



AOA Critical Issues in Education

Can Artificial Intelligence Fool Residency Selection Committees? Analysis of Personal Statements by Real Applicants and Generative AI, a Randomized, Single-Blind Multicenter Study

Zachary C. Lum, DO, Lohitha Guntupalli, BS, Augustine M. Saiz, MD, Holly Leshikar, MD, Hai V. Le, MD, John P. Meehan, MD, and Eric G. Huish, DO

Investigation performed at the Nova Southeastern University, Davie, Florida

Introduction: The potential capabilities of generative artificial intelligence (AI) tools have been relatively unexplored, particularly in the realm of creating personalized statements for medical students applying to residencies. This study aimed to investigate the ability of generative AI, specifically ChatGPT and Google BARD, to generate personal statements and assess whether faculty on residency selection committees could (1) evaluate differences between real and AI statements and (2) determine differences based on 13 defined and specific metrics of a personal statement.

Methods: Fifteen real personal statements were used to generate 15 unique and distinct personal statements from ChatGPT and BARD each, resulting in a total of 45 statements. Statements were then randomized, blinded, and presented to a group of faculty reviewers on residency selection committees. Reviewers assessed the statements by 14 metrics including if the personal statement was AI-generated or real. Comparison of all metrics was performed.

continued

Each author certifies that he or she has no commercial associations (e.g., consultancies, stock ownership, equity interest, and patent/licensing arrangements) that might pose a conflict of interest in connection with the submitted article.

In addition, the faculty reviewers who participated in the evaluation of these statements were provided with clear guidelines, and their involvement was voluntary and without any conflict of interest. The process of randomization and blinding was employed to maintain objectivity and minimize bias in the evaluation. The faculty reviewers received each statement in a randomized order to reduce sequential bias from repeat order. The reviewers were blinded as they did not know the identity of each group of statements. Each reviewer was assigned a randomized numerical generator and instructed to review each statement in that specific order. All data reviewed by the faculty had no identifying information.

This study utilized generative artificial intelligence and deidentified writing samples, so institutional review board approval was not obtained.

Disclosure: The **Disclosure of Potential Conflicts of Interest** forms are provided with the online version of the article (<http://links.lww.com/JBJSOA/A682>).

Copyright © 2024 The Authors. Published by The Journal of Bone and Joint Surgery, Incorporated. All rights reserved. This is an open access article distributed under the terms of the [Creative Commons Attribution-Non Commercial-No Derivatives License 4.0](https://creativecommons.org/licenses/by-nc-nd/4.0/) (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

Results: Faculty correctly identified 88% (79/90) real statements, 90% (81/90) BARD, and 44% (40/90) ChatGPT statements. Accuracy of identifying real and BARD statements was 89%, but this dropped to 74% when including ChatGPT. In addition, the accuracy did not increase as faculty members reviewed more personal statements (area under the curve [AUC] 0.498, $p = 0.966$). BARD performed poorer than both real and ChatGPT across all metrics ($p < 0.001$). Comparing real with ChatGPT, there was no difference in most metrics, except for Personal Interests, Reasons for Choosing Residency, Career Goals, Compelling Nature and Originality, and all favoring the real personal statements ($p = 0.001$, $p = 0.002$, $p < 0.001$, $p < 0.001$, and $p < 0.001$, respectively).

Conclusion: Faculty members accurately identified real and BARD statements, but ChatGPT deceived them 56% of the time. Although AI can craft convincing statements that are sometimes indistinguishable from real ones, replicating the humanistic experience, personal nuances, and individualistic elements found in real personal statements is difficult. Residency selection committees might want to prioritize these particular metrics while assessing personal statements, given the growing capabilities of AI in this arena.

Clinical Relevance: Residency selection committees may want to prioritize certain metrics unique to the human element such as personal interests, reasons for choosing residency, career goals, compelling nature, and originality when evaluating personal statements.

Introduction

The rapid introduction of generative artificial intelligence (AI) tools such as ChatGPT (OpenAI) and BARD (Google) has resulted in new era of synthetic writing, including the potential for generating personal statements for medical students applying to residency programs. With the environment of candidate evaluation shifting dramatically in recent years, this study explores the ability of generative AI (GAI), specifically large language models (LLM), in the generation of personal statements, with a particular focus on their application within the field of orthopaedic surgery residency applications.

Traditionally, selection committees for medical and surgical residencies relied heavily on metrics such as United States Medical Licensing Examination (USMLE) scores, letters of recommendation, research, and rotation performance to evaluate applicants and ultimately offer them an interview to their program¹. However, the shift of USMLE step 1 to a pass/fail system has altered the applicant evaluation capability of program committees, which may result in further nonobjective measurements of an applicant's ability to succeed in training^{2,3}. In addition, the introduction of standardized letters of recommendation has further decreased the margin for evaluation; most students now score within a narrow band, leading to difficulties in differentiating among candidates^{4,5}. This has led selection committees to increasingly consider other aspects of the application, such as personal statements, which are expected to reflect the applicants' motivations and reasons for pursuing a specific medical or surgical specialty.

With the public introduction of GAI, it is essential for selection committee members to understand the capabilities of AI-generated personal statements and their limitations. The goal of this study was to evaluate the ability of orthopaedic surgery residency selection committee members to accurately identify personal statements as real or AI-generated, and to compare 14 distinct metrics within these statements to detect any differences between real and AI-generated statements. We also aim to discover any differentiating characteristics that could hint at the origins of the statements.

Methods

This research study involving the use of personal statements from fourth-year medical students and GAI chatbots was conducted with ethical considerations. All personal statements, whether real or generated, were anonymized, and the privacy of individuals was protected. Real statements were obtained from Electronic Residency Application Service applications after medical student consent, at minimum 2 years after their match cycle to ensure voluntary participation. No personally identifiable information was disclosed or utilized in this research. An Institutional Review Board (IRB) protocol was written, and IRB review was approved.

Fifteen real personal statements from fourth-year medical students, comprised of statements that had received at minimum 1 orthopaedic surgery residency interview invitation. These statements were used to train both ChatGPT and Google BARD by entering the statements directly into the chatbox with the phrase "Please use this statement to prepare to write a subsequent personal statement". Subsequently, the GAI chatbots were prompted to generate 15 unique and distinct personal statements each, resulting in a total of 45 statements. A prompt of "Utilizing the text and personal statements within this chatbox, write a different and unique personal statement applying to orthopaedic surgery residency" was given. These statements were then saved and labeled. No other criteria were used to generate the personal statements except for the prompt criteria. Although the models output is based on foundational modeling, its outputs will also rely on the personal statements in the chatbox.

All statements, real and generated, were then randomized and blinded by a nontesting researcher, and presented to a group of faculty reviewers who have served or are currently serving on an orthopaedic surgery residency selection committee. Reviewers were selected from multiple institutions to decrease the risk of single-center institutional bias. The faculty members assessed the statements in an assigned sequential randomized order using a set of 13 metrics ranked by the Likert scale, ranging from 1 worst to 5

TABLE I Residency Selection Committee Evaluation of Personal Statements, Whether Real or Artificial Intelligence-Generated*

	Correct Identification	Incorrect Identification
Real PS	79	11
ChatGPT	40	50
BARD	81	9

* χ^2 test was used to detect differences between these variables. Chi-squared individually was $p < 0.00001$. When combining ChatGPT and BARD together as AI, χ^2 test $p = 0.00028$. PS = personal statement.

best. These metrics included grammar, word usage, punctuation, sentence/paragraph structure, overall organization, originality, articulation, compelling nature, English proficiency, reasons for choosing orthopaedic surgery, personal interests, career goals, and relevance to the residency program (Supplemental A). Metrics were determined based on previous studies evaluating personal statements⁶.

Finally, faculty were asked to determine whether each personal statement was AI-generated or real, written by a medical student. A comparison of all metrics was conducted between the personal statements by BARD, by ChatGPT, and those written by medical students. Analysis was also performed to identify whether reviewers' ability to identify AI-generated statements would improve with the increase in the number of statements they reviewed.

Statistical Analysis

Contingency table χ^2 testing was used to determine differences between correctly and incorrectly determined real, ChatGPT, and BARD statements. Accuracy, sensitivity, and specificity were calculated based on contingency tables. A receiver operator characteristic curve was created displaying true positive rates (sensitivity) against false positive rates (1-specificity) when varying the number of personal statements reviewed. For the metric data, multifactor ordinal logistic regression was performed comparing reviewer scores for 13 metrics based on author (real applicant, ChatGPT, and BARD) while controlling

for multiple reviewers and statements as fixed effects in the model using SPSS Statistics, version 25 (IBM), to determine the odds ratio of a personal statement receiving a higher score for each domain based on the author.

Power Analysis

On a 5-point Likert scale, a single-point difference with a standard deviation of 1.0, power of 80% and alpha 0.05, and enrollment ratio of 1:1:1, the cohort size was determined to be 15 per group.

Results

Faculty correctly identified 88% (79/90) real statements, 90% (81/90) BARD, and 44% (40/90) ChatGPT statements (Table I). Accuracy of identifying real and BARD statements was 89%, but this dropped to 74% when including ChatGPT. Reviewers identified statements written by AI (ChatGPT or BARD) with 67% sensitivity and 88% specificity. In addition, the accuracy did not increase as faculty members reviewed more personal statements with the AUC of 0.498 ($p = 0.966$) indicating that the number of statements reviewed yielded no additional ability to discriminate between real authors and AI resulting in an equal likelihood of a correct identification at any point during the review.

Metrics from the 3 groups of personal statements are presented in Table II. Using multifactor ordinal logistic regression, BARD performed poorer than both real and ChatGPT across all 13 metrics including grammar, word usage, punctuation, sentence/paragraph structure, overall organization, originality, articulation, compelling nature, English proficiency, reason for choosing specialty, personal interests, career goals, and relevance to residency program. The odds ratio of a faculty reviewer assigning a BARD written personal statement at a higher score was very low at odds ratio (OR) range 0.01 to 0.18, $p < 0.001$ (Tables III and IV).

When comparing ChatGPT statements with real ones, there was no difference in several metrics, including word usage (OR 0.53, $p = 0.06$), overall organization (OR 0.73, $p = 0.30$), articulation (OR 0.64, $p = 0.15$), English proficiency (OR 0.62, $p = 0.20$), sentence/paragraph structure (OR 0.89, $p = 0.71$), and relevance to residency program (OR 0.62, $p = 0.13$). Although ChatGPT had higher odds of receiving a higher rating in grammar (OR 1.45, $p = 0.28$) and punctuation (OR 1.63,

TABLE II Raw Value Metrics From Real, BARD, and ChatGPT Personal Statements

Personal Statement	Grammar	Word Usage	Punctuation	Sentence/Paragraph Structure	Overall Organization	Originality	Articulation	Compelling Nature	English Proficiency	Reasons for Choosing Orthopaedic Surgery	Personal Interests	Career Goals	Relevance to the Residency Program
Real (avg)	3.80	3.90	3.86	3.84	3.77	3.74	3.68	3.57	4.15	3.71	3.37	3.16	3.13
Real (SD)	0.63	0.68	0.58	0.74	0.87	1.00	0.82	1.07	0.79	0.93	0.96	0.98	1.18
ChatGPT (avg)	3.88	3.73	3.95	3.87	3.70	3.12	3.51	3.01	4.12	3.33	3.04	2.78	2.95
ChatGPT (SD)	0.58	0.76	0.61	0.77	0.95	1.20	1.00	1.14	0.79	1.05	1.02	0.99	0.99
BARD (avg)	3.25	2.74	3.31	2.63	2.43	2.03	2.31	1.98	3.20	2.23	2.00	1.97	1.87
BARD (SD)	0.82	0.94	0.85	0.95	1.11	0.91	0.96	0.88	1.12	0.90	0.80	0.86	0.85

TABLE III BARD vs. Real Personal Statements *

Odds Ratio of Bard Generated Personal Statements Scoring Higher Than Those Written by Real Applicants					
Metric	OR	Lower 95% CI	Upper 95% CI	p value	
Grammar	0.174	0.085	0.354	<0.001	
Word usage	0.025	0.011	0.055	<0.001	
Punctuation	0.182	0.087	0.378	<0.001	
Sentence/paragraph structure	0.026	0.012	0.057	<0.001	
Overall organization	0.043	0.021	0.088	<0.001	
Originality	0.023	0.011	0.048	<0.001	
Articulation	0.035	0.017	0.073	<0.001	
Compelling nature	0.026	0.013	0.055	<0.001	
English proficiency	0.019	0.007	0.046	<0.001	
Reason for choosing specialty	0.023	0.011	0.049	<0.001	
Personal interests	0.016	0.007	0.036	<0.001	
Career goals	0.034	0.016	0.074	<0.001	
Relevance to residency program	0.051	0.025	0.105	<0.001	

*BARD performed worse than Real statements in all metrics. CI = confidence interval, and OR = odds ratio.

$p = 0.18$), something that would be expected for a language-based AI system, the differences were not significant ($p > 0.05$).

ChatGPT performed significantly worse than real statements for 5 metrics including personal interests (OR 0.33, 95% confidence interval [CI] 0.18-0.62, $p = 0.001$), reasons for choosing orthopaedic surgery (OR 0.37, 95% CI 0.20-0.68, $p = 0.002$), career goals (OR 0.30, 95% CI 0.16-0.56, $p < 0.001$), compelling nature (OR 0.25, 95% CI 0.13-0.47, $p < 0.001$), and

originality (OR 0.22, 95% CI 0.11-0.41, $p < 0.001$), favoring the real personal statements (Table V).

Discussion

Faculty members accurately identified real and BARD-generated personal statements, but ChatGPT deceived them 56% of the time. In addition, metrics between BARD, ChatGPT, and real personal statements often showed

TABLE IV BARD vs. ChatGPT Personal Statements *

Odds Ratio of BARD Generated Personal Statements Scoring Higher Than Those Written by ChatGPT					
Metric	OR	Lower 95% CI	Upper 95% CI	p value	
Grammar	0.090	0.041	0.200	<0.001	
Word usage	0.023	0.009	0.060	<0.001	
Punctuation	0.069	0.029	0.164	<0.001	
Sentence/paragraph structure	0.019	0.008	0.047	<0.001	
Overall organization	0.042	0.019	0.090	<0.001	
Originality	0.090	0.044	0.184	<0.001	
Articulation	0.047	0.022	0.102	<0.001	
Compelling nature	0.090	0.043	0.186	<0.001	
English proficiency	0.012	0.004	0.036	<0.001	
Reason for choosing specialty	0.053	0.025	0.113	<0.001	
Personal interests	0.042	0.018	0.098	<0.001	
Career goals	0.070	0.031	0.156	<0.001	
Relevance to residency program	0.055	0.025	0.119	<0.001	

*BARD performed worse than ChatGPT statements in all metrics. CI = confidence interval, and OR = odds ratio.

TABLE V ChatGPT vs. Real Personal Statements*

Odds Ratio of ChatGPT Generated Personal Statements Scoring Higher Than Those Written by Real Applicants				
Metric	OR	Lower 95% CI	Upper 95% CI	p value
Grammar	1.450	0.744	2.828	0.275
Word usage	0.526	0.272	1.014	0.055
Punctuation	1.629	0.798	3.326	0.180
Sentence/paragraph structure	0.891	0.479	1.655	0.714
Overall organization	0.725	0.397	1.324	0.295
Originality	0.215	0.114	0.406	<0.001
Articulation	0.636	0.342	1.181	0.152
Compelling nature	0.251	0.134	0.469	<0.001
English proficiency	0.617	0.294	1.295	0.201
Reason for choosing specialty	0.365	0.195	0.684	0.002
Personal interests	0.328	0.175	0.615	0.001
Career goals	0.296	0.157	0.556	<0.001
Relevance to residency program	0.615	0.329	1.152	0.129

*ChatGPT performed worse than Real statements in metrics such as personal interests, reason for choosing specialty, career goals, originality, and compelling nature. Interestingly, ChatGPT performed better in metrics such as grammar and punctuation, something that would be expected for a language-based AI system; however, these were not statistically significant. CI = confidence interval and OR = odds ratio.

significant variation. BARD performed worse in all metrics compared with both real and ChatGPT ($p < 0.001$) (Table II). This discrepancy between BARD and ChatGPT may be due to the specific training data from each AI. BARD originates from Language Models for Dialog Applications, which comprises pretraining data from Infiniset, a combination of public web data and documents that total 1.56 trillion words and 137 billion parameters and includes areas such as dialog data from public forums (50%), code sites (12.5%), C4 data (12.5%), Wikipedia (12.5%), and English and non-English web documents (12.5%)^{7,8}. ChatGPT pretraining data come from public text data that are undisclosed, but include books, web text, and public forums with 175 billion parameters and more than 400 billion tokens⁹. Parameters, unlike words, are configurable variables in machine learning models, akin to contextual clues surrounding words that help predict or generate the next word based on previous words. Tokens are sequence of character instances grouped together usually corresponding to a word or punctuation. Furthermore, the earlier release of ChatGPT (November 30, 2022) compared with BARD (March 21, 2023) may have conferred an advantage to ChatGPT. The additional time and public interaction may have honed its ability to mimic human experiences and emotional nuances more convincingly.

Postdata collection feedback from the faculty reviewers (who remained blinded to the results) suggested they felt AI-generated statements lacked personal touch and real-life examples, rendering them rather generic. Interestingly, this feedback seemed to align with their ability to correctly identify real and BARD statements. However, distinguishing statements created by ChatGPT proved more challenging for the residency

selection committee members. One plausible reason could be that ChatGPT's statements were interspersed with personal details and real-world examples, adding to their authenticity. This might be attributable to the difference in the timeline of their respective releases. While a 4-month gap may not seem significant, it is noteworthy that editorials commenting on ChatGPT's ability to convincingly generate research abstracts and personal statements were published during this period^{10,11}. In addition, OpenAI utilized human expert feedback mechanisms to train ChatGPT in appropriate responses in an attempt to lower hallucination rates and increase accuracy of information which may have propelled its LLM above BARD¹².

Despite the challenges in distinguishing ChatGPT and real statements, subtle differences emerged in areas like reasons for choosing the residency, personal interests, career goals, originality, and compelling nature. These aspects were more favorably portrayed in real statements, highlighting ChatGPT's limitations in replicating nuanced human experiences and individuality. This suggests that future evaluations of personal statements should focus on these uniquely human elements. Moreover, the real personal statements did not perform significantly worse in any metric compared with both ChatGPT and BARD, indicating no potential advantage over real personal statements by any of the faculty reviewer metrics. This underscores the critical importance of authentic, individualized elements in personal statements, reflecting the genuine motivations and compassion that drive applicants to pursue a career in medicine and surgery.

Finally, the lack of improvement in discerning AI-generated statements, even with increased exposure, suggests a steep learning curve, highlighting the AI's sophistication and our

difficulty in detecting subtleties. This could have important implications in the future, especially considering the increasing reliance on personal statements in the absence of a scored USMLE step 1 score and the homogenization of letters of recommendation²⁻⁵. Given AI's increasing ability to simulate other evaluation metrics, attributes, such as personal interests, career goals, originality, and compelling reasons for choosing the residency, might demand a higher emphasis during personal statement assessments.


The strengths of our study include that it was a randomized blinded multicenter study, utilizing faculty reviewers from varying healthcare system size and geographical areas. We attempted to minimize bias by blinding all reviewers to the origin of the letters. We also randomized each reviewer to a unique random number to prevent training bias from repeated exposures. In addition, we performed AUC analysis to assess accuracy in relation to the number of statements, which did not show significant improvement in identification with increasing statement count.

This study had several limitations. Its primary aim was to investigate whether residency selection committee members could identify authentic personal statements vs. GAI statements. A more real-world scenario would involve a combination of GAI-generated content and genuine editing of an authentic personal statement. Our study protocol did not include this, but this could be considered for future research. In addition, because we felt that the personal statement was one singular aspect in an application packet, we did not ask the faculty reviewers if they would consider extending an interview or ranking the author of the statement to match into their program. Finally, some readers may question the necessity or purpose of this study. GAI is now widely and publicly available. While certain AI tools, such as dictation or grammar editing software has been in use for years, GAI has added a new level of technology that should be used with caution. GAI use has some ethical concerns, specifically that it may change the authors intent, has known biases, and could potentially misrepresent the authors intent, which raises authenticity concerns. In mediums such as an applicants' per-

sonal statement, this could be potentially construed as dishonest. In the medical and scientific field, where academic and professional standards of integrity are important, these concerns warrant attention.

In conclusion, it is important for faculty and decision-makers in orthopaedic surgery education to be aware of GAI in writing and consider their potential impacts in the application process. Our study highlights the capabilities of AI and the need for ongoing scrutiny and adaptation in our selection and application evaluation processes. Future studies may need to further investigate the ethical-based and practical-based implications of AI-generated writing and explore strategies to ensure authenticity in the application process.

Appendix

 Supporting material provided by the authors is posted with the online version of this article as a data supplement at [jbjs.org \(http://links.lww.com/JBJSOA/A683\)](http://links.lww.com/JBJSOA/A683). This content was not copy-edited or verified by JBJS. ■

Zachary C. Lum, DO^{1,2}
Lohitha Guntupalli, BS¹
Augustine M. Saiz, MD²
Holly Leshikar, MD²
Hai V. Le, MD²
John P. Meehan, MD²
Eric G. Huish, DO³

¹Department of Surgery, Kiran Patel School of Osteopathic and Allopathic Medicine, Nova Southeastern University, Davie, Florida

²Department of Orthopaedic Surgery, School of Medicine, University of California: Davis, Sacramento, California

³San Joaquin General Hospital, French Camp, California

E-mail address for H. Le: haile@ucdavis.edu

References

- Chen AF, Secrist ES, Scannell BP, Patt JC. Matching in orthopaedic surgery. *J Am Acad Orthop Surg*. 2020;28(4):135-44.
- Mun F, Jeong S, Juliano PJ, Hennrikus WL. Perceptions of USMLE step 1 pass/fail score reporting among orthopedic surgery residency program directors. *Orthopedics*. 2022;45(1):e30-e34.
- White-Dzuro CG, Makhoul AT, Pontell ME, Stephens BF II, Drolet BC, Abtahi AM. Perspectives of orthopedic surgery program directors on the USMLE step 1 scoring change. *Orthopedics*. 2022;45(5):e257-e262.
- Kang HP, Robertson DM, Levine WN, Lieberman JR. Evaluating the standardized letter of recommendation form in applicants to orthopaedic surgery residency. *J Am Acad Orthop Surg*. 2020;28(19):814-22.
- Samade R, Balch Samora J, Scharshmidt TJ, Goyal KS. Use of standardized letters of recommendation for orthopaedic surgery residency applications: a single-institution retrospective review. *J Bone Joint Surg Am*. 2020;102(4):e14.
- Max BA, Gelfand B, Brooks MR, Beckerly R, Segal S. Have personal statements become impersonal? An evaluation of personal statements in anesthesiology residency applications. *J Clin Anesth*. 2010;22(5):346-51.
- Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Machine Learn Res*. 2020;21:1-67.
- Thoppilan R, De Freitas D, Hall J, Shazeer N, Kulshreshtha A, Cheng H, Jin A, Bos T, Baker L, Du Y, Li YG, Lee H, Zheng H, Ghafouri A, Menegali M, Huang Y, Krikun M, Lepikhin D, Qin J, Chen D, Xu Y, Chen Z, Roberts A, Bosma M, Zhao V, Zhou Y, Chang C-C, Krivokon I, Rusch W, Pickett M, Srinivasan P, Man L, Meier-Hellstern K, Morris MR, Doshi T, DelosSantos R, Duke T, Soraker J, Zevenbergen B, Prabhakaran V, Diaz M, Hutchinson B, Olson K, Molina A, Hoffman-John E, Lee J, Aroyo L, Rajakumar R, Butryna A, Lamm M, Kuzmina V, Fenton J, Cohen A, Bernstein R, Kurzweil R, Aguerar-Arcas B, Cui C, Croak M, Chi E. LaMDA: language models for dialog applications. Available at: <https://arxiv.org/pdf/2201.08239.pdf>. Accessed May 29, 2023.
- Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler M, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language models are few-shot learners. Available at: <https://arxiv.org/pdf/2005.14165.pdf>. Accessed May 29, 2023.
- Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, Pearson AT. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digit Med*. 2023;6(1):75.
- Zumsteg JM, Junn C. Will ChatGPT match to your program? *Am J Phys Med Rehabil*. 2023;102(6):545-7.
- Metz C. The secret ingredient of ChatGPT is human advice. *The New York Times*. 2023. Available at: <https://www.nytimes.com/2023/09/25/technology/chatgpt-rlhf-human-tutors.html?smid=url-share>. Accessed January 11, 2024.