Quality, Accuracy, and Bias in ChatGPT-Based Summarization of Medical Abstracts

Joel Hake, MD
Miles Crowley, MD, MPH
Allison Coy, MD
Denton Shanks, DO, MPH
Aundria Eoff, MD
Kalee Kirmer-Voss, MD
Gurpreet Dhanda, MD
Daniel J. Parente, MD, PhD

Department of Family Medicine and Community Health, University of Kansas Medical Center, Kansas City, Kansas

MORE ONLINE
www.annfammed.org

VISUAL ABSTRACT

Conflicts of interest: authors report none.

CORRESPONDING AUTHOR

Daniel J. Parente
3901 Rainbow Blvd, MS 4010
Kansas City, KS 66160
dparente@kumc.edu

ABSTRACT

PURPOSE Worldwide clinical knowledge is expanding rapidly, but physicians have sparse time to review scientific literature. Large language models (eg, Chat Generative Pretrained Transformer [ChatGPT]), might help summarize and prioritize research articles to review. However, large language models sometimes "hallucinate" incorrect information.

METHODS We evaluated ChatGPT's ability to summarize 140 peer-reviewed abstracts from 14 journals. Physicians rated the quality, accuracy, and bias of the ChatGPT summaries. We also compared human ratings of relevance to various areas of medicine to ChatGPT relevance ratings.

RESULTS ChatGPT produced summaries that were 70% shorter (mean abstract length of 2,438 characters decreased to 739 characters). Summaries were nevertheless rated as high quality (median score 90, interquartile range [IQR] 87.0-92.5; scale 0-100), high accuracy (median 92.5, IQR 89.0-95.0), and low bias (median 0, IQR 0-7.5). Serious inaccuracies and hallucinations were uncommon. Classification of the relevance of entire journals to various fields of medicine closely mirrored physician classifications (nonlinear standard error of the regression [SER] 8.6 on a scale of 0-100). However, relevance classification for individual articles was much more modest (SER 22.3).

CONCLUSIONS Summaries generated by ChatGPT were 70% shorter than mean abstract length and were characterized by high quality, high accuracy, and low bias. Conversely, ChatGPT had modest ability to classify the relevance of articles to medical specialties. We suggest that ChatGPT can help family physicians accelerate review of the scientific literature and have developed software (pyJournalWatch) to support this application. Life-critical medical decisions should remain based on full, critical, and thoughtful evaluation of the full text of research articles in context with clinical guidelines.

Ann Fam Med 2024;22:113-120. https://doi.org/10.1370/afm.3075

INTRODUCTION

Nearly 1 million new journal articles were indexed by PubMed in 2020, and worldwide medical knowledge now doubles approximately every 73 days.[1] Meanwhile, care models emphasizing clinical productivity[2,3] leave clinicians with scant time to review the academic literature, even within their own specialty.

Recent developments in artificial intelligence (AI) and natural language processing might offer new tools to confront this problem. Large language models (LLMs) are neural network–based computer programs that use a detailed statistical understanding of written language to perform many tasks including text generation, summarization, software development, and prediction.[4-12] One LLM, Chat Generative Pretrained Transformer (ChatGPT) has recently garnered substantial attention in the popular press.[13-17] We wondered if LLMs could help physicians review the medical literature more systematically and efficiently.

Unfortunately, LLMs can also "hallucinate," producing text that, whereas often convincing and seemingly authoritative, is not fact based.[18-21] In addition, many concerns have been raised regarding the possibility of bias in AI models including LLMs. Bias in AI models can arise from both implicit and explicit biases present in their training data sets.[22,23] Additional biases might potentially arise during the fine-tuning process. Large language models can be fine-tuned via a reinforcement learning approach, which uses feedback from humans to improve model performance.[9] Such feedback might carry implicit and/or explicit biases of the humans providing feedback. Responsible use of LLMs at any stage of the clinical research process

therefore requires careful validation to ensure that specific uses are unlikely to exacerbate preexisting systemic inequalities in health care.

To perform tasks, LLMs are prompted with instructions and supporting information. We wondered if LLMs—when carefully instructed—could (1) help clinicians find articles relevant to their medical specialty and (2) produce reasonable summaries of the major findings without introducing inaccuracies as a result of hallucination. Specifically, we investigated whether ChatGPT-3.5 could produce (1) high quality, (2) accurate, and (3) bias-free summaries of medical abstracts, focusing on points that were most likely to be salient for practicing physicians. We also prompted ChatGPT to self-reflect on the quality, accuracy, and biasness of its own summaries and evaluated its performance in classifying articles' relevance to various medical specialties (eg, internal medicine, surgery, etc). Self-reflections have been used to improve the ability of LLMs to perform logical reasoning.[24] We compared these self-reflections and relevance classifications to annotations by human physicians.

METHODS

Article Selection

We analyzed 10 articles from each of 14 selected journals (Table 1). These journals were chosen to (1) include topics ranging across medicine, (2) include both structured and unstructured abstracts in the sample, and (3) span a large range of journal impact factors. We drew articles from research articles published in 2022 by simple random sampling of each journal. ChatGPT was trained on data assembled before 2022; therefore, we reasoned that these articles would not have been included in the training corpus (ie, not seen by ChatGPT previously). We included case series, observational studies, interventional studies, randomized controlled trials, systematic reviews, and meta-analyses. We excluded editorials, letters, perspectives, errata, nonsystematic reviews, and single-case reports.

ChatGPT Prompt and Data Extraction

We prompted ChatGPT with instructions (Supplemental Appendix 1) to summarize the abstract, self-reflect on the quality, accuracy, and degree of bias present in its summary, and classify the relevance of the abstract to 10 areas of medicine (cardiology, pulmonary medicine, family medicine, internal medicine, public health, primary care, neurology, psychiatry, obstetrics and gynecology, and general surgery). We instructed ChatGPT to adhere to a word limit of 125 words, but no efforts were otherwise undertaken to enforce this directive. Quality, accuracy, bias, and relevance were all evaluated on scales of 0-100. ChatGPT was given the full text of the abstract (as of February 2022 in PubMed) but was not provided with any other metadata (eg, journal, publication date) for the articles. We transcribed the ChatGPT-produced summary, along with the quality, accuracy, bias, and relevance scores, into Research Electronic Data Capture (REDCap; project-redcap.org) for data management.[25,26] We used the ChatGPT-3.5 (standard) model as of February 2022.

Physician Evaluation of ChatGPT Summaries

Seven physicians independently reviewed summaries. For each article, physician reviewers received the article's title, journal, PubMed ID, abstract, and the GPT-produced summary of the article via REDCap. The reviewers classified the quality, accuracy, and amount of bias present in the summaries on a 0-100 scale. Reviewers also evaluated the relevance of the articles to various areas of medicine on a 0-100 scale. To harmonize scores, reviewers used a common rubric to assign scores. For quality and accuracy, reviewers scored these on a 0-100 scale with anchors on the typical letter grade (A, B, C, D, or F) in the common American grading system, with corresponding ranges of 90-100, 80-89, 70-79, 60-69, and ≤59. For bias, reviewers used a common rubric ranging from "no bias" to "blatantly biased" (Supplemental Appendix 1). Reviewers also determined if the summary contained any evidence of bias on the basis of race, color, religion, sex, gender, sexual orientation, or national origin that was not present in the abstract. For relevance, reviewers were instructed to use a common rubric ranging from "clearly relevant" to "not relevant" (Supplemental Appendix 1). Each reviewer evaluated approximately 45 summaries. Reviews were randomly distributed to reviewers, and every summary was reviewed by 2 reviewers. The first 5 completed reviews for each reviewer were considered part of the reviewer burn-in phase and discarded. The review team comprised individuals of varied sexes, genders, races, religions, and national origins. The senior author (D.P.) acted as referee for the other reviewers and therefore did not review abstracts and summaries. Scores for quality, accuracy, bias, and relevance were averaged across all reviewers to produce final quality, accuracy, bias, and relevance scores for each summary. Reviewers also annotated both minor and serious factual inaccuracies. Serious factual inaccuracies were those that would change a major interpretation of an article. When substantial factual inaccuracies or biases were noted, reviewers supplied a text description of them.

Statistical and Qualitative Analyses

We performed statistical analyses using R version 4.2.2 (The R Foundation; r-project.org). To evaluate quality, accuracy, and bias, we calculated descriptive statistics (1) for the overall sample and (2) stratified by journal. We qualitatively compared the quality, accuracy, and bias score distributions for (1) ChatGPT, (2) individual human reviewers, and (3) all human reviewers in aggregate using violin plots and scatterplots.

For scores assigned by ChatGPT and human reviewers on how related each article was to various medical specialties, we conducted analyses at 2 levels: journal level and article level. First, we analyzed agreement at the journal level between

Table 1. Attributes of Journals Selected for Analysis

Journal	Field	Impact Factor	Abstract Type
American Journal of Epidemiology (*Am J Epidemiol*)	Public health	5.0	Unstructured
American Journal of Obstetrics and Gynecology (*Am J Obstet Gynecol*)	Obstetrics and gynecology	9.8	Structured
Annals of Family Medicine (*Ann Fam Med*)	General medicine	4.4	Structured
Annals of Internal Medicine (*Ann Intern Med*)	General medicine	39.2	Structured
Chest	Pulmonary medicine	9.6	Structured
JAMA Surgery (*JAMA Surg*)	Surgery	16.9	Structured
Journal of the American College of Cardiology (*J Am Coll Cardiol*)	Cardiovascular medicine	24.0	Structured
Journal of the American Medical Association (*JAMA*)	General medicine	120.7	Structured
Journal of General Internal Medicine (*J Gen Intern Med*)	General medicine	5.7	Structured
Nature Medicine (*Nat Med*)	General medicine	82.9	Unstructured
The Lancet Neurology (*Lancet Neurol*)	Neurology	48.0	Structured
The Lancet Psychiatry (*Lancet Psychiatry*)	Psychiatry	64.3	Structured
The Lancet Respiratory Medicine (*Lancet Respir Med*)	Pulmonary medicine	76.2	Structured
The New England Journal of Medicine (*N Engl J Med*)	General medicine	158.5	Structured

the ChatGPT-assigned relevance scores with (1) a priori expectations, and (2) human scores. For this journal-level analysis, we averaged across all articles for a given journal. For example, the "relevance to public health" score for *Annals of Family Medicine* is the average of all "relevance to public health" scores for all *Annals of Family Medicine* articles included. We expected a monotonic—although not necessarily linear—relation between the ChatGPT and human relevance scores. We could have asked ChatGPT and humans to make a dichotomous relevant/not relevant determination for each article. Instead, we collected more granular data on a scale of 0 (not relevant) to 100 (very relevant). In analogy to logistic regression analysis of categorical variables for classification, we modeled the nonlinear relation between ChatGPT (x) and human (y) relevance scores at the journal level using a 4-variable logistic function: $y \sim C + L/[1+e^{-k(x-x_0)}]$, where L models the difference in maximal and minimal human scores, x_0 models the difference in leniency between ChatGPT and humans, C is related to the difference in mean human score and mean ChatGPT score, and k describes the linear slope of the relation between human and ChatGPT scores near the midpoint of the fit. Nonlinear fits were computed using the nls function in the base R Stats package.

In addition, we defined the relevance profile for each journal as the vector of relevance scores for each specialty assigned to that journal. By this method, each journal received a ChatGPT-estimated relevance profile and a human-estimated relevance profile. The Euclidean distance between the relevance profile of 2 journals estimates their content similarity. We hierarchically clustered the journals via an agglomerative approach using both the ChatGPT and human relevance profiles and qualitatively compared the clustering dendrograms implied by ChatGPT- and human-assigned relevance scores.

At the article level, we evaluated the relation between ChatGPT relevance scores and human relevance scores (1) across all specialties in aggregate and (2) stratified by specialty. For the aggregate analysis, we again performed nonlinear regression with the 4-variable logistic function. A sensitivity analysis using a nonlinear mixed model, including reviewer identities as a random effect, was explored but did not substantially improve the quality of the nonlinear fit. For the analyses stratified by specialty, we used linear regression and calculated the coefficient of determination (R^2) for ChatGPT-predicted vs human-assigned scores within each specialty.

We used the Kruskal-Wallis rank-sum test to evaluate for differences in quality, accuracy, and bias scores stratified by (1) journal of origin and (2) structured vs unstructured abstracts.

Human Subjects Protection

This project was determined to be Not Human Subjects Research by the University of Kansas Medical Center Institutional Review Board.

RESULTS

Characteristics of ChatGPT Summaries of Medical Abstracts

We used ChatGPT to summarize 140 abstracts across 14 journals (Table 1). Most abstracts (n = 120) used a structured format. Abstracts included a mean of 2,438 characters. The ChatGPT summaries decreased this length by 70% to a mean of 739 characters. An example summary produced by ChatGPT is shown in Supplemental Appendix 2.

Summaries were scored by physician reviewers as high quality (median score 90.0, interquartile range [IQR] 87.0-92.5), high accuracy (median 92.5, IQR 89.0-95.0), and low bias (median 0, IQR 0-7.5) (Table 2). ChatGPT's self-reflections also rated the summaries as high quality, high accuracy, and low bias, concordant with the judgements of human

Table 2. Median Quality, Accuracy, and Bias Scores Assigned by Humans and ChatGPT to Articles Overall and Stratified by Journal

		Human Adjudicated, Median (IQR)			GPT Predicted, Median (IQR)		
	No.	Quality	Accuracy	Bias	Quality	Accuracy	Bias
Overall	140	90.0 (87.0-92.5)	92.5 (89.0-95.0)	0 (0-7.5)	90.0 (85.0-90.0)	90.0 (90.0-95.0)	0 (0)
Journal							
Ann Intern Med	10	90.0 (89.6-95.0)	93.75 (85.6-95.0)	0 (0)	90.0 (90.0-90.0)	95.0 (95.0-95.0)	0 (0)
Ann Fam Med	10	88.25 (86.2-90.0)	90.75 (87.5-93.3)	1.25 (0-10.6)	90.0 (90.0-93.8)	95.0 (91.25-95.0)	0 (0)
Chest	10	93.75 (88.8-95.0)	95.0 (90.6-95.0)	0 (0-3.8)	90.0 (81.25-90.0)	90.0 (90.0-93.8)	0 (0)
J Am Coll Cardiol	10	90.0 (88.0-92.5)	92.5 (89.5-95.0)	0 (0-7.5)	85.0 (85.0-90.0)	90.0 (90.0-90.0)	0 (0)
JAMA Surg	10	88.5 (85.5-90.0)	89.75 (87.5-90.0)	2.5 (0-7.5)	87.5 (80.0-90.0)	90.0 (90.0-90.0)	0 (0)
JAMA	10	90.0 (83.6-91.5)	91.75 (89.6-94.8)	0 (0)	90.0 (90.0-90.0)	90.0 (90.0-90.0)	0 (0)
J Gen Intern Med	10	95.0 (91.25-96.88)	95.0 (93.1-97.5)	0 (0-5.6)	90.0 (90.0-90.0)	90.0 (90.0-95.0)	0 (0)
Lancet Neurol	10	90.0 (84.6-94.4)	90.5 (89.3-94.4)	0 (0-3.8)	90.0 (85.0-90.0)	90.0 (90.0-90.0)	0 (0)
Lancet Psychiatry	10	87.5 (85.0-91.9)	90.0 (86.6-94.4)	0 (0-10.5)	90.0 (86.25-90.0)	95.0 (90.0-95.0)	0 (0)
Lancet Respir Med	10	87.5 (86.8-87.5)	89.75 (85.3-91.9)	0 (0-21.4)	90.0 (85.0-90.0)	90.0 (90.0-90.0)	0 (0)
Am J Obstet Gynecol	10	89.5 (85.1-90.0)	92.5 (88.9-94.3)	5.0 (0-10.5)	90.0 (90.0-90.0)	90.0 (90.0-90.0)	0 (0)
N Eng J Med	10	91.25 (87.1-92.5)	91.25 (89.2-92.5)	0 (0)	90.0 (90.0-93.8)	95.0 (90.0-95.0)	0 (0)
Nat Med	10	89.25 (87.5-92.9)	94.25 (91.4-96.4)	0 (0-1.9)	90.0 (90.0-90.0)	90.0 (90.0-93.8)	0 (0)
Am J Epidemiol	10	91.25 (88.1-96.2)	91.0 (90.0-96.9)	2.5 (0-9.4)	87.5 (80.0-93.8)	90.0 (90.0-93.8)	0 (0)

Am J Epidemiol = American Journal of Epidemiology; Am J Obstet Gynecol = American Journal of Obstetrics and Gynecology; Ann Fam Med = Annals of Family Medicine; Ann Intern Med = Annals of Internal Medicine; ChatGPT = Chat Generative Pretrained Transformer; GPT = Generative Pretrained Transformer; IQR = interquartile range; *J Am Coll Cardiol = Journal of the American College of Cardiology; JAMA = Journal of the American Medical Association; JAMA Surg = JAMA Surgery; J Gen Intern Med = Journal of General Internal Medicine; Lancet Neurol = The Lancet Neurology; Lancet Psychiatry = The Lancet Psychiatry; Lancet Respir Med = The Lancet Respiratory Medicine; N Engl J Med = The New England Journal of Medicine; Nat Med = Nature Medicine.*

evaluators. We found no difference in human-assigned scores when stratifying by (1) journal of origin or (2) structured vs unstructured abstracts.

Hallucinations and Inaccuracies

We qualitatively annotated instances of serious inaccuracies and hallucinations—defined as those that changed major interpretation of a study—in 4 of 140 summaries. One of these cases was due to omission; a significant risk factor (female gender) was found using a logistic regression model and was omitted from the summary whereas all other significant risk factors were reported. One of these cases was the result of apparent misunderstanding by ChatGPT of the semantic meaning of the abstract; a complicated study included 2 treatment arms with different primary outcomes, but the summary implied that the arms had the same primary outcomes. Two cases were due to hallucination; 1 summary stated that a randomized trial was double-blinded when the abstract clearly stated that it was open label, and 1 summary stated that results were consistent across subgroups, but only 1 of the many endpoints evaluated in that study was reported for the subgroup in the abstract. Minor inaccuracies were noted in 20 of 140 articles, related to the introduction of ambiguity in meaning (n = 2) or summarization away of details that would have provided additional content but not completely change the meaning (n = 18), for example, in cases in which an effect size was statistically significant but of questionable clinical significance.

ChatGPT Annotations of Journal Relevance Compared to Human Annotations

We included journals spanning many medical specialties. Our hypothesis was that ChatGPT would be able to classify articles drawn from a given journal as relevant to that journal's topical focus. For example, we would expect *The Lancet Neurology* to have high relevance to neurology and low relevance to obstetrics and gynecology. At the aggregate level, this hypothesis was borne out; the ChatGPT relevance profile of journals agreed with a priori expectations (Table 3). Likewise, there was a strong nonlinear association between physician and ChatGPT relevance scores at the journal level (standard error of the regression 8.6) (Figure 1). The standard error of the regression is related to the typical expected prediction error for the regression model. On our relevance scale, ranging from 0 to 100, a standard error of the regression of 8.6 can therefore be interpreted as an 8.6% expected error in the predicted human-assigned relevance given the ChatGPT-assigned relevance. Clustering analysis (Figure 2) revealed similarity of the general structure of the clustering dendrograms in both the ChatGPT and human dendrograms. General medicine and cardiovascular medicine journals cluster together, with specialist journals (psychiatry, neurology) branching out from the root of the tree. Likewise, 3 of the 4 strongest associations (*American Journal of Epidemiology* with *Annals of Internal Medicine, Journal of General Internal Medicine* with *Annals of Family Medicine, The Lancet Respiratory Medicine* with *Chest*) inferred from the human relevance profiles were also

Table 3. Journal Relevance Profiles; Topical Content of Each Journal as Annotated by ChatGPT

Journal	PrimCare	FM	IM	Pulm	Card	GenSurg	Neuro	Psych	Ob/Gyn	PubHealth
General medicine										
Ann Intern Med	76	76	74	46	60	49	47	50	59	77
Ann Fam Med	94	94	92	51	45	35	43	51	38	76
JAMA	66	66	63	40	52	42	60	32	36	64
J Gen Intern Med	85	86	85	38	49	40	41	55	47	74
N Engl J Med	73	73	84	40	29	25	20	17	13	67
Nat Med	68	68	79	48	57	32	38	36	24	66
Pulmonary medicine										
Chest	64	64	77	80	56	28	33	36	12	61
Lancet Respir Med	70	70	81	80	50	34	34	36	28	62
Cardiovascular medicine										
J Am Coll Cardiol	67	67	80	26	91	38	19	18	13	56
Surgery										
JAMA Surg	50	51	63	23	27	83	23	20	16	59
Neurology										
Lancet Neurol	54	54	66	25	27	35	87	37	6	49
Psychiatry										
Lancet Psychiatry	68	69	60	18	20	24	50	93	23	73
Ob/Gyn										
Am J Obstet Gynecol	66	66	65	29	41	45	39	38	86	70
Public health										
Am J Epidemiol	79	79	77	46	52	38	50	47	54	78

Am J Epidemiol = American Journal of Epidemiology; Am J Obstet Gynecol = American Journal of Obstetrics and Gynecology; Ann Fam Med = Annals of Family Medicine; Ann Intern Med = Annals of Internal Medicine; Card = cardiovascular medicine; ChatGPT = Chat Generative Pretrained Transformer; FM = family medicine; GenSurg = general surgery; IM = internal medicine; *J Am Coll Cardiol = Journal of the American College of Cardiology; JAMA = Journal of the American Medical Association; JAMA Surg = JAMA Surgery; J Gen Intern Med = Journal of General Internal Medicine; Lancet Neurol = The Lancet Neurology; Lancet Psychiatry = The Lancet Psychiatry; Lancet Respir Med = The Lancet Respiratory Medicine; N Engl J Med = The New England Journal of Medicine; Nat Med = Nature Medicine;* Neuro = neurology; Ob/Gyn = obstetrics and gynecology; PrimCare = primary care; Psych = psychiatry; PubHealth = public health; Pulm = pulmonary medicine.

Note: ChatGPT relevance profiles for each journal and specialty combination are calculated on a 0-100 scale and colored according to strength of relevance (strong = yellow; weak = blue).

present in the ChatGPT dendrogram (Figure 2; red-, yellow-, and blue-shaded areas). We thus conclude that ChatGPT can infer the relevance profile of journals based on analysis of the abstracts of its articles.

Ability of ChatGPT to Classify Relevance of Individual Articles to Various Disciplines of Medicine

We next evaluated whether ChatGPT could classify the relevance of individual articles. Within individual specialties, the relation between ChatGPT scores and human scores was much more modest at the article level than at the journal level (Supplemental Figure 1). Coefficients of determination (R^2) for linear regression ranged from 0.26 (general surgery) to 0.58 (obstetrics and gynecology). Likewise, global analysis of relevance scores assigned across all specialties revealed a clear but much weaker relation between human and ChatGPT-assigned relevance scores based on the standard error of the regression (Supplemental Figure 2). The standard error of the regression at the article level was 22.3, which is 2.5-fold greater than the journal-level analysis. We conclude that ChatGPT has only modest ability to classify the relevance of individual articles to specific domains of medicine.

Sensitivity and Quality Analyses

We visually inspected the distribution of scores for quality, accuracy, and bias produced by individual human reviewers, human reviewers in aggregate, and ChatGPT (Supplemental Figure 3). Score distributions were broadly similar, suggesting that harmonization instructions (Supplemental Appendix 1) given to reviewers were largely effective at standardizing scores of various human reviewers, without obvious variability due to individual reviewer leniency.

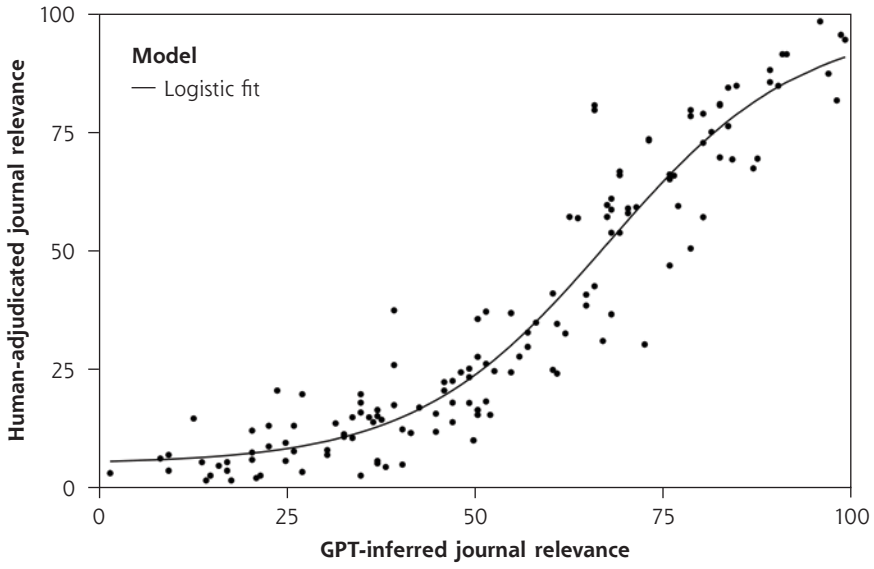
DISCUSSION

We evaluated whether the GPT-3.5 model, implemented as ChatGPT, could summarize medical research abstracts and determine the relevance of these articles to various medical specialties. Our analyses reveal that ChatGPT can produce high-quality, high-accuracy, and low-bias summaries of abstracts despite being given a word limit. We conclude that because ChatGPT summaries were 70% shorter than abstracts and usually of high quality, high accuracy, and low bias, they are likely to be useful as a screening tool to help busy clinicians and scientists more rapidly evaluate whether further review of an article is likely to be worthwhile. In Supplemental Appendix 3, we describe software—pyJournalWatch—that might enable this kind of application.[27-31] Life-critical medical decisions should for obvious reasons remain based on full, critical, and thoughtful evaluation of the full text of articles in context with available evidence from meta-analyses and professional guidelines. Our data also show that ChatGPT was much less able to classify the relevance of specific articles to various medical specialties. We had hoped to build a digital agent with the goal of consistently surveilling the medical literature, identifying relevant articles of interest to a given specialty, and forwarding them to a user. ChatGPT's inability to reliably classify the relevance of specific articles limits our ability to construct such an agent. We hope that in future iterations of LLMs, these tools will become more capable of relevance classification.

We are not aware of prior studies that have systematically evaluated GPT-3.5's ability to summarize medical abstracts with a focus on quality, accuracy, and bias. However, our data are concordant with prior evidence suggesting reasonable performance for summarization in other domains (eg, news).[32,33] Contrary to our expectations that hallucinations would limit the utility of ChatGPT for abstract summarization, this occurred in only 2 of 140 abstracts and was mainly limited to small (but important) methodologic or result details. Serious inaccuracies were likewise uncommon, occurring only in a further 2 of 140 articles. We conclude that ChatGPT summaries have rare but important inaccuracies that preclude them from being considered a definitive source of truth. Clinicians are strongly cautioned against relying solely on ChatGPT-based summaries to understand study methods and study results, especially in high-risk situations. Likewise, we noted at least 1 example in which the summary introduced bias by omitting gender as a significant risk factor in a logistic regression model, whereas all other significant risk factors were reported. In addition, concerns have been reasonably raised regarding biases inherent in LLMs.[22,23] Here, we investigated whether—if carefully prompted—ChatGPT could nevertheless be used to produce low-bias summaries despite this known theoretical limitation.

The present study has limitations. First, we considered only a limited number of journals, and all abstracts focused on clinical medicine. Summarization performance on biomedical research at earlier stages of translational research (eg, articles describing fundamental mechanisms of cellular biology or biochemistry) was not evaluated by our analysis. We also focused exclusively on primary research reports, systematic reviews, and meta-analyses. We did not evaluate the performance of ChatGPT on abstracts from many other article types that are important to the scientific process including nonsystematic reviews, perspectives, commentaries, and letters to the editor. Second, because most journals now use a structured abstract, we included a small number of unstructured abstracts in our data set. Although we found no difference in ChatGPT performance in summarizing structured vs unstructured abstracts, it could be that a sample including more unstructured abstracts might detect performance differences with smaller effect sizes. Third, although we included journals with a broad range of impact factors (4.4-158.5), our analyses focused mostly on high-impact journals or journals that are particularly well regarded in their own specialty. Abstracts written for high-quality journals might be easier (or harder) to summarize than articles published in lower-tier journals. Performance in lower-impact journals could be interrogated in future studies. Fourth, all articles were chosen using simple random sampling except that we belatedly

Figure 1. Agreement between human and GPT relevance scores at the journal level.

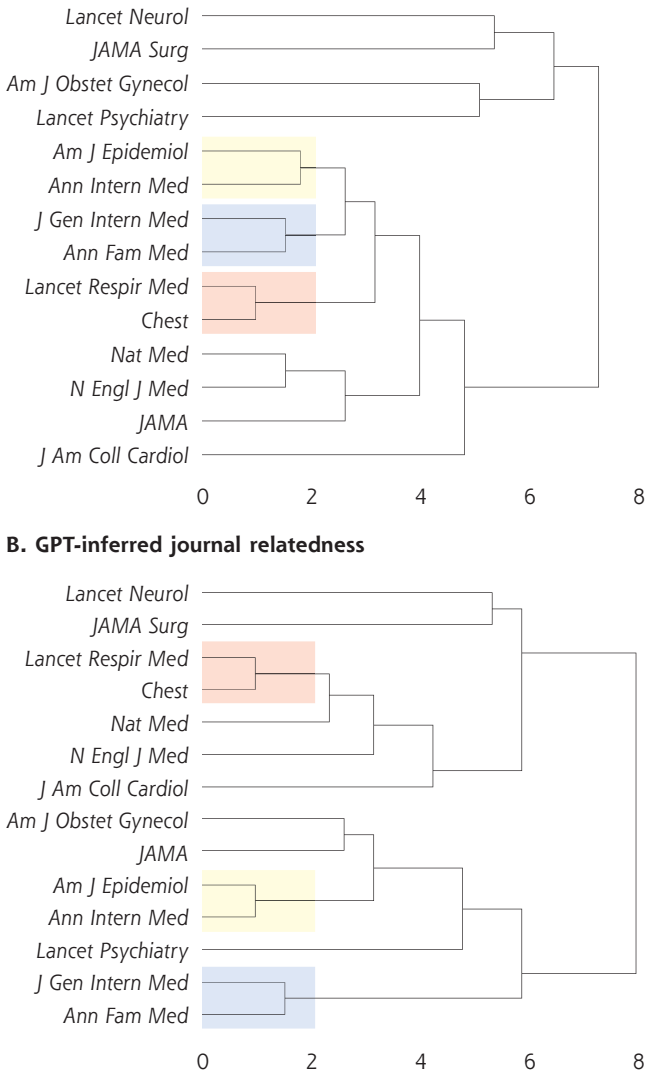


GPT = Generative Pretrained Transformer.

realized articles from *Nature Medicine* were sampled based on the order in which they were exported from PubMed rather than fully randomized. Sensitivity analyses excluding *Nature Medicine* did not change our conclusions; therefore, we kept articles from *Nature Medicine* in our analyses.

Large language models will continue to improve in quality. As we were finishing our analysis of ChatGPT based on the GPT-3.5 model, OpenAI began a limited beta release of the next generation in the GPT models, GPT-4. We suspect that as these models improve, summarization performance will be preserved and continue to improve. In addition, because the ChatGPT model was trained on pre-2022 data, it is possible that its slightly out-of-date medical knowledge decreased its ability to produce summaries or to self-assess the accuracy of its own summaries. In Supplemental Appendix 3, we report software that allows clinicians and scientists to immediately begin using GPT-3.5 and GPT-4 to systematically and rapidly review the clinical literature augmented by the advances in LLMs evaluated in this article. As LLMs evolve, future analyses should determine whether further iterations of the GPT language models have better performance in classifying the relevance of individual articles to various domains of medicine. In addition, in our analyses, we did not provide the LLMs with any article metadata such as the journal title or author list. Future analyses might investigate how performance varies when these metadata are provided. We encourage robust discussion within the family medicine research and clinical community on the responsible use of AI LLMs in family medicine research and primary care practice.

Figure 2. Clustering dendrograms for journal relatedness.



Am J Epidemiol = American Journal of Epidemiology; Am J Obstet Gynecol = American Journal of Obstetrics and Gynecology; Ann Fam Med = Annals of Family Medicine; Ann Intern Med = Annals of Internal Medicine; ChatGPT = Chat Generative Pretrained Transformer; GPT = Generative Pretrained Transformer; *J Am Coll Cardiol = Journal of the American College of Cardiology; JAMA = Journal of the American Medical Association; JAMA Surg = JAMA Surgery; J Gen Intern Med = Journal of General Internal Medicine; Lancet Neurol = The Lancet Neurology; Lancet Psychiatry = The Lancet Psychiatry; Lancet Respir Med = The Lancet Respiratory Medicine; N Engl J Med = The New England Journal of Medicine; Nat Med = Nature Medicine.*

Note: Human (panel A) and ChatGPT (panel B) dendrograms are shown. Strong relations between journals that are preserved in both dendrograms are highlighted in blue, yellow, and red.

Read or post commentaries in response to this article.

Key words: artificial intelligence; large language models; ChatGPT; primary care research; critical assessment of scientific literature; bias; text mining; text analysis

Submitted April 21, 2023; submitted, revised, October 13, 2023; accepted November 17, 2023.

Funding support: This work was not directly funded but used the REDCap data management platform at the University of Kansas Medical Center, which was supported by a Clinical and Translational Science Awards grant from the National Center for Advancing Translational Sciences (NCATS) awarded to the University of Kansas for Frontiers: University of Kansas Clinical and Translational Science Institute (#UL1TR002366). This work is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or NCATS. This agency had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Supplemental materials

References

1. Densen P. Challenges and opportunities facing medical education. *Trans Am Clin Climatol Assoc.* 2011;122:48-58.
2. Bradley EA, Winchester D, Alfonso CE, et al; American Heart Association Fellows in Training and Early Career Committee of the Council on Clinical Cardiology; Council on Cardiovascular Surgery and Anesthesia; Council on Lifelong Congenital Heart Disease and Heart Health in the Young; and Stroke Council. Physician wellness in academic cardiovascular medicine: a scientific statement from the American Heart Association. Circ*ulation.* 2022;146(16):e229-e241. 10.1161/cir.0000000000001093
3. Dillon EC, Tai-Seale M, Meehan A, et al. Frontline perspectives on physician burnout and strategies to improve well-being: interviews with physicians and health system leaders. *J Gen Intern Med.* 2020;35(1):261-267. 10.1007/s11606-019-05381-0
4. Amatriain X, Sankar A, Bing J, Bodigutla PK, Hazen TJ, Kazi M. Transformer models: an introduction and catalog. arXiv:2302.07730. Published Feb 12, 2023. Last updated May 25, 2023. Accessed Jan 5, 2024. https://arxiv.org/abs/2302.07730

5. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Presented at: 31st Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA. Accessed Jan 5, 2024. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
6. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Presented at: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020; December 6-12, 2020; Vancouver, Canada (virtual). Accessed Jan 5, 2024. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
7. Neelakantan A, Xu T, Puri R, et al. Text and code embeddings by contrastive pre-training. arXiv:2201.10005. Published Jan 24, 2022. Accessed Jan 5, 2024. https://arxiv.org/abs/2201.10005
8. Stiennon N, Ouyang L, Wu J, et al. Learning to summarize with human feedback. Presented at: 34th Conference on Neural Information Processing Systems. December 6-12, 2020; Vancouver, Canada (virtual). Accessed Jan 5, 2024. https://proceedings.neurips.cc/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf
9. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. arXiv:2203.02155. Published Mar 4, 2022. Accessed Jan 5, 2024. https://arxiv.org/abs/2203.02155
10. 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*. 2023;2(2):e0000198. 10.1371/journal.pdig.0000198
11. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. arXiv:2205.11916. Published May 24, 2022. Last updated Jan 29, 2023. Accessed Jan 5, 2024. https://arxiv.org/abs/2205.11916
12. Chen M, Tworek J, Jun H, et al. Evaluating large language models trained on code. arXiv:2107.03374. Presented Jul 7, 2021. Last updated Jul 14, 2021. Accessed Jan 5, 2024. https://arxiv.org/abs/2107.03374
13. O'Brien M. Microsoft invests billions in ChatGPT-maker OpenAI. *AP News*. Published Jan 23, 2023. Accessed Jan 5, 2024. https://apnews.com/article/technology-science-microsoft-corp-business-artificial-intelligence-03f157ddc482ef76f4999b929eaac7bf
14. Browne R. All you need to know about ChatGPT, the A.I. chatbot that's got the world talking and tech giants clashing. *CNBC*. Published Feb 28, 2023. Updated Apr 17, 2023. Accessed Jan 5, 2024. https://www.cnbc.com/2023/02/08/what-is-chatgpt-viral-ai-chatbot-at-heart-of-microsoft-google-fight.html
15. Agrawal A, Gans J, Goldfarb A. ChatGPT and how AI disrupts industries. *Harv Bus Rev*. Published Dec 12, 2022. Accessed Jan 5, 2024. https://hbr.org/2022/12/chatgpt-and-how-ai-disrupts-industries
16. Wunker S. Disruptive innovation and ChatGPT – three lessons from the smartphone's emergence. *Forbes*. Published Feb 16, 2023. Accessed Jan 5, 2024. https://www.forbes.com/sites/stephenwunker/2023/02/16/disruptive-innovation-and-chatgpt--three-lessons-from-the-smartphones-emergence/?sh=13c0ee1f61aa
17. Warzel C. Is this the week AI changed everything? *The Atlantic*. Published Feb 9, 2023. Accessed Jan 5, 2024. https://www.theatlantic.com/technology/archive/2023/02/google-bing-race-to-launch-ai-chatbot-powered-search-engines/673006/
18. Sun W, Shi Z, Gao S, Ren P, de Rijke M, Ren Z. Contrastive learning reduces hallucination in conversations. arXiv:2212.10400. Published Dec 20, 2022. Accessed Jan 5, 2024. https://arxiv.org/abs/2212.10400
19. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. *Crit Care*. 2023;27(1):120. 10.1186/s13054-023-04393-x.
20. Roller S, Dinan E, Goyal N, et al. Recipes for building an open-domain chatbot. arXiv:2004.13637. Published Apr 28, 2020. Last updated Apr 30, 2020. Accessed Jan 5, 2024. https://arxiv.org/abs/2004.13637
21. Shuster K, Poff S, Chen M, Kiela D, Weston J. Retrieval augmentation reduces hallucination in conversation. arXiv:2104.07567. Published Apr 15, 2021. Accessed Jan 5, 2024. https://arxiv.org/abs/2104.07567
22. Kirk H, Jun Y, Iqbal H, et al. Bias out-of-the-box: an empirical analysis of intersectional occupational biases in popular generative language models. arXiv:2102.04130. Published Feb 8, 2021. Last updated Oct 27, 2021. Accessed Jan 5, 2024. https://arxiv.org/abs/2102.04130
23. Nozza D, Bianchi F, Hovy D. Pipelines for social bias testing of large language models. Presented at: BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models. May 2022. Accessed Jan 5, 2024. https://aclanthology.org/2022.bigscience-1.6
24. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv:2201.11903. Published Jan 28, 2022. Last updated Jan 10, 2023. Accessed Jan 5, 2024. https://arxiv.org/abs/2201.11903
25. Harris PA, Taylor R, Minor BL, et al; REDCap Consortium. The REDCap consortium: building an international community of software platform partners. *J Biomed Inform*. 2019;95:103208. 10.1016/j.jbi.2019.103208
26. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377-381. 10.1016/j.jbi.2008.08.010
27. Sayers E. The E-utilities in-depth: parameters, syntax and more. In: Sayers E, ed. *Entrez Programming Utilities Help*. National Center for Biotechnology Information; 2009. Last updated Nov 30, 2022. Accessed Jan 5, 2024. https://www.ncbi.nlm.nih.gov/books/NBK25499
28. Open AI. API Reference - OpenAI API 2023. Accessed Apr 20, 2023. https://platform.openai.com/docs/api-reference
29. Wobben G. Releasing PyMed: the PubMed library for Python. *Medium*. Published Jul 2, 2018. Accessed Apr 19, 2023. https://medium.com/@gijswobben/releasing-pymed-7429826ed0a
30. Scao TL, Fan A, Akiki C, et al; BigScience Workshop. Bloom: a 176b-parameter open-access multilingual language model. arXiv:2211.05100. Published Nov 9, 2022. Last updated Jun 27, 2022. Accessed Jan 5, 2024. https://arxiv.org/abs/2211.05100
31. Thoppilan R, De Freitas D, Hall J, et al. LaMDA: language models for dialog applications. arXiv:2201.08239. Published Jan 20, 2022. Last updated Feb 10, 2022. Accessed Jan 5, 2024. https://arxiv.org/abs/2201.08239
32. Goyal T, Li JJ, Durrett G. News summarization and evaluation in the era of GPT-3. arXiv:2209.12356. Published Sep 26, 2022. Last updated May 23, 2022. Accessed Jan 5, 2024. https://arxiv.org/abs/2209.12356
33. Gupta A, Chugh D, Anjum A, Katarya R. Automated news summarization using transformers. arXiv:2108.01064. Published Apr 23, 2021. Accessed Jan 5, 2024. https://arxiv.org/abs/2108.01064