Should my recommendation letter be written by artificial intelligence?

Jad Mansour, MD
Mark Burman, MD
Mitchell Bernstein, MD
Emilie Sandman, MD
Kaisssar Yammine, MD
Mohammad Daher, BSc
Paul Andre Martineau, MD

Correspondence to:
J. Mansour
McGill University Health Centre
1001 Decarie Blvd
Montréal QC H4A 3J1
Jad.mansour09@gmail.com

Letters of recommendation are increasingly important for the residency match. We assessed whether an artificial intelligence (AI) tool could help in writing letters of recommendation by analyzing recommendation letters written by 3 academic staff and AI duplicate versions for 13 applicants. The preferred letters were selected by 3 blinded orthopedic program directors based on a predetermined set of criteria. The first orthopedic program director selected the AI letter for 31% of applicants, and the 2 remaining program directors selected the AI letter for 38% of applicants, with the staff-written versions selected more often by all of the program directors (p < 0.05). The first program director recognized only 15% of the AI-written letters, the second was able to identify 92%, and the third director identified 77% of AI-written letters (p < 0.05).

SELECTING QUALIFIED CANDIDATES FOR MEDICAL SCHOOL OR RESIDENCY PROGRAMS IS A CRUCIAL AND CHALLENGING Undertaking that determines a person’s career track and future. Recent modifications in medical education and training — such as the pass/fail concept in board exams and the limited availability of clerkships — have made it more challenging to establish objective standards that distinguish different applicants. Therefore, other aspects of the residency application have gained importance in differentiating between students. One of the most crucial elements taken into account by residency programs and their program directors is the letter of recommendation, which provides important supplementary content, offering a perspective of a candidate’s skills and traits from a reliable source. These letters are usually submitted by mentors and educators, including faculty members or supervisors who worked closely with the candidate. They can provide important details and features of a person’s clinical abilities, work ethic, and promise as a future doctor.

The medical community is reaping the benefits of the recent advancements in artificial intelligence (AI), particularly chatbot technology. Chatbots are now used to aid in activities such as patient education, appointment scheduling, and symptom management. ChatGPT, a language model developed by OpenAI, has rapidly gained popularity since its launch on Nov. 30, 2022. This model can quickly generate content and respond to inquiries quickly and efficiently, making it an ideal tool for assisting medical professionals to write well-structured and informative articles, book chapters, or recommendation letters. The effect of this technology will be felt in the medical community for many years to come and its use can have a profound effect on medical writing or academic promotions, particularly for elements such as cover or recommendation letters.

We evaluated the efficacy of AI language models in writing recommendation letters for doctors applying to orthopedic residency programs, compared with conventional letters written by academic staff. We also explored the ability of program directors to differentiate staff-written letters from those written by the AI chatbot. We used a prospective series of recommendation letters written by 3 academic staff from the same university hospital in Montréal,
Canada, between January 2019 and January 2023. We collected 13 recommendation letters and created AI versions of the letters using ChatGPT 4.0. The duplicates contained detailed, objective content that appeared relevant and pertinent to the applicant, such as the number of research years or personal experiences with the mentor or staff (Figure 1).

The duplicates were all created by a single orthopedic surgeon (J.M.) and all applicants’ names, letterheads, and contact information were removed from both versions to eliminate any decision bias. We included 2 letters for each of the 13 applicants ($n = 26$ letters total), 1 written by academic staff and 1 written by the AI chatbot.
The study consisted of 2 parts (Figure 2). The first part requested 3 orthopedic program directors from 3 different universities to choose a preferred letter from the 2 versions for each applicant, without the awareness that 1 of those versions was written by an AI robot. Program directors used a predetermined set of criteria of selection set by each university program (based on clinical knowledge and experience, publications, and personality) to select their preferred letters of reference. In the second part of the study, we informed the same program directors that 1 version of each applicant’s letters was written by ChatGPT and asked them to judge whether each letter was written by either an attending staff or the AI model.

We analyzed data using StatsDirect software. For continuous variables, we reported mean values with their standard deviations (SDs). We performed univariate and multivariate regression analyses. Significance was set for p values of less than 0.05.

In the first part of the study, 1 orthopedic program director preferred the AI-written letters for 4 applicants (31%) and preferred the staff-written letters for the remaining 9 applicants (69%) (p < 0.05). The 2 remaining program directors each selected 5 (38%) letters written by the AI robot and 8 (62%) written by attending staff (p < 0.05) (Table 1). The AI-written letters that were selected by the 3 program directors were all different, meaning that AI versions were not selected by more than 1 program director.

In the second part of the study, in which program directors were informed about an AI version, the director who selected 4 AI versions in the first part identified only 2 AI-written versions of the 13 (15%). The second program director identified 12 (92%) and the third director identified 10 (77%) of the 13 AI-written versions (Table 1).

On average, 36% of AI-written recommendation letters were selected (14 times of the 36 total selections) and 67% were recognized as written by AI (24 of 36) (p < 0.05).

We found that around one-third of recommendation letters written by ChatGPT were selected by the program directors. In addition, this result was consistent and repetitive in the choices of all 3 program directors as all 3 of them selected 4–5 of 13 AI-written recommendation letters. These choices can be attributed to the fact that the AI chatbot provides categorical, articulate, and well-organized recommendation letters, which, in certain applications, could be written in better English and in a more appealing structure than a letter written by a human. In addition, the AI-written letters selected by the different program directors were all different, meaning there was no consistency in whether an AI- or human-written letter was chosen. This suggests that none of the recommendation letters written by the AI language model was superior and that all of the letters were comparable in quality. Nevertheless, this AI tool presents many limitations, including the generation of content with incorrect citations and details, as well as the risk of bias and plagiarism. Those limitations can rationalize the description of ChatGPT as a black-box technology.7-8

<table>
<thead>
<tr>
<th>Program director</th>
<th>Written by artificial intelligence</th>
<th>Written by staff</th>
<th>No. of identified letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>8</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 1. Summary of the letters chosen and identified by program directors.

FIG. 2. Flowchart of the study design. AI = artificial intelligence.
The probability of an AI-written letter being identified is also important. A recently published article that studied the ability of medical reviewers to identify medical abstracts written by ChatGPT revealed that 67% of those abstracts were labelled as fakes and identified as having been written by a robot. These results align with the second part of our study, in which 67% of AI-written letters were identified once program directors were informed that each applicant had an AI-written letter. This means that, even though a letter written by an AI language model could be more appealing than a human-written letter, if program directors are aware of such AI models, there is a strong possibility of those letters being recognizable. Moreover, the program director who was able to recognize only 2 of the 13 AI-written recommendation letters in the second part of the study was not previously aware of the existence of such AI technology and the true capability of ChatGPT in particular. This may have explained the discrepancy between this program director and the 2 others who had been previously exposed to this technology. When asked about the method of identification of the AI-written letters, program directors mentioned a similar writing pattern between all the different letters. An analogous outline of writing, as well as identical, proficient words (e.g., wholeheartedly, preceptor) can guide program directors in identifying the AI-written letters. It is important for AI letters to be identifiable to push those responsible for letters to not delegate this task to such programs. Letters of recommendation should be personalized and truthful. They should be written by a human who knows and has interacted with the concerned person for a certain period of time.

Artificial intelligence will now be involved in many aspects of health care and medical education systems. Program directors and admission committees should be aware of the presence of such technologies and learn to be able to identify any recommendation or cover letter written by an AI language model. A solution could be to involve more direct and live interview systems to differentiate between applicants. In addition, new preset recommendation letters have been implemented in many universities. These involve an application form rather than an essay, which obliges mentors and staff to fill recommendatory forms without using AI language models. This system reduces the influence of the writing abilities and focuses mainly on the applicant’s skills and capacity.

CONCLUSION

Artificial intelligence language models have gained popularity because of their ability to rapidly generate content and respond to inquiries. Such technology allows academic staff to write very informative and well-structured recommendation letters that may have more superior writing than human-written letters. However, if program directors are aware of the capability of such a technology, the AI-written letters can generally be identified. Further studies will be required to improve the power of results and reach more robust conclusions; these should involve more medical schools and residency programs.

Affiliations: From the Division of Orthopaedic Surgery and Sports Medicine, McGill University Health Centre, Montréal, Que. (Mansour, Burman, Bernstein, Martineau); the Department of Orthopedic Surgery, Hôpital du Sacré-Cœur de Montréal, Canada; Université de Montréal, Montréal, Que. (Sandman); the Department of Orthopedic Surgery, Lebanese American University Medical Center – Rizk Hospital, Achrafieh, Beirut, Lebanon (Yammine); the Department of Orthopedic Surgery, Brown University, Providence, RI, USA (Daher).

Competing interests: Mark Burman is a representative on the Council of Delegates with the American Orthopaedic Society for Sports Medicine. Mitchell Bernstein reports consulting fees from Smith & Nephew, NuVasive, OrthoPediatrics, Synthes, Restor3D, Resolute Medical, and Orthofx, as well as stock in Restor3D and Resolute Medical. Emilie Sandman reports funding from Depuy, Zimmer, Stryker, Smith & Nephew, Wright Medical, Medacta, and Johnson & Johnson. She is a member of the executive committees of the Canadian Shoulder and Elbow Society and the Association orthopédique du Québec. No other competing interests were declared.

Contributors: Mark Burman and Paul Andre Martineau contributed to the conception and design of the work. Kassar Yammine contributed to data acquisition, and Jad Mansour, Mitchell Bernstein, Emilie Sandman, and Mohammad Daher contributed to data analysis and interpretation. Jad Mansour and Mohammad Daher drafted the manuscript. All of the authors revised it critically for important intellectual content, gave final approval of the version to be published, and agreed to be accountable for all aspects of the work.

Content licence: This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY-NC-ND 4.0) licence, which permits use, distribution and reproduction in any medium, provided that the original publication is properly cited, the use is noncommercial (i.e., research or educational use), and no modifications or adaptations are made. See: https://creativecommons.org/licenses/by-nc-nd/4.0/

References