interpreting results, and identifying study limitations.

Conclusions: Although a statistically significant overall effect of appraisal condition was observed for free-text accuracy, no individual pairwise comparison remained statistically significant following Holm adjustment, with the largest observed difference occurring between students without AI assistance and AI alone. No statistically significant differences were detected in binary correctness, completeness, or overall quality. These findings support cautious investigation of GPT-4o as an adjunct educational tool for structured critical appraisal while emphasizing the continued importance of independent critical reasoning and human oversight. Larger studies are needed to determine the generalizability of these findings.

JMIR Form Res 2026;10:e99015; doi: 10.2196/99015

Keywords: critical appraisal; literature appraisal; medical education; otolaryngology; large language models

Introduction

Critical appraisal of scientific literature is a vital skill in the practice of evidence-based medicine, which physicians develop over time. With increased exposure to scientific literature throughout their medical education, physicians in training are expected to confidently assess the trustworthiness and methodological integrity of a study and its findings [1]. Such appraisals address a study's credibility and quality of the evidence presented based on multiple factors including risk of bias, consistency of results, and precision of the measurements taken [2]. "Appraisal of the credibility of information sources" is a medical school accreditation requirement enforced by the United States Liaison Committee on Medical Education (LCME); however, there is no standardized curriculum recommended to teach such competencies [3]. Systematic reviews from 2001 and 2009 suggested that critical appraisal skills can be improved with training [4-6].

Current literature on the practice of critical appraisal is limited, with an absence of studies evaluating the ability of AI, specifically large language models (LLMs), to perform such skills or enhance medical education in this area. With the rapid democratization of AI technology via free and accessible software such as ChatGPT, these tools have become increasingly common in various aspects of medical education. For example, a 2023 study found that most residency applicants were interested in using AI throughout residency [7], and a 2024 study discovered that nearly half of US medical students studied with the assistance of AI tools [8]. Additionally, a 2024 study by Ganjavi et al [9] found that 96% of medical students had heard of ChatGPT and 52% endorsed using it for medical school coursework, a number that has likely increased alongside the increasing prevalence of AI use globally.

The potential for LLM involvement in review processes for medical literature has been discussed but not extensively studied [10,11]. The present study seeks to describe the ability of AI technology to appraise literature and explore its potential role in medical education. Specifically, the objective of this study is to compare medical students, medical students with AI, and AI alone in their ability to appraise medical literature in otolaryngology using a standardized critical appraisal tool.

Methods

Ethical Considerations

This study was approved by the Texas Christian University Institutional Review Board (IRB; 2025-181) and determined to meet criteria for exempt research under 2 exemption categories. The study involved normal educational practices conducted within an established educational setting, and disclosure of participants' responses outside the research was not reasonably expected to place participants at risk of criminal or civil liability or adversely affect their financial standing, employability, educational advancement, or reputation. Accordingly, the IRB determined that informed consent was not required. All study data were deidentified prior to analysis. No personally identifiable information was included in the study dataset, and data were stored securely with access limited to the study team. The participating medical students did not receive financial compensation for their participation.

Data Collection

This prospective randomized crossover study compared structured critical appraisals of randomized controlled trials (RCTs) completed by medical students with AI assistance, medical students without AI assistance, and ChatGPT (GPT-4o; OpenAI) alone. The search term "randomized controlled trial" was queried in *The Laryngoscope*, an otolaryngology journal, with parameters set to limit results to articles published between January 1, 2024, and December 31, 2024. The journal, *The Laryngoscope*, was chosen due to its high-impact factor, relevance to our field of interest, reputation within otolaryngology, and quality of studies within the journal [12]. This search yielded 163 studies; after excluding those whose abstracts did not describe an RCT, 12 remained for further evaluation. RCTs were selected because they represent a high level of evidence for therapeutic interventions and are specifically aligned with the Critical Appraisal Skills Programme (CASP) RCT checklist used in this study. Full-text articles were accessed through our institution's library subscriptions or were publicly available through open-access publication, as applicable, and were uploaded to ChatGPT solely for the purpose of study-related critical appraisal.

Medical Student Appraisal

Two second-year medical students were randomly assigned to either use AI in their appraisal or to perform the appraisal on their own. The participating students were recruited from a second-year medical school class and volunteered to participate in the study. Second-year medical students were selected because they had completed the preclinical evidence-based medicine curriculum, which includes instruction in interpreting study design, critically appraising primary literature, and evaluating evidence for clinical decision-making. This level of training is consistent with LCME accreditation standards requiring instruction in the scientific method and the development of evidence-based critical judgment [3]. They were instructed to use ChatGPT as they found most helpful to best represent the software most likely to be used by medical students given its widespread popularity and accessibility [13]. Two students were selected to permit each participant to complete repeated appraisals across all 12 RCTs under both study conditions, thereby enabling

within-participant comparison while maintaining a feasible appraisal workload.

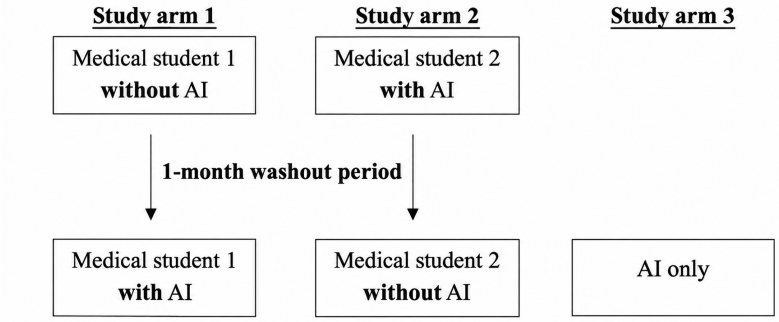
Over 2 weeks, the 2 students each performed critical appraisals of the 12 articles following a standardized checklist developed by the CASP for RCTs [14]. We determined it necessary to structure appraisals with the CASP checklist to optimize data extraction and more objectively evaluate the components that we expected to be included in a literary appraisal [15]. Structured literary appraisals have also been suggested to provide more thorough and critical evaluations of literature [16-18]. The CASP checklist was chosen due to its specific design for the appraisal of RCTs, its combination of binary and free-response questions, and its free availability and accessibility for students to use online [14]. After the students appraised the articles, they did not access them for a 1-month period. The purpose of this "washout period" was to allow students time to forget details of the studies they appraised, reducing the potential for memory bias during the second round of appraisal. One month was determined to be an adequate amount of time based on published retention periods [19-21]. After the 1-month "washout period," the 2 students appraised the 12 articles again with AI-use conditions opposite to those previously assigned.

An additional round of critical appraisal was performed by AI without any human input, other than a researcher inserting the prompts into the software. This researcher was not involved in the evaluation of the critical appraisals. ChatGPT with GPT-4o was chosen as the AI tool in this study due to its free availability and accessibility to medical students. It is also the most used AI tool worldwide with an estimated 4.7 billion monthly visits [22]. Each research article was uploaded to ChatGPT alongside a blank CASP questionnaire and the following prompt:

> *A medical journal is looking to appraise published research for the scientific community at the level of a scientific journal editor. Complete, in detail, the entire questionnaire available in the attached 'CASP' checklist for the attached research article published by The Laryngoscope. Ensure answers to the 'considerations' recommended in the CASP checklist.*

This prompt was thoughtfully designed based on published literature on optimizing AI inputs [23,24]. The order in which articles were appraised was randomized for each appraiser (student-authored, student with AI, and AI-generated) and in each round of appraisal to reduce sequence bias. See Figure 1 for a visual representation of the 3 arms observed in this study.

Figure 1. Visual representation of the 3 research arms performing critical appraisals in this study.



Appraisal Evaluation

All critical appraisals (5 per article; Figure 1) were randomly organized, and the formatting was made consistent to prevent bias from variable text size, spacing, and bolding. No changes to word choice or content were made in this process. One board-certified, fellowship-trained academic otolaryngologist and 1 third-year otolaryngology resident, both with expertise in clinical research, evidence-based medicine, and critical appraisal of the medical literature, were selected as expert evaluators. Both have extensive experience conducting and publishing clinical research, serving as peer reviewers for otolaryngology journals, and evaluating scientific evidence as part of their academic and clinical practice. They began by independently completing CASP checklists for each article, then discussed any differences in answers to ultimately create 1 master key for each study with a consensus on the best and most accurate answer. Then, they independently evaluated each article using the master key for reference. As demonstrated in Checklist 1, every CASP question was evaluated based on correctness of binary responses, correctness of text-based responses (1 indicating no inaccuracies and 4 indicating multiple major inaccuracies), completeness (1 indicating no details missing and 4 indicating that major points for critical assessment were missing), and overall quality (1 representing a failure to critically appraise the study and 5 representing an exceptional appraisal). In addition to these outcome ratings, each evaluator independently selected the strongest appraisal for each article and classified the perceived source of each appraisal as AI-generated, student-authored, or student-authored with AI assistance. These selections were used to determine the frequency with which each appraisal condition was selected as the strongest and the frequency with which evaluators correctly identified AI and student involvement.

Perception of AI Assistance

Following completion of all appraisal exercises, the participating medical students completed a 7-item electronic questionnaire administered through Google Forms. The survey assessed prior familiarity with generative AI models, previous AI use cases, self-reported confidence in critically appraising medical literature, the perceived impact of AI on appraisal time and appraisal quality, intentions regarding future AI use for critical appraisal, and the aspects of the appraisal process in which AI was perceived to be most helpful. Responses included multiple-choice, checkbox, and 5-point Likert scale items, with optional free-text responses where applicable.

Following completion of the appraisal evaluations, the 2 otolaryngologists completed a 3-item electronic questionnaire administered through Google Forms. The survey explored the characteristics that increased or decreased their suspicion of AI involvement and the factors considered when selecting the strongest appraisal and assigning overall quality ratings. Responses consisted of checkbox items with optional free-text responses. The complete survey instruments are provided in Multimedia Appendices 1 and 2, respectively.

Statistical Analysis

Microsoft Excel (version 16.103; 25110922) was used for descriptive statistical analyses in this study [25]. Descriptive statistics, including means, SDs, percentages, and frequencies, were calculated for study outcomes. Evaluator classifications of perceived AI involvement and selections of the strongest appraisal were summarized descriptively using frequencies and percentages.

To account for the repeated and nonindependent structure of the data, inferential analyses were performed at the article level. For each of the 12 articles, scores within each appraisal condition were averaged across the 2 physician evaluators and, for the student conditions, across the 2 student appraisers, thereby preventing repeated ratings from the same students and evaluators from being treated as independent observations. This approach was applied to binary correctness, free-text accuracy, completeness, and overall quality. Binary correctness was summarized for each article and appraisal condition as the percentage of binary CASP items scored correctly. The resulting article-level summary scores for the 3 appraisal conditions were compared using 1-way repeated-measures ANOVA, with the article serving as the repeated unit. Partial eta squared (η^2p) was calculated as the effect size for each omnibus ANOVA. When a statistically significant omnibus condition effect was identified, paired post hoc comparisons were performed with Holm adjustment for multiple comparisons, with Cohen d_z calculated as the effect size for paired comparisons. Statistical significance was defined as $P<.05$.

Results

Physicians chose which appraiser performed the strongest overall CASP appraisal for each study with an interrater agreement of 83%. Students with AI assistance were selected as the strongest appraisers 75% (18/24) of the time, students without AI were selected as the strongest 12.5% (n=3) of the time, and AI alone was selected 12.5% (n=3) of the time. According to the evaluators, the appraisals determined to be the strongest had greater detail in free responses, increased accuracy, and more thorough and comprehensive responses.

As demonstrated in Table 1, no statistically significant effect of appraisal condition was observed for binary correctness ($F_{2,22}$=0.07, P=.93; η^2p=0.006). Article-level mean correctness was 69.7% (SD 8.4%) for students with AI assistance, 69.6% (SD 8.3%) for students without AI assistance, and 70.8% (SD 15.0%) for AI alone.

Table 1. Comparison of correctness, accuracy, completeness, and overall quality of critical appraisals performed by students with AI assistance, students without AI assistance, and AI alone.

Outcome	Students with AI, mean (SD)	Students without AI, mean (SD)	AI alone, mean (SD)	*F* test (*df*)	*P* value[a]	Partial η^2
Binary correctness (%)[b]	69.7 (8.4)	69.6 (8.3)	70.8 (15.0)	0.07 (2, 22)	.93	0.006
Free-text accuracy[c]	1.45 (0.36)	1.37 (0.15)	1.85 (0.65)	3.94 (2, 22)	.04	0.263
Completeness[c]	1.74 (0.64)	1.69 (0.21)	1.81 (0.49)	0.18 (2, 22)	.84	0.016
Overall quality[d]	3.19 (1.38)	2.96 (0.53)	2.71 (1.34)	0.52 (2, 22)	.60	0.045

[a]*P* values were determined using 1-way repeated-measures ANOVA, with each article serving as the repeated unit.
[b]Binary correctness is expressed as the percentage of correctly answered questions, with higher percentages indicating better performance.
[c]Free-text accuracy and completeness are rated on scales from 1 to 4, with lower scores indicating greater accuracy and completeness.
[d]Overall quality is rated on a scale from 1 to 5, with higher scores indicating better appraisal quality.

A statistically significant overall effect of appraisal condition was observed for free-text accuracy ($F_{2,22}$=3.94; P=.04; η^2p=0.263). Mean accuracy ratings were 1.45 (SD 0.36) for students with AI assistance, 1.37 (SD 0.15) for students without AI assistance, and 1.85 (SD 0.65) for AI alone, with lower scores representing greater accuracy. However, no individual pairwise comparison remained statistically significant following Holm adjustment: students with AI assistance versus students without AI assistance (t_{11}=0.69; adjusted P=.506; Cohen d_z=0.20), students with AI assistance versus AI alone (t_{11}=–1.93; adjusted P=.158; Cohen d_z=–0.56), and students without AI assistance versus AI alone (t_{11}=–2.52; adjusted P=.086; Cohen d_z=–0.73). The largest observed difference was between students without AI assistance and AI alone, although this comparison did not reach statistical significance after adjustment for multiple

comparisons. No statistically significant effect of appraisal condition was detected for completeness ($F_{2,22}$=0.18; P=.84; η^2p=0.016), or overall quality ($F_{2,22}$=0.52; P=.60; η^2p=0.045).

Evaluators indicated that signs decreasing their suspicion of AI use included less detail in responses and the absence of sentence structure or wording that they identify with AI-generated content. They also indicated that signs that increased their suspicion of AI use included increased detail in responses, longer responses, and the presence of wording or sentence structure that they associated with AI-generated content. Physician evaluators identified AI involvement in 56% of student appraisals completed with AI assistance and 58% of AI-only appraisals. Interrater agreement was 75% for student appraisals completed with AI assistance, 92% for student appraisals completed without AI assistance, and 58% for AI-only appraisals (Table 2).

Table 2. Physician evaluators' ability to identify student and AI involvement in critical appraisals[a].

Evaluator perception of AI involvement	Students with AI	Students without AI	AI alone
Guessed correctly[b]	0.25	0.40	0.13
Thought no AI was used	0.44	0.40	0.42
Thought student performed CASP[c] with AI assistance	0.25	0.50	0.46
Thought AI was used alone	0.31	0.06	0.13
Identified any form of student involvement[d]	0.69	0.90	0.88
Identified any form of AI involvement[e]	0.56	0.56	0.58
Interrater agreement[f]	0.75	0.92	0.58

[a]Values represent proportions ranging from 0 to 1 (1.00=100%). For evaluator perceptions, proportions were calculated by dividing the number of classifications meeting each criterion by the total number of evaluator classifications within the corresponding actual appraisal condition.
[b]Accurate identification of the exact appraisal condition (students with AI, students without AI, or GPT-4o alone).
[c]CASP: Critical Appraisal Skills Programme.
[d]Includes classifications of either student only or student with AI appraisals.
[e]Includes classifications of either student with AI or AI-only appraisals, regardless of whether the exact condition was correctly identified.
[f]Proportion of appraisals for which both evaluators independently assigned the same classification.

When asked about the level of AI involvement they believed was present in each CASP, evaluators were correct between 13% and 40% of the time.

The 2 medical student appraisers both endorsed experience with ChatGPT; 1 reported experience with Perplexity (Perplexity AI, Inc), and 1 reported experience with Gemini (Google AI). They also reported previous use of AI to generate writing, edit existing writing, use AI as a search engine, summarize writing, and clarify confusing topics. The students reported neutral and medium confidence in their critical appraisal skills (ranked 3 and 4 on a scale of 5). Both appraisers reported that AI assistance decreased the amount of time it took to complete their CASP appraisals; however, they disagreed on the perceived quality of their appraisals. One student reported slightly worse quality of their appraisal, while the other appraiser reported much better perceived quality with AI assistance. One student reported being unsure of their likelihood of using AI in future critical appraisal, while the other reported that they would definitely use it. The students self-reported that AI helped them understand the topic of the article, interpret results, identify study limitations, and improve their confidence in the overall appraisal.

Discussion

Principal Findings and Interpretation

No statistically significant differences among the 3 appraisal conditions were detected in binary correctness, completeness, or overall quality. Although a statistically significant overall effect of appraisal condition was observed for free-text accuracy, no individual pairwise comparison remained statistically significant following Holm adjustment. The largest observed difference was between students without AI assistance and AI alone, with AI-only appraisals receiving less favorable accuracy ratings. This finding suggests a potential limitation of AI-generated free-text responses, although the lack of significant adjusted pairwise comparisons precludes definitive conclusions regarding differences between individual conditions. Previous studies demonstrated that AI models can perform comparably to or outperform medical students on board-style multiple-choice questions [26-30]. However, AI-generated content remains susceptible to misinformation and fabricated references [31, 32], underscoring the importance of human oversight when interpreting medical literature.

Another subsection of this study evaluated student perceptions of AI as a tool in critical appraisal as well as evaluator perceptions of indicators that an appraisal may be AI-generated. Students reported that AI assistance decreased the total amount of time needed for them to conduct an appraisal, though they disagreed on whether it improved or worsened the quality of their appraisals. This perception aligns with our finding that CASP appraisals showed no significant difference in perceived quality between the 3 study arms, though importantly, AI-generated appraisals were potentially inferior in their correctness of generated texts. This difference may reflect the greater cognitive demands of free-text appraisal, which requires synthesis, interpretation, and justification rather than selection of predefined responses.

These higher-order reasoning tasks may be more susceptible to unsupported inferences or inaccurate reasoning generated by current LLMs. A 2021 study evaluating 3133 medical and dental students' views on AI identified that the majority believed that AI technology should act as a partner in medical care and had the potential to revolutionize medicine [33]. Additionally, a 2025 study of 702 medical students revealed that most students had a favorable attitude toward AI, believing it to be an effective learning tool in medical education, which optimized their study time and helped them better grasp new material [34]. With most trainees endorsing the continued use and exploration of AI in medicine, the technology's role in critical appraisal of scientific literature is an important yet understudied topic.

Otolaryngology evaluators in this study reported that longer, more detailed responses with wording they associate with AI were clues they used to distinguish the presence of AI use. However, AI involvement was correctly identified only approximately half of the time. This aligns with previous studies demonstrating that when asked to compare human content to AI-generated content, the use of AI was often misidentified or unidentified [35]. Similar findings have demonstrated that advanced AI detection tools inconsistently and incorrectly identify content as human versus AI-generated [36]. Despite this, a 2024 study found evidence of distinctive vocabulary and paragraph structures in AI-generated texts compared with human-authored texts [37]. Human detection rates of AI-generated content have yet to be studied, though most medical publishers have clear policies on the use of generative AI including no AI authorship, mandatory disclosure of AI use, and required human oversight of research projects [38-40].

Implications

With the increasing prevalence and democratization of AI technology, we are continuously evaluating the strengths and weaknesses of its involvement in medicine. In the setting of literature review and editing, AI has been found to improve manuscript processing times, reduce the frequency of grammatical errors, and increase global accessibility of scholarly articles [41]. The idea of leveraging AI for rapid manuscript review has been introduced, though not thoroughly studied, and must be considered with caution [42,43]. Given the results of this study, we support cautious implementation of AI technology in medical student education on critical appraisal to improve the understanding of published literature and confidence in student appraisals. Critical appraisal is a foundational competency of evidence-based medicine and underpins clinicians' abilities to interpret emerging evidence and apply it appropriately to patient care. AI may reduce the logistical burden associated with learning structured appraisal, potentially allowing trainees to devote more effort to understanding study methodology and clinical interpretation. However, reliance on inaccurate AI-generated reasoning could undermine the development of independent appraisal skills if outputs are accepted without critical evaluation. These findings also raise the possibility of automation bias, whereby learners may place undue trust in AI-generated outputs despite inaccuracies. Educational strategies should therefore emphasize verification of AI-generated appraisals and encourage AI to function as a supplement to, rather than a replacement for, independent critical reasoning. Collectively, these findings suggest that current LLMs may be most appropriately used to support structured components of critical appraisal while preserving human oversight for nuanced interpretation and clinical reasoning.

Strengths and Limitations

The present study assessed the ability of AI to appraise medical literature and discussed its potential role in related medical education. The evaluations of CASP appraisals performed by 2 otolaryngologists using predetermined grading scales helped facilitate an objective review of qualitative CASP responses with minimal bias. Additionally, the master CASP key was created through physicians' independent appraisal of each article followed by discussion of any disagreements. The use of a structured appraisal process with multiple types of responses allowed for a more objective, standardized evaluation of appraisal abilities in each study arm as well as an improved understanding of which tasks AI can be most useful for. To address the inherent bias in the subjective nature of this study, the medical student appraisers and otolaryngology evaluators independently performed their roles before all data were combined for analysis.

We decided to include RCTs from a single, high-impact journal to eliminate the variability that would result from a random array of study types and quality. However, this decreased the generalizability of our results and highlighted a need for future research evaluating various study types. Thorough critical appraisal of literature is a labor-intensive process. Therefore, we determined it was of greater importance to have multiple articles with few appraisers than multiple appraisers reviewing 1 or 2 articles each. With the current study design, we better addressed our research goal of how often appraisals with varying AI involvement agree in their responses.

We also recognize that medical students vary in their critical appraisal abilities. Second-year medical students were specifically chosen as they are required to have learned some degree of critical appraisal as a part of their required medical curriculum, providing a relatively consistent baseline level of training while allowing assessment of AI-assisted appraisal early in medical education [3]. However, the study included only 2 second-year medical student appraisers. Although these students independently completed all appraisals before analysis, their critical appraisal abilities and experience with AI tools may not be representative of second-year medical students more broadly. Consequently, the findings should be interpreted as exploratory and may not be generalizable to other populations of learners or educational settings. Future research would benefit from additional studies with larger and more diverse cohorts of appraisers and evaluators and a larger number of articles. The primary inferential analyses treated the 12 articles as independent repeated units, resulting in limited statistical power. No a priori power calculation

was performed, as the number of eligible RCTs identified during the prespecified study period determined the available sample. A sensitivity analysis based on 12 repeated units, 3 appraisal conditions, α=.05, and 80% power indicated that the study was approximately powered to detect an omnibus effect of Cohen f=0.68 (partial $\eta^2 \approx 0.32$), representing a very large effect. Therefore, smaller differences among appraisal conditions may have gone undetected, and nonsignificant findings should not be interpreted as evidence of equivalence.

Although a 1-month washout period was implemented to minimize recall bias, we cannot exclude the possibility that participants retained memorable details from individual studies. Such residual recall may have influenced subsequent appraisals, despite the crossover design and randomization of article order in this study. We selected the washout interval based on published evidence regarding retention periods; however, complete elimination of memory effects cannot be assured.

In this study, only RCTs were evaluated. RCTs are typically reported using standardized frameworks, such as the CONSORT (Consolidated Standards of Reporting Trials) guidelines, resulting in relatively consistent organization and reporting of study methods and outcomes. This standardized structure may facilitate information extraction and interpretation by LLMs when compared with less structured study designs. Consequently, the performance of GPT-4o observed in this study may not be generalizable to observational studies, diagnostic accuracy studies, qualitative research, systematic reviews, or other forms of medical literature. We recommend that future studies evaluate AI-assisted critical appraisal across a broader range of study designs.

An additional limitation of this study is the implementation of a new, updated ChatGPT model since this research was conducted. This is an expected limitation of any research using online AI software, as such tools are rapidly advancing with frequent updates. In turn, the advancement of AI outpaces our more traditional methods of research, abstract submission, manuscript writing, and journal article review. We still find this research important and relevant as it compares critical appraisals with varying levels of human authorship and generative AI. The authoring team believes that the quick progression of AI technology must not be a reason to avoid researching it and provides an even greater reason for thorough evaluation of its many roles in medicine.

Conclusions

GPT-4o demonstrated no statistically significant differences from medical student appraisals in binary correctness, completeness, or overall quality. Although an overall difference was observed in free-text accuracy, no individual pairwise comparison remained statistically significant following adjustment for multiple comparisons. These findings suggest that generative AI may serve as an adjunct to structured critical appraisal in medical education but should not replace independent critical reasoning or human oversight at this time. Larger studies involving diverse participants and study designs are needed to further evaluate the educational utility and limitations of generative AI in literature appraisal.

Generative AI Use

The authors attest that there was no use of generative AI technology in the generation of text, figures, or other informational content of this manuscript.

Funding

The authors declared no financial support was received for this work

Data Availability

The datasets generated and analyzed during this study, including deidentified appraisal responses, scoring rubrics, and study materials, are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: SO, JL, NJ
Formal analysis: SO
Investigation: SO, NDM, RB, FB, NJ
Methodology: SO, NDM, RB, FB, NJ
Writing—original draft: SO, NDM, RB, FB
Writing—review and editing: SO, NDM, RB, FB, JL, NJ

Conflicts of Interest

None declared.

Multimedia Appendix 1

Medical student and physician poststudy questionnaires. The following questionnaire was administered electronically using Google Forms after participating medical students completed all critical appraisal exercises.
[DOCX File (Microsoft Word File), 17 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Physician evaluator poststudy questionnaire. The following questionnaire was administered electronically using Google Forms after completion of all appraisal evaluations.

[DOCX File (Microsoft Word File), 17 KB-Multimedia Appendix 2]

Checklist 1

CASP checklist.
[DOCX File (Microsoft Word File), 101 KB-Checklist 1]

References

1. Tod D, Booth A, Smith B. Critical appraisal. Int Rev Sport Exerc Psychol. 2022;15(1):52-72. [doi: 10.1080/1750984X.2021.1952471]
2. Patel JJ, Hill A, Lee ZY, Heyland DK, Stoppe C. Critical appraisal of a systematic review: a concise review. Crit Care Med. Sep 1, 2022;50(9):1371-1379. [doi: 10.1097/CCM.0000000000005602] [Medline: 35853198]
3. Functions and structure of a medical school: standards for accreditation of medical education programs leading to the M.D. degree. Liaison Committee on Medical Education; 2023. URL: https://education.med.wustl.edu/app/uploads/2023/07/2024-25-Functions-and-Structure-of-a-Medical-School.pdf [Accessed 2026-11-05]
4. Taylor R, Reeves B, Ewings P, Binns S, Keast J, Mears R. A systematic review of the effectiveness of critical appraisal skills training for clinicians. Med Educ. Feb 2000;34(2):120-125. [doi: 10.1046/j.1365-2923.2000.00574.x] [Medline: 10652064]
5. Coomarasamy A, Taylor R, Khan KS. A systematic review of postgraduate teaching in evidence-based medicine and critical appraisal. Med Teach. Jan 2003;25(1):77-81. [doi: 10.1080/0142159021000061468] [Medline: 14741863]
6. Macrae HM, Regehr G, McKenzie M, et al. Teaching practicing surgeons critical appraisal skills with an internet-based journal club: a randomized, controlled trial. Surgery. Sep 2004;136(3):641-646. [doi: 10.1016/j.surg.2004.02.003] [Medline: 15349113]
7. Alkhaaldi SM, Kassab CH, Dimassi Z, et al. Medical student experiences and perceptions of ChatGPT and artificial intelligence: cross-sectional study. JMIR Med Educ. Dec 22, 2023;9:e51302. [doi: 10.2196/51302] [Medline: 38133911]
8. Zhang JS, Yoon C, Williams DK, Pinkas A. Exploring the usage of ChatGPT among medical students in the United States. J Med Educ Curric Dev. 2024;11:23821205241264695. [doi: 10.1177/23821205241264695] [Medline: 39092290]
9. Ganjavi C, Eppler M, O'Brien D, et al. ChatGPT and large language models (LLMs) awareness and use. A prospective cross-sectional survey of U.S. medical students. PLOS Digit Health. Sep 2024;3(9):e0000596. [doi: 10.1371/journal.pdig.0000596] [Medline: 39236008]
10. Hosseini M, Horbach SP. Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. Res Integr Peer Rev. May 18, 2023;8(1):4. [doi: 10.1186/s41073-023-00133-5] [Medline: 37198671]
11. Gilat R, Cole BJ. How will artificial intelligence affect scientific writing, reviewing and editing? The future is here …. Arthroscopy. May 2023;39(5):1119-1120. [doi: 10.1016/j.arthro.2023.01.014] [Medline: 36774329]
12. The Laryngoscope. Wiley Online Library. 2006. URL: https://onlinelibrary.wiley.com/journal/15314995 [Accessed 2025-10-08]
13. Chatterji A, Cunningham T, Deming DJ, et al. How people use ChatGPT. National Bureau of Economic Research. 2025. URL: https://ideas.repec.org/p/nbr/nberwo/34255.html [Accessed 2026-11-05]
14. CASP checklist. Critical Appraisal Skills Programme. URL: https://casp-uk.net/casp-tools-checklists/ [Accessed 2025-11-14]
15. Deenadayalan Y, Grimmer-Somers K, Prior M, Kumar S. How to run an effective journal club: a systematic review. J Eval Clin Pract. Oct 2008;14(5):898-911. [doi: 10.1111/j.1365-2753.2008.01050.x] [Medline: 19018924]
16. Carley SD, Mackway-Jones K, Jones A, et al. Moving towards evidence based emergency medicine: use of a structured critical appraisal journal club. J Accid Emerg Med. Jul 1998;15(4):220-222. [doi: 10.1136/emj.15.4.220] [Medline: 9681303]
17. Crowe M, Sheppard L. A review of critical appraisal tools show they lack rigor: alternative tool structure is proposed. J Clin Epidemiol. Jan 2011;64(1):79-89. [doi: 10.1016/j.jclinepi.2010.02.008] [Medline: 21130354]
18. Katrak P, Bialocerkowski AE, Massy-Westropp N, Kumar S, Grimmer KA. A systematic review of the content of critical appraisal tools. BMC Med Res Methodol. Sep 16, 2004;4:22. [doi: 10.1186/1471-2288-4-22] [Medline: 15369598]
19. Yu H, Liu Z, Wu J. Forgetting in order to remember better. arXiv. Preprint posted online on Dec 12, 2018. [doi: 10.48550/arXiv.1812.05668]
20. Vlach HA, Ankowski AA, Sandhofer CM. At the same time or apart in time? The role of presentation timing and retrieval dynamics in generalization. J Exp Psychol Learn Mem Cogn. Jan 2012;38(1):246-254. [doi: 10.1037/a0025260] [Medline: 21895392]

21. Ferreira JÂ, Valmorbida EL, Sato BG, Fuentes BP, Botti R. Forgetting curve models: a systematic review aimed at consolidating the main models and outlining possibilities for future research in production. Expert Syst. Feb 2024;41(2):e13405. [doi: 10.1111/exsy.13405]
22. Zhu K. Ranked: most popular AI tools by monthly site visits. Visual Capitalist. 2025. URL: https://www.visualcapitalist.com/ranked-most-popular-ai-tools-by-monthly-site-visits/ [Accessed 2026-11-05]
23. Ranade N, Saravia M, Johri A. Using rhetorical strategies to design prompts: a human-in-the-loop approach to make AI useful. AI Soc. 2024;40(2):711-732. [doi: 10.1007/s00146-024-01905-3]
24. Heston TF, Khun C. Prompt engineering in medical education. Int Med Educ. 2023;2(3):198-205. [doi: 10.3390/ime2030019]
25. Microsoft Excel. Microsoft Corporation. URL: https://www.microsoft.com/en-in/microsoft-365/excel [Accessed 2026-11-05]
26. Bourmand R, Olsson SE, Baloch FZ, Sonti A, Fijany A. Analyzing the accuracy of large language models in United States Medical Licensing Exam social-science preparation question banks. Cureus J Comput Sci. 2025;2. [doi: 10.7759/s44389-025-09114-9]
27. Guerra GA, Hofmann H, Sobhani S, et al. GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. World Neurosurg. Nov 2023;179:e160-e165. [doi: 10.1016/j.wneu.2023.08.042] [Medline: 37597659]
28. Chen TC, Multala E, Kearns P, et al. Assessment of ChatGPT's performance on neurology written board examination questions. BMJ Neurol Open. 2023;5(2):e000530. [doi: 10.1136/bmjno-2023-000530] [Medline: 37936648]
29. Knoedler L, Knoedler S, Hoch CC, et al. In-depth analysis of ChatGPT's performance based on specific signaling words and phrases in the question stem of 2377 USMLE step 1 style questions. Sci Rep. Jun 12, 2024;14(1):13553. [doi: 10.1038/s41598-024-63997-7] [Medline: 38866891]
30. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. Feb 2023;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
31. Camp NT, Bengtson JA, Sandstrom JC. The citation catastrophe: propagation of AI-generated counterfeit citations in scholarship. J Acad Librariansh. 2025;51(4):103065. [doi: 10.1016/j.acalib.2025.103065]
32. Kim SG. Using ChatGPT for language editing in scientific articles. Maxillofac Plast Reconstr Surg. Mar 8, 2023;45(1):13. [doi: 10.1186/s40902-023-00381-x] [Medline: 36882591]
33. Bisdas S, Topriceanu CC, Zakrzewska Z, et al. Artificial intelligence in medicine: a multinational multi-center survey on the medical and dental students' perception. Front Public Health. 2021;9:795284. [doi: 10.3389/fpubh.2021.795284] [Medline: 35004598]
34. Sami A, Tanveer F, Sajwani K, et al. Medical students' attitudes toward AI in education: perception, effectiveness, and its credibility. BMC Med Educ. Jan 17, 2025;25(1):82. [doi: 10.1186/s12909-025-06704-y] [Medline: 39833834]
35. Fine KS, Zona EE, O'Shea AW, et al. Evaluating the detection accuracy of AI-generated content in plastic surgery: a comparative study of medical professionals and AI tools. Plast Reconstr Surg. Jan 1, 2026;157(1):129e-136e. [doi: 10.1097/PRS.0000000000012242] [Medline: 40492637]
36. Bellini V, Semeraro F, Montomoli J, Cascella M, Bignami E. Between human and AI: assessing the reliability of AI text detection tools. Curr Med Res Opin. Mar 2024;40(3):353-358. [doi: 10.1080/03007995.2024.2310086] [Medline: 38265047]
37. Zhao Y, Borelli A, Martinez F, Xue H, Weiss GM. Admissions in the age of AI: detecting AI-generated application materials in higher education. Sci Rep. 2024;14(1):26411. [doi: 10.1038/s41598-024-77847-z] [Medline: 39488584]
38. Wiley releases AI guidelines for authors. Wiley. 2025. URL: https://newsroom.wiley.com/press-releases/press-release-details/2025/Wiley-Releases-AI-Guidelines-for-Authors/default.aspx [Accessed 2026-10-06]
39. The use of generative AI and AI-assisted technologies in the writing process. Elsevier. URL: https://www.elsevier.com/about/policies-and-standards/the-use-of-generative-ai-and-ai-assisted-technologies-in-writing-for-elsevier [Accessed 2025-10-06]
40. Piccirillo JF. *JAMA Otolaryngology–Head & Neck Surgery* —the year in review, 2024. JAMA Otolaryngol Head Neck Surg. May 2025;151(5):429-431. [doi: 10.1001/jamaoto.2025.0365]
41. Siddique MK, Rahim SA, Rahman MM, Lasker SP, Rahim M. Artificial intelligence in journal editing: expanding horizons for scholarly publications - a mixed method study. Health Sci Q. 2025;5(2):109-119. [doi: 10.26900/hsq.2674]
42. Fiorillo L, Mehta V. Accelerating editorial processes in scientific journals: leveraging AI for rapid manuscript review. Oral Oncol Rep. Jun 2024;10:100511. [doi: 10.1016/j.oor.2024.100511]
43. Leung TI, de Azevedo Cardoso T, Mavragani A, Eysenbach G. Best practices for using AI tools as an author, peer reviewer, or editor. J Med Internet Res. Aug 31, 2023;25:e51584. [doi: 10.2196/51584] [Medline: 37651164]

Abbreviations

CASP: Critical Appraisal Skills Programme
CONSORT: Consolidated Standards of Reporting Trials
IRB: institutional review board
LCME: Liaison Committee on Medical Education
LLM: large language model
RCT: randomized controlled trial

Edited by Luke MacNeill; peer-reviewed by Sven Benson; submitted 21.Apr.2026; initial comments to author 24.Jul.2026; final revised version received 18.Oct.2026; accepted 21.Oct.2026; published 20.Nov.2026

Please cite as:
Olsson S, Del Mundo N, Bourmand R, Baloch F, Liang J, Jiang N
Large Language Models for Structured Critical Appraisal of Randomized Controlled Trials in Otolaryngology: Randomized Crossover Study Comparing GPT-4o and Medical Students
JMIR Form Res 2026;10:e99015
URL: https://formative.jmir.org/2026/1/e99015
doi: 10.2196/99015

© Sofia Olsson, Nicholas Del Mundo, Raika Bourmand, Fiza Baloch, Jonathan Liang, Nancy Jiang. Originally published in JMIR Formative Research (https://formative.jmir.org), 20.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.