Canadian Medical Education Journal

Brief Reports

Re-evaluating the role of personal statements in pediatric residency admissions in the era of artificial intelligence: comparing faculty ratings of human and AI-generated statements

Réévaluation du rôle des lettres de motivation dans l'admission en résidence de pédiatrie à l'ère de l'intelligence artificielle : comparaison des évaluations par le corps professoral des lettres rédigées par des humains et de celles générées par l'IA

Brittany Curry,[1] Amrit Kirpalani,[2] Mia Remington,[1] Tamara Van Hooren,[2] Ye Shen,[3] Erin R Peebles,[1,4]

[1]Department of Pediatrics, University of British Columbia, British Columbia, Canada; [2]Department of Pediatrics, Western University, Ontario, Canada; [3]BC Children's Hospital Research Institute, British Columbia, Canada; [4]Centre for Health Education Scholarship, University of British Columbia, British Columbia, Canada

Correspondence to: Brittany Curry, Department of Pediatrics, University of British Columbia, Vancouver, BC; phone: 604-875-2345; email: Brittany.curry@phsa.ca,

Published: Dec 22, 2025. CMEJ 2025, 16(6) Available at https://doi.org/10.36834/cmej.81345

Abstract

Background: Personal statements play a large role in pediatric residency applications, providing insights into candidates' motivations, experiences, and fit for the program. With large language models (LLMs) such as Chat Generative Pre-trained Transformer (ChatGPT), concerns have arisen regarding how this may influence the authenticity of statements in evaluating candidates. This study investigates the efficacy and perceived authenticity of LLM-generated personal statements compared to human-generated statements in residency applications.

Methods: We conducted a blinded study comparing 30 ChatGPT-generated personal statements with 30 human-written statements. Four pediatric faculty raters assessed each statement using a standardized 10-point rubric. We analyzed the data using linear mixed-effects models, a chi-square sensitivity analysis, an evaluation of rater accuracy in identifying statement origin as well as consistency of scores amongst raters using intraclass correlation coefficients (ICC).

Results: There was no significant difference in mean scores between AI and human-written statements. Raters could only identify the source of a letter (AI or human) with 59% accuracy. There was considerable disagreement in scores between raters as indicated by negative ICCs.

Conclusions: AI-generated statements were rated similarly to human-authored statements and were indistinguishable by reviewers, highlighting the sophistication of these LLM models and the challenge in detecting their use. Furthermore, scores varied substantially between reviewers. As AI becomes increasingly used in application processes, it is imperative to examine its implications in the overall evaluation of applicants.

Résumé

Contexte : Les lettres de motivation jouent un rôle crucial dans les candidatures aux résidences en pédiatrie, car elles permettent de mieux comprendre les motivations, l'expérience et l'adéquation des candidats au programme. L'utilisation de grands modèles de langage (GML), tels que ChatGPT (Chat Generative Pre-trained Transformer), a soulevé des interrogations quant à l'authenticité des lettres lors de l'évaluation des candidats. Cette étude examine l'efficacité et l'authenticité perçue des lettres de motivation générées par un GML par rapport à celles rédigées par des humains dans le cadre des candidatures aux résidences.

Méthodes : Nous avons mené une étude en aveugle comparant 30 lettres générées par ChatGPT à 30 lettres rédigées par des humains. Quatre membres du corps professoral en pédiatrie ont évalué chaque lettre à l'aide d'une grille d'évaluation standardisée sur 10 points. Les données ont été analysées à l'aide de modèles linéaires mixtes, d'une analyse de sensibilité par le test du χ^2, d'une évaluation de la précision des évaluateurs quant à l'identification de l'origine des lettres, ainsi que de la cohérence des scores entre évaluateurs à l'aide des coefficients de corrélation intraclasse (CCI).

Résultats : Aucune différence significative n'a été observée entre les scores moyens des lettres générées par l'IA et ceux des lettres rédigées par des humains. Les évaluateurs n'ont pu identifier la source d'une lettre (IA ou humain) qu'avec une précision de 59 %. Un désaccord considérable a été constaté entre les évaluateurs concernant les scores, comme l'indiquent les CCI négatifs.

Conclusions : Les lettres générées par l'IA ont été évaluées de manière similaire à celles rédigées par des humains et étaient indiscernables pour les évaluateurs, ce qui souligne la sophistication de ces modèles GML et la difficulté à détecter leur utilisation. De plus, les scores variaient considérablement d'un évaluateur à l'autre. À mesure que l'IA se généralise dans les processus de candidature, il est impératif d'examiner ses implications dans l'évaluation globale des candidats.

Introduction

Personal statements play a substantial role in candidate selection for pediatric residency programs and are a part of the Canadian residency matching application.[1] Applicants use these statements to describe their competencies, experience, and desire to study in their program of choice. These statements are reviewed during the initial screening and ranking phases of applications and may influence interview offers and the final selection of candidates. The subjective nature of scoring these statements is a longstanding concern: In 2022, Dirschl showed the intraclass correlation coefficients for individual elements within an orthopedic residency scoring system ranged from 0.28 to 0.98 with lowest scores for subjective elements like personal statements and letters of recommendation.[2]

The emergence of large language models (LLMs) has further complicated the issue by raising questions surrounding how LLMs are used to write or modify personal statements and how this may affect the role of these statements in evaluating candidates.[3,4] Discussions around the acceptability of using AI in residency applications have been sparse, particularly in a context where personal statement evaluation is already highly subjective and inconsistent.[2,5] Early research in this area has shown that some AI tools can distinguish between AI- and human-authored psychiatry personal statements,[6] while a small study in plastic surgery admissions found no significant differences in perceived quality between AI-generated and human-written personal statements as judged by faculty reviewers.[7] A small 2024 study in general surgery found that neither faculty nor residents could reliably distinguish between human versus AI written statements.[8] This is an important and emerging issue.

Despite the longstanding concerns regarding subjectivity, and newer concerns regarding AI use, personal statements remain widely used in residency selection. There is a lack of empirical data on the ability of evaluators to distinguish whether differences in quality are perceived between AI-generated and human-authored statements. This study addressed this knowledge gap by evaluating the use of AI-generated personal statements in pediatric residency applications with a standardized residency application scoring rubric used by selection committee members from two pediatric residency programs.

Following our exploration of how faculty perceive AI-generated statements relative to authentic personal statements, we offer insights that may guide evolving residency admissions requirements related to these statements.

Methods

We received written permission from the Canadian Residency Match Service (CaRMS) to extract key traits, themes, and experiences from 30 de-identified human-generated personal statements that were previously submitted to a Canadian pediatrics residency program and used these elements as input for ChatGPT (GPT-4)©. Using standardized prompts, we instructed the model to generate a pediatric residency personal statement incorporating the identified content, thereby creating 30 human-generated and 30 corresponding AI-generated statements. We coded the statements and randomly assigned them to four raters, representing residency selection committees from two different pediatric programs. We evaluated these statements using a standardized 10-point scoring tool already in routine use at one of the participating programs. The tool was developed locally for residency application file review and had not undergone external validation; however, its use in this study reflects its authentic application within the program. The scoring tool evaluates an applicant's motivation for the field of pediatrics and for the specific training program, as well as key strengths of the applicant. The primary objective was to compare scores between human- and AI-generated statements using linear mixed-effects models. We modeled score as the outcome, with human- and AI-generated as the fixed effect. Random intercepts were included for raters and statements to account for repeated measures. This particular method was used as each AI-generated personal statement was created based upon the contents of the corresponding human-generated personal statement. A sensitivity analysis was conducted using a chi-square test to account for inter-rater variability, comparing letter scores classified as High (>=rater's median score) and Low (<rater's median score). The secondary objective was to evaluate whether raters could correctly identify a letter as human- or AI-generated. In addition, we used intraclass correlation coefficients (ICC) using a two-way mixed effects model to compare the ratings of our reviewers on the same letters. We obtained ethics approval from the University of British Columbia (H23-01998).

Results

Each rater evaluated 30 statements (15 AI- and 15 human-generated) and each letter was evaluated by two raters.

Mean scores (out of 10) for statements were: Rater 1 (AI 7.00, Human 5.87), Rater 2 (AI 7.53, Human 6.47), Rater 3 (AI 4.93, Human 7.07), Rater 4 (AI 6.87, Human 7.47). We found no significant difference in scores between AI and human statements among correctly (coefficient 0.65, 95% CI -0.22 to 1.50, *p* = 0.14) and incorrectly (coefficient -0.69, 95% CI -1.70 to 0.36, *p* = 0.20) identified statements using linear mixed-effects models. The chi-square sensitivity analysis confirmed no significant difference in AI- or human-generated statement scores (*p* = 0.70 among correctly identified statements, *p* = 0.12 among incorrectly identified statements).

Raters could only identify the source of a letter (AI or human) with 59% (70/120) accuracy (Rater 1: 60% (18/30), Rater 2: 60% (18/30), Rater 3: 67% (20/30), Rater 4: 47% (14/30)).

There was substantial variability between raters when scoring the statements, with negative ICCs indicating poor agreement in evaluating how well each statement aligned with the rubric criteria. The ICC between Rater 1 and Rater 3 who rated the same statements was -.236, and between Rater 2 and Rater 4 was -0.133.

Discussion

Using a site-specific pediatric residency admissions scoring rubric, we found that ChatGPT©-generated statements did not score significantly differently from human-written statements. This finding aligns with existing literature demonstrating that AI-generated content meets or exceeds human standards across various professional writing contexts.[9,10] However, the interpretation of this result was constrained by substantial variability in reviewer ratings. There has been previous research describing this variability,[5] which substantiates concerns about the reliability of the personal statement in the selection process. These results suggest that programs may need to reassess their evaluation processes to enhance consistency, potentially through the development of empirically evaluated scoring instruments, more robust rater training to decrease inter-rater variability, and the reconsideration of the role of the personal statement in selection decisions.

Reviewers in our study were unable to reliably determine whether a statement was written by AI or a human. This degree of inaccuracy is comparable to findings from two other residency admission studies, where ChatGPT deceived reviewers in 34-56% of letters.[11,12] This underscores the sophistication of LLMs and raises further concerns about the authenticity of these submissions.

As financial and social capital are known to provide advantages to candidates in the residency matching process,[13] emerging AI tools may serve as an additional resource by supporting applicants who lack access to paid editing services, established mentorship networks, or fluency in academic English.[14] In this sense, AI has the potential to modestly lessen existing inequities in how some applicants prepare their materials but does not negate the need for broader system reform. There are important concerns regarding authenticity and the shifting role of the personal statement. The value, purpose, and interpretability of the personal statement have long varied across programs,[15,16] and evidence of its ability to meaningfully distinguish among similarly qualified applicants remains limited.[5] Given these longstanding challenges - and the additional complexity introduced by AI - some have suggested that reducing the weight of the personal statement is warranted.[17,18] Others propose that, rather than eliminating it outright, the field may benefit from clearer guidance regarding expectations and evaluation in the AI-era.[6,19]

One could consider shifting its focus from a narrative assessment of character to a tool for assessing specific professional attributes, such as empathy or communication skills, that are harder for AI to convincingly simulate,[20] and remove the poor human judgement that makes personal statements so unreliable.

Limitations of our study include a small sample size, and a scoring tool that was developed locally without external validation. Although its use allowed us to evaluate statements in the context of real-world residency file review, the lack of formal validation and absence of structured rater training likely contributed to variability in scoring.

Our findings provide evidence from a real-world pediatric residency selection context, underscoring the sophistication of AI-generated statements and the challenge in distinguishing them from human-authored submissions. Further research could explore how selection committees perceive the use of AI in application materials and whether programs are adapting their evaluation processes in response to its growing prevalence to support the modernization of residency selection in the era of AI. As well, further research involving additional programs across a range of disciplines and training contexts (including both academic and community settings) would help deepen understanding of how AI-generated materials intersect with selection practices and norms across the residency landscape. At the very least, our findings should prompt all residency program selection committees to ensure all reviewers are trained on the use of the scoring tool, and initiate or continue discussions around the utility of the personal statement in the AI era.

Conflicts of Interest: None
Funding: None
Authorship: All authors have each signed their own attestation statement that they meet the requirements of authors. The lead author, in addition to their own attestation as an author, has signed an attestation that all authors listed on this paper meet the requirements for authors.
Edited by: Cindy Schmidt (senior section editor); Marcel D'Eon (editor-in-chief)

References

1. Whalen A. *CaRMS*. 2024. Available from https://www.carms.ca/ [Accessed Oct 30, 2024].
2. Dirschl DR. MD. Scoring of orthopaedic residency applicants: Is a scoring system reliable? *Clin Orthop Relat Res.* 2002;399:260-264. https://doi.org/10.1097/00003086-200206000-00033
3. Hostetter L, Kelm D, Nelson D. Ethics of writing personal statements and letters of recommendations with large language models. *ATS Sch*. 2024;0038PS. https://doi.org/10.34197/ats-scholar.2024-0038PS
4. Zumsteg JM, Junn C. Will ChatGPT match to your program. *Am J Phys Med Rehabil*. 2023;102(6):545-547. https://doi.org/10.1097/PHM.0000000000002238
5. White BA, Sadoski M, Thomas S, Shabahang M. Is the evaluation of the personal statement a reliable component of the general surgery residency application? *J Surg Educ.* 2012;69(3):340-343. https://doi.org/10.1016/j.jsurg.2011.12.003
6. Burke H, Kazinka R, Gandhi R, et al. Artificial intelligence-generated writing in the ERAS personal statement: an emerging quandary for post-graduate medical education. *Acad Psychiatry.* 2025; 49:13-17. https://doi.org/10.1007/s40596-024-02080-9
7. Patel V, Deleonibus A, Wells MW, Bernard SL, Schwarz GS. Distinguishing authentic voices in the age of ChatGPT: comparing AI-generated and applicant-written personal statements for plastic surgery residency application. *Ann Plast Surg.* 2023;91(3):324-325. https://doi.org/10.1097/SAP.0000000000003653
8. Whitrock J, Pratt C, Carter M, et al. Does using artificial intelligence take the person out of personal statements? We can't tell. *Surg*. 2024;176(6):1610-1616. https://doi.org/10.1016/j.surg.2024.08.018
9. Johnstone RE, Neely G, Sizemore DC. Artificial intelligence software can generate residency application personal statements that program directors find acceptable and difficult to distinguish from applicant compositions. *J Clin Anesth.* 2023;89:111185. https://doi.org/10.1016/j.jclinane.2023.111185
10. Gao CA, Howard FM, Markov N.S., et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digit Med.* 2023;6(75). https://doi.org/10.1038/s41746-023-00819-6
11. Chen J, Tao BK, Park S, Bovill E. Can ChatGPT fool the match? Artificial intelligence personal statements for plastic surgery residency applications: a comparative study. *Plastic Surg*. 2024;33(2):348-353. https://doi.org/10.1177/22925503241264832
12. Lum ZC, Guntupalli L, Saiz AM, et al. Can artificial intelligence fool residency selection committees? Analysis of personal statements by real applicants and generative AI, a randomized, single-blind multicenter study. *JB JS Open Access.* 2024;9(4):e24.00028. https://doi.org/10.2106/JBJS.OA.24.00028
13. Christophers B, Marr MC, Pendergrast TR. Medical school admission policies disadvantage low-income applicants. *Perm J.* 2022;26(2):172-176. https://doi.org/10.7812/TPP/21.181
14. Shadan M, Chhapra HU, Mashooq FN. Navigating challenges: Supporting non-native speaking medical students with AI and mentorship. *Cogent Educ*. 2024;12(1). https://doi.org/10.1080/2331186X.2025.2563991
15. Taylor C, Weinstein L, Mayhew H. The process of resident selection: A view from the residency director's desk. *Obstet Gynecol.* 1995;85(2):299-303. https://doi.org/10.1016/0029-7844(94)00388-T
16. 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):346-351. https://doi.org/10.1016/j.jclinane.2009.10.007
17. Matsubara, S. Comment on "Artificial intelligence-generated writing in the ERAS personal statement: an emerging quandary for post-graduate medical education". *Acad Psych.* 2025;49,200-201. https://doi.org/10.1007/s40596-025-02123-9
18. Matsubara S, Matsubara D. Letter regarding: "Digital ink and surgical dreams: perceptions of artificial intelligence-generated essays in residency applications." *J Surg Res*. 2024;303:797-8. https://doi.org/10.1016/j.jss.2024.08.025
19. Subillaga O, Coulter AP, Tashjian D, Seymour N, Hubbs D. Artificial intelligence-assisted narratives: analysis of surgical residency personal statements. *J Surg Educ.* 2025;18:103566. https://doi.org/10.1016/j.jsurg.2025.103566
20. Montemayor C, Halpern J, Fairweather A. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. *AI Soc.* 2022;37(4):1353-1359. https://doi.org/10.1007/s00146-021-01230-z