Received: 25 June 2023 | Revised: 16 September 2023 | Accepted: 18 September 2023

DOI: 10.1002/aet2.10924

AEM Education and Training

ORIGINAL CONTRIBUTION

Brain versus bot: Distinguishing letters of recommendation authored by humans compared with artificial intelligence

Carl Preiksaitis MD[1] | Christopher Nash MD, EdM[2] | Michael Gottlieb MD[3] | Teresa M. Chan MD, MHPE[4] | Al'ai Alvarez MD[1] | Adaira Landry MD[5]

[1]Department of Emergency Medicine, Stanford School of Medicine, Stanford, California, USA

[2]Department of Emergency Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA

[3]Department of Emergency Medicine, Rush University Medical Center, Chicago, Illinois, USA

[4]Division of Emergency Medicine, Department of Medicine, McMaster University, Hamilton, Ontario, Canada

[5]Department of Emergency Medicine, Harvard Medical School, Boston, Massachusetts, USA

Correspondence

Carl Preiksaitis, MD, Department of Emergency Medicine, Stanford School of Medicine, 900 Welch Road, Suite 350, Palo Alto, CA 94304, USA.
Email: cpreiksaitis@stanford.edu

Abstract

Objectives: Letters of recommendation (LORs) are essential within academic medicine, affecting a number of important decisions regarding advancement, yet these letters take significant amounts of time and labor to prepare. The use of generative artificial intelligence (AI) tools, such as ChatGPT, are gaining popularity for a variety of academic writing tasks and offer an innovative solution to relieve the burden of letter writing. It is yet to be determined if ChatGPT could aid in crafting LORs, particularly in high-stakes contexts like faculty promotion. To determine the feasibility of this process and whether there is a significant difference between AI and human-authored letters, we conducted a study aimed at determining whether academic physicians can distinguish between the two.

Methods: A quasi-experimental study was conducted using a single-blind design. Academic physicians with experience in reviewing LORs were presented with LORs for promotion to associate professor, written by either humans or AI. Participants reviewed LORs and identified the authorship. Statistical analysis was performed to determine accuracy in distinguishing between human and AI-authored LORs. Additionally, the perceived quality and persuasiveness of the LORs were compared based on suspected and actual authorship.

Results: A total of 32 participants completed letter review. The mean accuracy of distinguishing between human- versus AI-authored LORs was 59.4%. The reviewer's certainty and time spent deliberating did not significantly impact accuracy. LORs suspected to be human-authored were rated more favorably in terms of quality and persuasiveness. A difference in gender-biased language was observed in our letters: human-authored letters contained significantly more female-associated words, while the majority of AI-authored letters tended to use more male-associated words.

Conclusions: Participants were unable to reliably differentiate between human- and AI-authored LORs for promotion. AI may be able to generate LORs and relieve the burden of letter writing for academicians. New strategies, policies, and guidelines are

Present address Duke University, Durham, NC, USA

Supervising Editor: Jeffrey N Siegelman

needed to balance the benefits of AI while preserving integrity and fairness in academic promotion decisions.

INTRODUCTION

Letters of recommendation (LORs) are essential within academic medicine. They offer a thorough evaluation of a candidate's credentials, significantly impacting everything from admission decisions to considerations for promotion. Specifically, LORs for promotion and tenure (P&T) are integral in determining a candidate's eligibility for promotion in line with institutional criteria.[1] These P&T letters have long been foundational in faculty promotion determinations. However, preparing these letters is often time-consuming and labor-intensive for academics, who are usually required to write several throughout the year, further adding to their extensive workload.[1] Consequently, due to these challenges, LORs often suffer from shortcomings like generic remarks, inconsistent evaluations, and bias in assessments.[2–4]

The emergence of generative artificial intelligence (AI) technology, such as ChatGPT, could potentially offer an innovative solution to ease the burden on those tasked with letter writing. Yet, a significant research gap remains regarding AI's effectiveness and role in such professional evaluations within academia.

ChatGPT is an advanced large language model that employs machine learning algorithms to generate text of near human quality on a multitude of topics.[5] Recent iterations have excelled in knowledge benchmarks such as the Uniform Bar Examination and even produced academic writing virtually indistinguishable from human-authored work.[6,7] It is yet to be determined if ChatGPT could aid in crafting LORs, particularly in high-stakes contexts like faculty promotion. To determine the feasibility of this process and whether there is a significant difference between AI- and human-authored letters, we conducted a study aimed at determining whether academic physicians can distinguish between the two.

METHODS

Study design and selection of promotion criteria

We conducted a single-blind quasi-experimental repeated-measures study to assess the ability of academic physicians to distinguish between LORs for promotion to associate professor, written by either humans or AI. We adhered to the Strengthening the Reporting of Observational Studies in Epidemiology guidelines.[8] This study was deemed exempt by the institutional review board at Stanford University (#68837). The rank of associate professor was deliberately chosen as it represents a significant career milestone, bridging the gap between early career (assistant professor) and more advanced faculty roles (professor). Criteria for promotion to this rank were adapted from two authors' (CP, AA) home institution and are available as supplemental material accompanying the online article. The criteria categories were clinical care, scholarship, teaching, and service.

Candidate selection and data abstraction

We purposively selected four candidates, each either on the brink of applying for promotion from assistant to associate professor or having recently undergone promotion (Figure 1). Two team members (AA, AL) were responsible for anonymizing and abstracting salient accomplishments that fulfilled promotion criteria. These achievements were classified into three categories: education and training history, employment, and commendations. The four anonymized candidates were assigned pseudonyms, ranging from Dr. ChatGPT1 to Dr. ChatGPT4. Any gender identifiers were removed and letter authors were blinded to the gender of candidates. Detailed lists of achievements for each candidate are available as supplemental material accompanying the online article.

Letter of recommendation creation

Four team members composed LORs, drawing on each candidate's array of achievements. TC and MG, having collectively written over 100 LORs for promotion, were designated as human-author controls. Simultaneously, two junior team members (CP and CN) were selected to create a comparison group of letters, using the assistance of ChatGPT. Neither CN nor CP had any prior experience in P&T procedures nor had they previously written a LOR for a P&T committee. Using the GPT-3.5-turbo and GPT-4 models of ChatGPT, respectively, CN and CP generated AI-authored LORs. The AI-authored LORs were developed using prompts derived from the candidates' achievements, with reprompting available for fine-tuning (prompts and responses available as supplemental material accompanying the online article). Human additions to the text were not permitted, and formatting suggestions from ChatGPT were accepted. To ensure uniformity, all LORs adhered to the same font and overall presentation, leaving the content of the text unaltered.

Survey design and distribution

A web-based survey, created through the Qualtrics survey tool (Qualtrics, Inc.), was used to evaluate the LORs. Each participant was

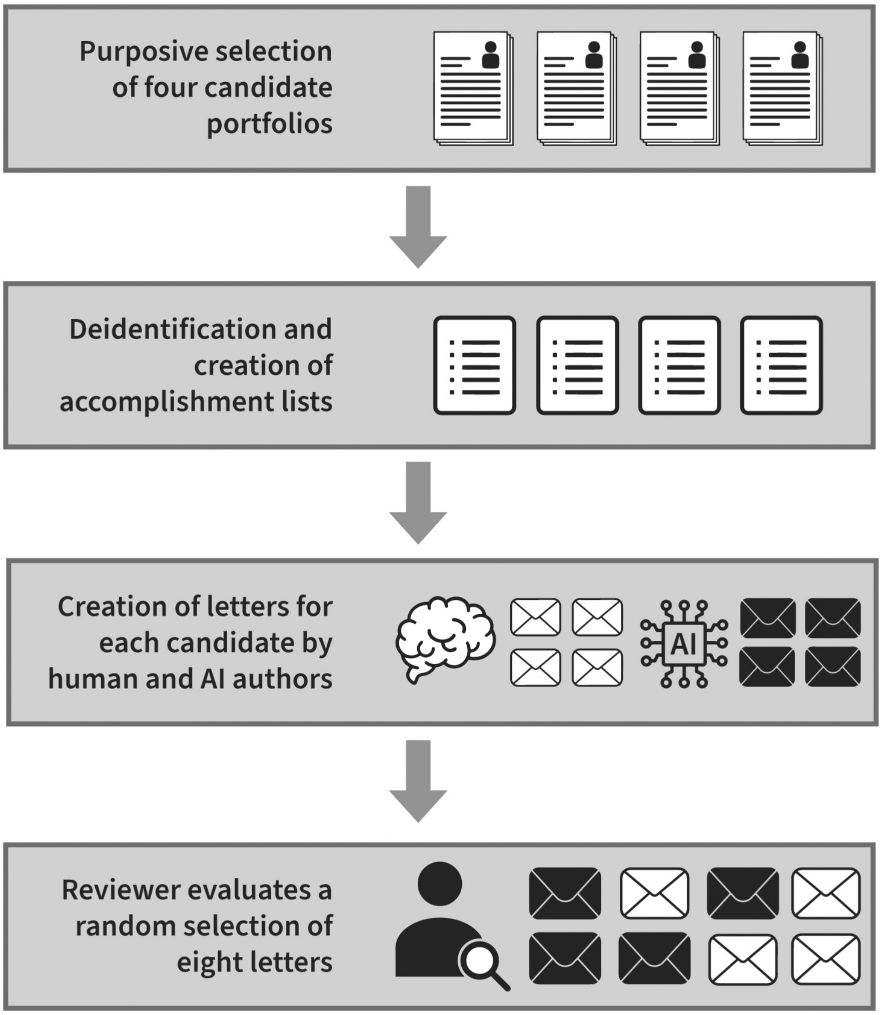
Purposive selection of four candidate portfolios
Deidentification and creation of accomplishment lists
Creation of letters for each candidate by human and AI authors
AI
Reviewer evaluates a random selection of eight letters

FIGURE 1 Methodology of letter generation and review.

randomly presented with eight out of 16 LORs in no specific order. Following each LOR, participants were asked about their perception of authorship (human-only or ChatGPT-assisted), the quality, and the persuasiveness of the LOR concerning promotion. Authorship identification was a binary choice (human vs. AI), with certainty assessed on a 0–100 scale. The quality of the LOR was evaluated on a 1–5 scale (ranging from "one of the worst" to "one of the best"), while the persuasiveness of the LOR in terms of promotion was gauged on a 1–3 scale ("decreases chances of promotion," "no change," or "increases chances of promotion"). As a proxy for deliberation about answer choices, letters and survey questions appeared on separate pages so that time required to answer questions could be differentiated from letter review time. We defined deliberation time as time from arrival on the question page to submission of that page's answer choices. Initially, the survey was tested within the research team and further refined based on their feedback. Cognitive interviews were conducted utilizing a think aloud technique to ensure response process validity, which led to additional survey refinement including clarification of language, adjusting the quality rating scale, and ensuring consistency in letter formatting.[9] An illustrative survey question is available as supplemental material accompanying the online article.

Study population

Our study population was recruited using a snowball sampling technique. The survey and recruitment materials were disseminated via academic emergency medicine (EM) listservs, via the medical education community on Twitter, and during a national EM conference. We initially recruited participants with prior experience on P&T committees and later included individuals who had prior experience reviewing LORs in any capacity (medical school, residency, fellowship, faculty selection). Subjects without P&T experience were included to allow for comparison between less experienced and more experienced letter reviewers. The candidates who shared promotion materials and study team members were excluded. Participants were required to complete a preliminary form about their experience reviewing LORs and provide consent. All responses were kept anonymous.

Primary outcome measure

The primary outcome of interest was the accuracy in distinguishing human-authored from AI-authored LORs. To establish the minimum number of participants, we conducted a power calculation using a one-sample, one-tail *t*-test, aiming to detect an ability to differentiate at least 10% better than chance with 90% power at a 0.05 significance level. This resulted in a requirement of at least 214 measurements, or a minimum of 27 participants.

Data analysis

We included only complete survey responses in our study. Preliminary analysis of the data demonstrated violation of several assumptions required for typical parametric testing; therefore, we implemented a generalized linear mixed model with a logistic link function and an unstructured covariance pattern. This allowed us to account for repeated measures and the effects of individual authors and candidates. Our outcome variable was the accurate identification of a letter, with the author and the candidate acting as fixed effects and the subject identifier as a random-effects variable. We assumed random intercepts for the random effects, considering participants as the random intercepts. We did not incorporate any random slopes in our analysis and conducted fixed-effects estimation using conditional likelihood. We verified fit using the DHARMa package (v. 0.1.2) for diagnostic tests of model residuals. Visual inspection of the simulated residuals indicated uniform distribution, suggesting good data fit. The Kolmogorov–Smirnov test confirmed this, failing to reject the null hypothesis that the residuals follow a uniform distribution ($p=0.83$). To determine participants' accuracy in discerning human- from AI-authored LORs, we used our model to simulate responses and conducted a one-sample t-test against a null hypothesis with a mean accuracy of 50%.

Our secondary analysis compared perceived quality and persuasiveness of promotion LORs based on suspected and actual LOR sources using linear mixed-effects models. We estimated and contrasted mean values of quality and persuasiveness using estimated marginal means from the fitted models. Additionally, we studied deliberation time and reported certainty in the identification to explore their effect on accuracy using further generalized linear mixed models. All statistical analyses were conducted using R statistical software, version 4.3.0 (R Foundation for Statistical Computing).

We used a publicly available calculator for gender-biased language (https://slowe.github.io/genderbias/) to evaluate each letter for signs of gender-biased language. This tool has been used in past research on gender bias in LORs.[10] The calculator determines the percentage of bias toward male or female language by counting words commonly associated with either gender, derived from a corpus of LORs for male and female biochemistry job applicants.[11] The calculator's results range from 100% male-biased language to 100% female-biased language, with 0% indicating neutrality. For instance, a document with 10% male-biased language contains more male-associated than female-associated words. In contrast, one with 5% female-biased language contains a majority of female-associated words, but the magnitude of this bias is half as strong as the 10% bias toward male-associated words. For detection of AI-authored text, we used two AI-based tools: GPTZero (https://gptzero.me/) and OpenAI's Text Classifier (https://platform.openai.com/ai-text-classifier). GPTZero calculates text *perplexity* and *burstiness*. These metrics quantify the randomness of a sequence of words and variation in length and structure of sentences, respectively. OpenAI's AI Text Classifier was trained to distinguish between human-written and AI-written text using a large data set of pairs of human and AI text. We applied these tools to each LOR to identify differences in gender-biased language and compare AI detection tool performance to our human results.

RESULTS

Participant characteristics and survey response rate

Out of the 44 subjects who began the survey, 32 participants (72.7%) completed all reviews. The majority of these participants were full professors, followed by associate professors and assistant professors. Three quarters of participants identified as male. The primary specialty of the majority of participants was EM, followed by internal medicine and family medicine. Half of the participants had served on P&T committees in their careers. Their experience was evenly divided between having less than 1 year, 2 to 5 years, or more than 5 years. Participants with experience reviewing letters in other capacities most commonly reviewed letters for residency, fellowship, and faculty selection. Most participants reported reviewing over 100 LORs during their careers. Specific figures can be found in Table 1.

Accuracy of differentiating between AI- and human-authored letters

On average, participants correctly identified whether a letter of recommendation was human or AI-authored 59.4% of the time (95% confidence interval [CI] 58.6%–60.2%). In the model, only one human LOR author (Human 1) significantly predicted accurate identification, having an odds ratio of 2.41 (95% CI 1.14–5.07) compared to the reference AI Author 1. The individual Dr. ChatGPT candidate persona did not significantly predict accurate identification. Furthermore, deliberation time did not have a significant effect on accuracy (regression coefficient 0.011 [95% CI −0.016 to 0.037], $p=0.43$). Additionally, reviewer certainty did not significantly predict accuracy (regression coefficient 0.990 [95% CI −0.849 to 2.83], $p=0.29$). Interestingly, even among the subset of participants with extensive experience in reviewing LORs (more than 100 letters), there was no substantial improvement in accuracy (regression coefficient − 0.136 [95% CI −0.718 to 0.445], $p=0.64$).

Quality and persuasiveness of letters

The perceived quality and persuasiveness of LORs for promotion seemed to be significantly impacted by the reviewers' assumptions

TABLE 1 Characteristics of study participants ($n = 32$).

Characteristic	$n = 32$
Gender	
Male	24 (75)
Female	8 (25)
Rank	
Assistant professor	5 (16)
Associate professor	13 (41)
Professor	14 (44)
Specialty	
EM	23 (72)
Internal medicine	4 (13)
Family medicine	3 (9)
Anesthesiology	1 (3)
Pediatrics	1 (3)
Physical medicine and rehabilitation	1 (3)
P&T experience	16 (50)
<1 year	4 (13)
2–5 years	6 (19)
>5 years	6 (19)
Other letter-reviewing experience	16 (50)
Medical school selection	5 (16)
Residency selection	14 (44)
Fellowship selection	12 (38)
Faculty selection	12 (38)
Number of letters reviewed	
<20	3 (9)
20–50	4 (13)
50–100	5 (16)
>100	20 (63)

Note: Data are reported as *n* (%). All participants are academic physicians with experience in LOR review.

Abbreviations: LOR, letter of recommendation; P&T, promotion and tenure.

concerning the authorship of the letters. When reviewers believed that a human had written a LOR, they consistently assigned higher ratings for quality compared to those LORs thought to be composed by AI (estimated mean rating 2.49 [95% CI 2.35–2.62] vs. 1.72 [95% CI 1.58–1.87], $p < 0.001$). However, when assessing LORs based on their actual source of authorship, the perceived quality difference between human and AI-written LORs became statistically insignificant (estimated mean rating 2.21 [95% CI 2.05–2.37] vs. 2.08 [95% CI 1.92–2.24], $p = 0.17$). A parallel pattern was observed when evaluating the persuasiveness of LORs. LORs perceived as human-authored were deemed more persuasive than those considered to be AI-authored (estimated mean rating 1.79 [95% CI 1.66–1.92] vs. 1.31 [95% CI 1.17–1.44], $p < 0.001$). Yet, when the actual authorship was taken into account, the difference in persuasiveness ratings diminished and was not statistically significant (estimated mean rating 1.61 [95% CI 1.46–1.75] vs. 1.54 [95% CI 1.40–1.68], $p = 0.34$) (Figure 2).

Other evaluations: Gender-biased language and AI-detection tools performance

Results from the Gender Bias Calculator revealed that human-authored letters contained a higher percentage of female-associated words. Specifically, the median score was 29.5% female-associated words (IQR 15.3% female-associated words to 33% female-associated words). Contrarily, the majority of AI-authored letters exhibited more male-associated words, with five out of eight letters (63%) showing a median of 3% male-associated words (IQR 6% male-associated words to 11.5% female-associated words; Table 2).

Assessments of AI detection tools demonstrated varying results. GPTZero misclassified all eight AI-authored LORs as human-authored. However, it correctly identified all eight human-authored LORs, for an overall accuracy of 50%. Despite the misclassification, there was a significant difference in the average "perplexity" (517 vs. 222, $p < 0.001$) and "burstiness" (1472 vs. 354, $p = 0.02$) scores between AI-authored and human-authored letters. OpenAI's AI Text Classifier labeled two out of eight AI-authored LOR as "very unlikely AI-generated" and the remaining six as "unlikely AI-generated." All eight of the human-authored LORs were classified as "very unlikely AI-generated," which similarly resulted in an accuracy of 50% (Table 2).

DISCUSSION

In our study, participants were unable to reliably differentiate between human-authored and AI-authored LORs for promotion in academic medicine, with an average accuracy of only 59.4%. Neither reviewer's certainty in the authorship of the letter nor the deliberation time significantly influenced the accuracy of identification. As to our motivation for this study, these findings suggest that AI may be able to partially or completely generate LORs for promotion and relieve the burden of letter writing for academicians.

The potential role for AI in crafting LORs has been previously suggested; however, this is the first study that demonstrates the feasibility of this application for AI.[12,13] Human-authored letters have documented, problematic tendencies toward bias, and AI is suggested as a potential solution to reducing bias in LOR, though contrary arguments and data exist.[3,11,14–18] Although our study only examined biased language related to gender and was not powered to examine this aspect in detail, we observed signal of gender-biased language in both human and AI-authored LORs. Current and future work in understanding AI systems' bias will hopefully further determine whether AI can mitigate or exacerbate bias in letter writing.

The potential role for AI to assist in letter writing further highlights a need for increased education among faculty in how to use AI safely and effectively. Although AI may be an aid for more rapid generation

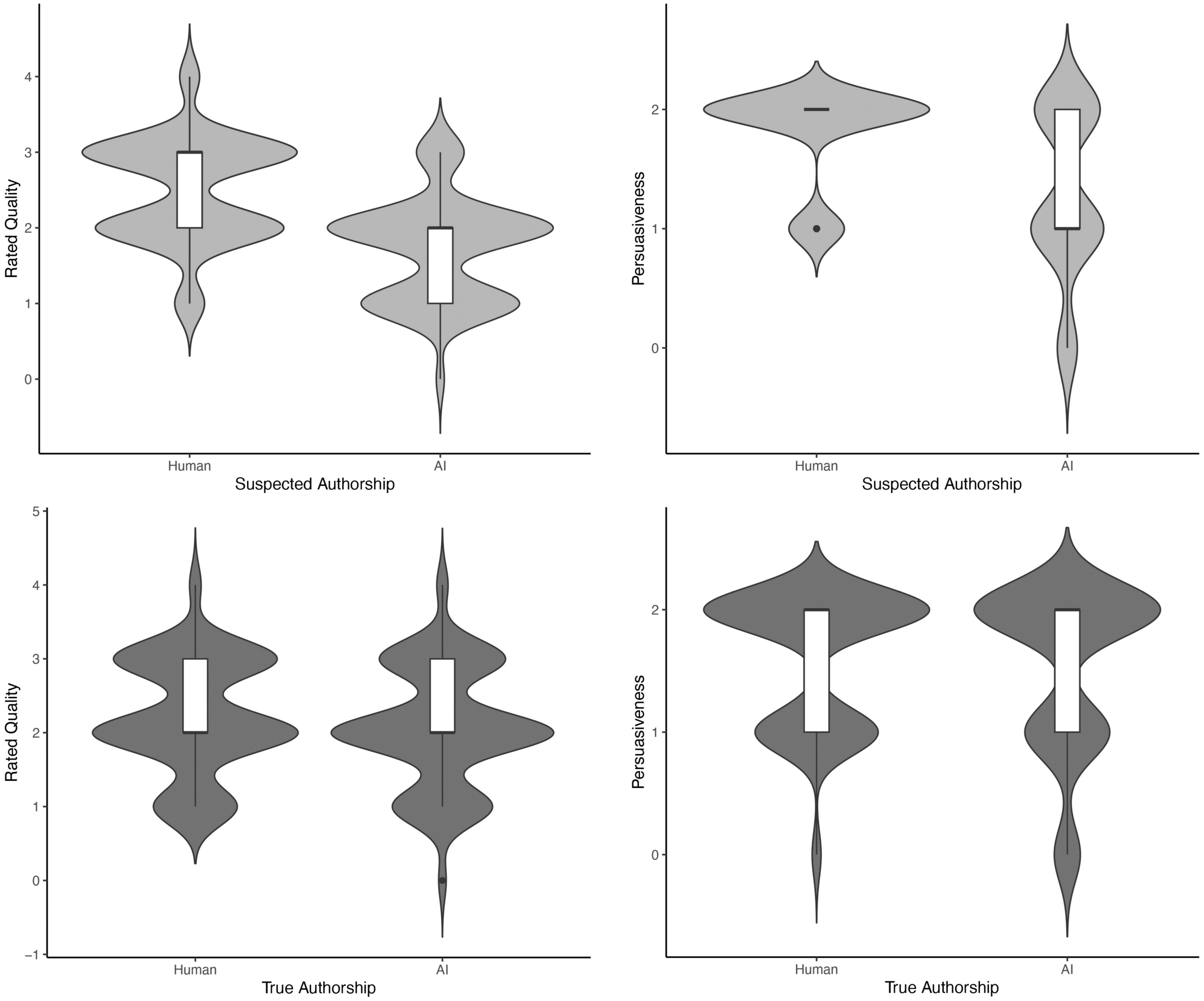

FIGURE 2 Violin plot comparing perceived letter quality and persuasiveness based on perceived and actual authorship as determined by academic physicians with experience in reviewing LORs (n = 32). Quality: 0 = "one of the worst," 1 = "below average," 2 = "average," 3 = "above average," 4 = "one of the best"; persuasiveness: 0 = "decreases chances of promotion," 1 = "no change," 2 = "increases chances of promotion." LORs, letters of recommendation.

of letters, faculty members must be aware of the potential for errors and inaccuracies, notably the phenomenon referred to as "hallucinations."[19 20] Furthermore, the quality of output from large language models is contingent on the quality of input data and instructions (prompts) to the models. Training in how to effectively engineer effective prompts may improve the quality of letters that they are able to generate.[21]

Ethically, faculty members or institutions will need to determine how the use of AI is disclosed, if at all. The degree to which faculty members are responsible for the content of letters that are generated by AI will need to be determined, although ethical debates surrounding letter writing are not new.[2,22,23] It is possible that the bias against AI-authored letters we observed was amplified by reviewer knowledge that AI was used for letter generation. This finding may complicate the decision to disclose whether AI was used for letter generation. Based on our results regarding the use of AI-detection software, disclosure appears to be the only reliable way to identify AI-generated letters at this time.

More broadly, our findings demonstrate a threat to the current status quo of the letter-writing process. If a letter of recommendation for a candidate can be generated by AI simply from their CV and personal statement, what new or additional information can it offer the review committee? Our results are considered a drive to change the system currently in place. We went to great lengths to anonymize the fictional promotion candidates and remove the possibility of any personalization, which may not be true of letters in reality. However, many letter writers are provided with only promotion packet materials and are deliberately selected to not have significant personal knowledge of the candidate, similar to the paradigm of our study. Some literature suggests that referees offer their opinions and perspectives on the accomplishments and reputation of candidates;

TABLE 2 Letter-specific results of gender bias calculator (https://slowe.github.io/genderbias/), AI-detector software (https://gptzero.me/; https://platform.openai.com/ai-text-classifier), and classification by academic physicians with experience in LOR review ($n = 32$).

			GPTZero				
Author	Candidate	Percent gender-biased language	Classification	"Perplexity"	"Burstiness"	OpenAI classification	Reviewer classification
AI 1	Dr. Chat GPT1	2% male	Human	195.529	255.381	Very unlikely AI	46.7% AI
AI 1	Dr. Chat GPT2	20% male	Human	148.611	130.822	Very unlikely AI	43.8% AI
AI 1	Dr. Chat GPT3	5% male	Human	144.647	149.835	Unlikely AI	52.6% AI
AI 1	Dr. Chat GPT4	9% male	Human	180.529	260.337	Unlikely AI	43.8% AI
AI 2	Dr. Chat GPT1	38% female	Human	295.56	530.609	Unlikely AI	60% AI
AI 2	Dr. Chat GPT2	16% female	Human	319.68	534.872	Unlikely AI	64.3% AI
AI 2	Dr. Chat GPT3	10% female	Human	249.7	462.411	Unlikely AI	62.5% AI
AI 2	Dr. Chat GPT4	4% male	Human	244.708	511.034	Unlikely AI	50.0% AI
Human 1	Dr. Chat GPT1	33% female	Human	283.688	467.376	Very unlikely AI	44.4% AI
Human 1	Dr. Chat GPT2	3% female	Human	297.343	456.176	Very unlikely AI	40.0% AI
Human 1	Dr. Chat GPT3	19% female	Human	314.303	485.842	Very unlikely AI	26.7% AI
Human 1	Dr. Chat GPT4	14% female	Human	319.294	485.003	Very unlikely AI	31.3% AI
Human 2	Dr. Chat GPT1	33% female	Human	649.5	2480.154	Very unlikely AI	29.4% AI
Human 2	Dr. Chat GPT2	27% female	Human	750.651	2458.046	Very unlikely AI	38.9% AI
Human 2	Dr. Chat GPT3	32% female	Human	779	2515.031	Very unlikely AI	44.4% AI
Human 2	Dr. Chat GPT4	38% female	Human	743.214	2426.486	Very unlikely AI	37.5% AI

Abbreviations: AI, artificial intelligence; LOR, letter of recommendation; P&T, promotion and tenure.

however, it is not clear if this is the norm or whether this information is useful to P&T committees.[24] In a broader sense, several authors have commented on the lack of guidance and training that faculty members receive in how to craft letters.[1,4,24] All of these aspects demonstrate potential problems or inconsistencies with the LOR process as a whole. Perhaps with clearer instructions and further guidance to faculty members, letters may focus more on aspects not evident in other elements of the candidate packet, which may be of more use to a P&T committee and may merit investigation of its own. Deliberations about using AI to write letters may offer an opportunity for P&T committees to clarify their goals and expectations for letter writers to ensure the data they receive is valuable.

Taken further, the rising use of AI in academic writing may be an opportunity to consider a different approach for the recommendation process, such as adopting a standardized LOR format, which has provided a more useful alternative to narrative letters in some contexts.[25,26] The challenge of how and whether to use AI in recommendation letters might present an opportunity to reconsider the utility, fairness, and transparency of including LORs in candidate evaluation for hiring or promotion.

LIMITATIONS

This study is not without limitations. First, our approach involved the use of standardized data for all letters, which was presented in a summary format to the letter writers (see supplemental material accompanying the online article). In real-world scenarios, awareness of the candidate's reputation, personal identifiers (such as gender and race), and history may lead to more specific comments that are unique to letters authored by humans and make them more recognizable. Second, our recruitment strategy could introduce bias and compromise data reliability. Specifically, our letter reviewers were not randomly sampled, leading to an overrepresentation of male gender and EM physician reviewers. Additionally, we could not independently verify the identities of the respondents, which may affect the reliability of our data. Our reviewers reported a range of experience in letter review, and half of them had previous experience on P&T committees. We included a diversity of reviewers to see whether there was any signal toward greater accuracy in more experienced reviewers. There did not appear to be evidence of difference in review ability; however, our study was not powered to detect this, and different results may be found with a more experienced reviewer pool. Finally, our study did not explore the factors reviewers used to distinguish between human- and AI-authored LORs. We attempted to ascertain the difficulty in discerning between the letter types by recording time spent on the question page (deliberation time) and reviewer certainty; however, the accuracy of these proxies is limited by our data collection abilities in our survey. Future research could endeavor to measure the factors related to differentiation more comprehensively.

CONCLUSIONS

This study illustrated that academic physicians have only a slightly better ability than chance at identifying AI-authored LORs. This finding brings promise that some of the administrative burden of writing LORs could be offloaded to AI in the future. As AI continues to advance and integrate into various facets of professional and academic workflows, concrete strategies are needed to balance the benefits of AI while preserving integrity and fairness in academic promotion decisions. Furthermore, the provocative abilities of AI in this space may offer an opportunity to reflect on existing practices and reexamine the value, equity, and transparency of letters of recommendation for promotion.

AUTHOR CONTRIBUTIONS

All authors contributed to the conception and design of the study. Al'ai Alvarez and Adaira Landry were responsible for preparation of letter materials. Carl Preiksaitis, Christopher Nash, Teresa M. Chan, and Michael Gottlieb created the letters used for review. Carl Preiksaitis was responsible for data collection, analysis, and statistical expertise. Study procedures were overseen by Adaira Landry, Michael Gottlieb, and Teresa M. Chan. Carl Preiksaitis drafted the manuscript and all authors significantly contributed to its revision. Carl Preiksaitis takes responsibility for the manuscript as a whole.

ACKNOWLEDGMENTS

The authors thank the following individuals who served as article reviewers in the study: Laryssa Patti, MD; Harsh Sule, MD, MPP; Tom Morrissey, MD, PhD; Robert Steele, MD, MBA; Lars Beattie, MD, MS; Steven Nazario, MD; Craig Krausz, MD; Jonathan Sherbino, MD, MEd; Pinaki Mukherji, MD; Moshe Weizberg, MD; Julien Poitras, MD; Tiffany Moadel, MD; Loice, Anne Swisher, MD; Michele L. Dorfsman, MD; Todd Taylor, MD; Moira Davenport, MD; Chris Belcher, MD; Elizabeth Dubey, MD; Matthew Silver, MD; Umbereen S. Nehal, MD, MPH; Eddy Lang, MD; Sally Henin, MD; Cullen Hegarty, MD; David Jones, MD; Christina Shenvi, MD, PhD, MBA; and Stephen John Cico, MD, MEd.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ORCID

Carl Preiksaitis https://orcid.org/0000-0002-3856-0068
Christopher Nash https://orcid.org/0000-0002-0738-409X
Michael Gottlieb https://orcid.org/0000-0003-3276-8375
Teresa M. Chan https://orcid.org/0000-0001-6104-462X
Al'ai Alvarez https://orcid.org/0000-0002-5438-2476
Adaira Landry https://orcid.org/0000-0002-5299-679X

REFERENCES

1. Gottlieb M, Chan TM, Yarris LM, Linden JA, Coates WC. Promotion and tenure letters: a guide for faculty. *AEM Educ Train*. 2022;6(3):e10759. doi:10.1002/aet2.10759
2. Roberts LW, Termuehlen G. (Honest) letters of recommendation. *Acad Psychiatry*. 2013;37(1):55-59.
3. Trix F, Psenka C. Exploring the color of glass: letters of recommendation for female and male medical faculty. *Discourse Soc*. 2003;14(2):191-220.
4. Agarwal V, Mullins ME, Mainiero MB, Suh RD, Chetlen AL, Lewis PJ. The ADVICER template for faculty reviewer letters for promotion and appointment. *Acad Radiol*. 2022;29(9):1413-1416.
5. Gottlieb M, Kline JA, Schneider AJ, Coates WC. ChatGPT and conversational artificial intelligence: friend, foe, or future of research? *Am J Emerg Med*. 2023;70:81-83.
6. OpenAI. GPT-4 Technical Report. arXiv.org, Cornell University. 2023. Accessed June 1, 2023. https://arxiv.org/abs/2303.08774v3.
7. Gao CA, Howard FM, Markov NS, et al. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *npj Digit Med*. 2023;6:75. doi:10.1101/2022.12.23.521610v1
8. von Elm E, Altman DG, Egger M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med*. 2007;147(8):573-577.
9. Willis GB, Artino AR Jr. What do our respondents think We're asking? Using cognitive interviewing to improve medical education surveys. *J Grad Med Educ*. 2013;5(3):353-356.
10. Chapman BV, Rooney MK, Ludmir EB, et al. Linguistic biases in letters of recommendation for radiation oncology residency applicants from 2015 to 2019. *J Cancer Educ*. 2022;37(4):965-972.
11. Schmader T, Whitehead J, Wysocki VH. A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex Roles*. 2007;57(7–8):509-514.
12. Leung TI, Sagar A, Shroff S, Henry TL. Can AI mitigate bias in writing letters of recommendation? *JMIR Med Educ*. 2023;9(1):e51494.
13. Bogost I. The end of recommendation letters [Internet]. *Atlantic*. 2023 [cited 2023 Sep 1];Available from. https://www.theatlantic.com/technology/archive/2023/04/chatgpt-ai-college-professors/673796/
14. A4BL Anti-racist tenure letter working group. A guide for writing anti-racist tenure and promotion letters. *Elife* 2022;11:e79892.
15. Madera JM, Hebl MR, Dial H, Martin R, Valian V. Raising doubt in letters of recommendation for academia: gender differences and their impact. *J Bus Psychol*. 2019;34(3):287-303.
16. Mittermaier M, Raza MM, Kvedar JC. Bias in AI-based models for medical applications: challenges and mitigation strategies. *NPJ Digit Med*. 2023;6(1):113.
17. Drage E, Mackereth K. Does AI debias recruitment? Race, gender, and AI's "eradication of difference". *Philos Technol*. 2022;35(4):89.
18. Preiksaitis C, Sinsky CA, Rose C. Chatgpt is not the solution to physicians' documentation burden. *Nat Med*. 2023;29:1-2.
19. Emsley R. ChatGPT: these are not hallucinations – they're fabrications and falsifications. *Schizophrenia*. 2023;9(1):1-2.
20. Salvagno M, Taccone FS, Gerli AG. Artificial intelligence hallucinations. *Crit Care*. 2023;27(1):180.
21. Giray L. Prompt engineering with ChatGPT: a guide for academic writers. *Ann Biomed Eng*. 2023.
22. Larkin GL, Marco CA. Ethics seminars: beyond authorship requirements-ethical considerations in writing letters of recommendation. *Acad Emerg Med*. 2001;8(1):70-73.
23. Balon R. What about letters in support of academic promotion? *Acad Psychiatry*. 2013;37(2):142-143.
24. Dunnick NR. Letters of recommendation for promotion. *Acad Radiol*. 2022;29(1):1-3.
25. Girzadas DV, Harwood RC, Dearie J, Garrett S. A comparison of standardized and narrative letters of recommendation. *Acad Emerg Med*. 1998;5(11):1101-1104.

26. Jackson JS, Bond M, Love JN, Hegarty C. Emergency medicine standardized letter of evaluation (SLOE): findings from the new electronic SLOE format. *J Grad Med Educ*. 2019;11(2):182-186.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Preiksaitis C, Nash C, Gottlieb M, Chan TM, Alvarez A, Landry A. Brain versus bot: Distinguishing letters of recommendation authored by humans compared with artificial intelligence. *AEM Educ Train*. 2023;7:1-9. doi:10.1002/aet2.10924