Canadian Medical Education Journal

Original Research

Investigating the threat of AI to undergraduate medical school admissions: a study of its potential impact on the rating of applicant essays

Enquête sur la menace que représente l'IA pour les admissions dans les facultés de médecine : une étude de son impact potentiel sur l'évaluation des lettres de motivation des candidats

Joshua Choi,[1,2] *Jenny Zhao,*[1,2] *Thuy-Anh Ngo,*[3] *Lawrence Grierson*[1,4]

[1]Department of Family Medicine, Faculty of Health Sciences, McMaster University, Ontario, Canada; [2]Faculty of Health Sciences, McMaster University, Ontario, Canada; [3]Educational Developer, Office of Teaching and Learning, University of Guelph, Ontario, Canada; [4]McMaster Education Research, Innovation and Theory, Faculty of Health Sciences, McMaster University, Ontario, Canada

Correspondence to: Lawrence Grierson, PhD, Department of Family Medicine, McMaster University, 100 Main St. W., Hamilton, ON, L8P 1H6; phone: (905) 525-9140 x22738; email: griersle@mcmaster.ca

Published ahead of issue: Jan 6, 2025 CMEJ 2025 Available at https://doi.org/10.36834/cmej.79690

Abstract

Background: Medical school applications often require short written essays or personal statements, which are purportedly used to assess professional qualities related to the practice of medicine. With generative artificial intelligence (AI) tools capable of supplementing or replacing inputs by human applicants, concerns about how these tools impact written assessments are growing. This study explores how AI influences the ratings of essays used for medical school admissions

Methods: A within-subject experimental design was employed. Eight participants (academic clinicians, faculty researchers, medical students, and a community member) rated essays written by 24 undergraduate students and recent graduates from McMaster University. The students were divided into four groups: medical school aspirants with AI assistance (ASP-AI), aspirants without AI assistance (ASP), non-aspirants with AI assistance (NASP-AI), and essays generated solely by ChatGPT 3.5 (AI-ONLY). Participants were provided training in the application of single Likert scale tool before rating. Differences in ratings by writer group were determined via one-way between group ANOVA.

Results: Analyses revealed no statistically significant differences in ratings across the four writer groups (p = .358). The intraclass correlation coefficient was .147.

Conclusion: The proliferation of AI adds to prevailing questions about the value personal statements and essays have in supporting applicant selection. We speculate that these assessments hold less value than ever in providing authentic insight into applicant attributes. In this context, we suggest that medical schools move away from the use of essays in their admissions processes.

Résumé

Contexte : Les demandes d'admission dans les facultés de médecine exigent souvent de courtes lettres de motivation écrites ou des lettres de présentation, qui devraient être utilisées pour évaluer les qualités professionnelles liées à la pratique de la médecine. Les outils génératifs d'intelligence artificielle (IA) étant capables de compléter ou de remplacer les données fournies par les candidats humains, l'impact de ces outils sur les évaluations écrites suscite de plus en plus d'inquiétudes. Cette étude explore l'influence de l'IA sur l'évaluation des lettres de motivation utilisées pour les admissions dans les facultés de médecine.

Méthodes : Un plan expérimental à l'intérieur d'un sujet a été utilisé. Huit participants (cliniciens universitaires, chercheurs de la faculté, étudiants en médecine et un membre de la communauté) ont évalué des lettres de motivation rédigées par 24 étudiants de premier cycle et diplômés récents de l'Université McMaster. Les étudiants ont été répartis en quatre groupes : les aspirants à la faculté de médecine avec l'aide de l'IA (ASP-IA), les aspirants sans l'aide de l'IA (ASP), les non-aspirants avec l'aide de l'IA (NASP-IA), et les lettres de motivation générées uniquement par ChatGPT 3.5 (IA-UNIQUEMENT). Les participants ont reçu une formation pour l'application de l'échelle de Likert unique avant d'évaluer. Les différences d'évaluation selon le groupe de rédacteurs ont été déterminées au moyen d'une ANOVA à sens unique entre les groupes.

Résultats : Les analyses n'ont révélé aucune différence statistiquement significative dans les évaluations entre les quatre groupes de rédacteurs (p = 0,358). Le coefficient de corrélation intraclasse était de 0,147.

Conclusion : La prolifération de l'IA renforce les questions qui se posent présentement sur la valeur des lettres de présentation et des lettres de motivation dans la sélection des candidats. Nous supposons que ces évaluations ont moins de valeur que jamais pour ce qui est de fournir un aperçu authentique des attributs des candidats. Dans ce contexte, nous suggérons que les facultés de médecine abandonnent l'utilisation des lettres de motivation dans leurs processus d'admission.

Introduction

Artificial intelligence (AI) technology has been rapidly advancing in medicine and healthcare since the release of the large language model, ChatGPT 3.5, in November 2022.[1,2] Within medical education, a notable application that has garnered interest is the way applicants to medical school may use AI technology in support of their admissions materials. On the positive side, aspiring medical students may leverage the technology to reduce language barriers, or as a substitute for assistance from personal connections, mentors, or costly premedical consultant services.[3] However, given its ability to supplement or replace inputs traditionally produced by humans, there is mounting concern that the technology may confound the integrity and authenticity of current selection systems; particularly, as it pertains to application materials that are submitted in written form.[4]

Although there is considerable evidence (predating the proliferation of generative AI technology) that admissions essays and personal statements are not rated reliably by admissions committees,[5–7] even when assessors are trained on established rubrics,[8] they are still used widely as part of holistic admissions assessments at institutions across Canada.[9–13] Presumably, training institutions see value in the insight these writing samples offer into important applicant attributes—motivation, collaboration, integrity, and empathy, to name a few.[14] However, research on the impact of AI in academic writing should stand to reinvigorate concerns about the reliability, and ultimately the validity, of these essays as medical school selection tools.[15–22] Numerous studies highlight how AI generates believable high-quality texts that can improve upon or outperform human writing,[15–18] and that can replicate the type of empathy we desire in physicians.[19,23] Furthermore, AI-generated content is frequently undetectable to human raters.[17,21,22]

Accordingly, this study considered the way in which AI influences the ratings of medical school admissions essays designed to evaluate the personal or professional qualities of applicants. Through a rigorous experimental investigation of rating outcomes, we aimed to understand the ways in which different approaches to AI use influenced the appraisal of materials created by authentic and inauthentic applicants to medical school. This is an important endeavour as medical student selection has significant ramifications for both applicants and society. Training institutions that use essays in their admissions systems need to understand how AI technology may disrupt their utility in differentiating candidates for matriculation.

Methods

Participants

Eight participants (four women, four men (self-identified)) were recruited via a combination of purposive and convenience sampling to act as essay raters. The sampling was purposive insofar that we sought a collection of individuals with professional and/or community status that would be commiserate with the constitution of a typical undergraduate medical training program's admissions committee. In this regard, our participants included four academic clinicians, one faculty researcher, two medical school students, and one community member. All participants had affiliations or active academic relationships with the Undergraduate MD Program and/or Faculty of Health Sciences at McMaster University (Hamilton, Canada).

Writing samples

24 participants (17 women, seven men) at McMaster University were recruited via convenience and purposive sampling to act as essay writers. Participants included current undergraduate students, recent graduates from an undergraduate program, and newly admitted postgraduate students. The sampling was purposive insofar that we sought 16 essay writers who were aspirants to medical school (i.e., individuals who had recently or were actively curating applications for submission for consideration) and eight essay writers who were not medical school aspirants (i.e., individuals pursuing careers in other non-healthcare fields).

Upon recruitment, these individuals were asked to generate a 250-word essay in response to a prompt used during the 2023-24 admissions cycle at a Canadian medical school that collects writing samples as part of their selection process:

> *The Russian poet Marina Tsvetaeva said: "it isn't that you need time to think, you need time to feel." How does this statement connect with your future career in the field of medicine?*

Participants were instructed to write their essays with the goal of developing a written statement that they believed would be rated favourably by a medical school admissions committee. Notably, the essay writers were placed into groups differentiated by the use of AI-assistance in

producing the essay. The 16 medical school aspirants were randomly assigned to a group that used AI-assistance (ASP-AI; *n* = 8) or a group that was prohibited from using AI-assistance (ASP; *n* = 8). The eight non-aspirants were assigned to a group that used AI-assistance in writing their essay (NASP-AI; *n* = 8). AI-assistance was provided by ChatGPT 3.5. The use of AI-assistance involved integrating outputs from the generative AI tool with personally developed writing material. The writer could either start with the AI and refine the essay from there, use the AI to refine an essay that they had generated, or toggle between AI and personally developed outputs. The research team did not provide essay writers any direction, instruction, or training on using the generative AI tool.

The research team also used ChatGPT 3.5 to generate eight additional 250-word essays in response to the prompt. These essays were produced without any human input beyond that offered to the generative AI tool as the imperative for the writing. This created a set of essays associated with a fourth "AI-only" group (AI-ONLY). To ensure some contextual nuance in these AI-generated responses, the researchers selected eight essays at random from the other three groups and reviewed them for relevant personal characteristics. These characteristics were then yoked into the instructions provided to ChatGPT 3.5. For example, if one of the randomly selected essays indicated that the author was the child of immigrant parents, then a similar detail was provided to the generative AI tool. A representative input for the AI-ONLY essays read as follows:

> *You are a radiation therapy student. During your placement, you encountered a patient who was diagnosed with breast cancer. Through this experience, you witnessed the power of empathy in healing when offering moments of genuine connections and shared humanity. Please write a 250 word response to this prompt: "The Russian poet Marina Tsvetaeva said: "it isn't that you need time to think, you need time to feel." How does this statement connect with your future career in the field of medicine?". Do not directly quote the prompt.*

In total, 32 essays were generated by essay writers and the generative AI tool, with eight each associated with the ASP-AI, ASP, NASP-AI, and AI-ONLY groups.

Data collection procedures

Each participant (i.e., essay rater) was tasked with rating eight essays, which were assigned to them in a pseudorandom fashion that ensured that each essay received two independent assessments. Their ratings were provided via a single 7-point Likert scale that was used by McMaster University's (Hamilton, Canada) medical school admissions committee when autobiographical statements were still part of the institution's selection process (Figure 1).

Upon reading the essay, please use the following scale to indicate the suitability of this candidate for admission into medical school

unsuitable	clearly unsuitable	marginally less suitable	suitable	marginally more suitable	clearly more suitable	outstanding
1	2	3	4	5	6	7

Figure 1. Assessment tool used to rate essays.

Raters were unaware that there were different categories of essay writers or that AI-technology may have been used to develop the essays.

Before beginning their assessments, all essay raters attended an hour-long, researcher-led group training session that introduced the essay rating tool. During the training session, raters used the scale to independently assess three exemplar essays (i.e., not selected from the 32 essays described above) and discussed how they applied the scale. Co-authors JC and JZ facilitated the discussion. Through the session, participants agreed that essays were to be rated with consideration for the sentiment that the author conveyed, the relevance that previous experiences described by the author had for the medical profession, the personality conveyed by the author, and the coherence of the writing. Following the session, participants were assigned their essays and given a week to complete the rating. Essays were read and rated via a secure RedCap survey.[24,25]

Data Analyses

Interrater reliability was determined by calculating an intraclass correlation coefficient estimate and 95% confident intervals based on a mean-rating (k = 2), two-way mixed-effects model. A single score representing the mean of the ratings provided by each essay's two raters was generated for each essay. These scores were then appraised via a one-way analysis of variance (ANOVA) with group as the only factor (ASP, ASP-AI, NASP-AI, AI-ONLY). Alpha was set at $p < .05$. All analyses were performed using SPSS statistical package version 23 (SPSS Inc, Chicago, IL).

Ethics

This research was approved by the Hamilton Integrated Research Ethics Board (HIREB; #16937). All participants and essay writers provided informed consent prior to study initiation or the submission of written work as experimental stimuli, respectively, as per the guidelines set out by the HIREB and the Declaration of Helsinki (2018).

Results

The single measures ICC was .147 with a 95% confidence interval from -.207 to .468, F (31,31) = 1.346, p = .207.

The ANOVA revealed no statistically significant differences in ratings as a function of writer groups (p = .358; grand mean (± SD) = 4.47 ± 0.18). Mean (± SD) scores for the ASP, ASP-AI, NASP-AI, and AI-ONLY groups appear in Table 1.

Table 1. Mean ratings as a function of writing group

Writing Group	Mean Rating (SD)
ASP	4.13 (.41)
ASP-AI	4.63 (.23)
NASP-AI	4.19 (.41)
AI-ONLY	4.94 (.37)

Discussion

This study sought to determine the influence that the use of generative AI assistance had on ratings of short medical school admissions essays. Our analysis revealed that essays written by medical school applicants with or without AI assistance, non-aspirants with AI assistance, and exclusively by ChatGPT 3.5 were rated similarly by our participants. That no differences were noted in the assessments of essays written by aspirants who used and did not use the AI support suggests that the technology offers little additional advantage for those pursuing medical training. Furthermore, that no differences were noted between aspirants and non-aspirants who used the AI-support suggests that the technology offers augmentation to essay writing that obviates the assumed advantage conveyed by dedicated experience and authentic orientation to a career in medicine. Together, these findings lend credence to the idea that generative AI technology has the potential to disrupt the utility of the short essay or autobiographical statement as an effective admission tool. This is consistent with research from other fields that shows how generative AI can support the development of text-based responses typically associated with competent and empathetic human writers.[15–19; 23]

To understand this potential, however, requires some consideration for the objective of the short essay in selection. If assessment of a writing sample is intended to present to admissions committees some evidence of the candidate's aptitude or attitude to the medical profession, their ability to empathize or communicate, and/or their understanding of the work, their community, or society, then these results suggest that AI technology may help essay writers generate content despite any real relationship to these constructs. This is evidenced in the way essays of similar quality are generated by the AI exclusively as well as by AI-assisted individuals who have expended little energy or thought in curating a medical school admissions package. That individuals may use this technology to perform as well as legitimate aspirants even though they have no genuine interest is a major threat to the selection tool.

Notably, the hope for a personal essay to offer an authentic view on candidate values may be fundamentally misconceived, even before accounting for the proliferation of AI. Given essays are typically completed without supervision, in the applicant's own time, over the course of months, admissions committees cannot ensure that they were written by applicants alone. It is quite possible that they received external input from family members, mentors, or paid consultants. Moreover, it is well established that medical schools signal, explicitly and implicitly, what they are seeking from applicants. These signals clearly influence how applicants behave with recent studies showing that they are often inclined to forgo their authentic positions in order to write what they perceive admissions teams want to hear.[28,29] A phenomenon that can be further complicated when applicant perceptions of what medical schools value are inaccurate or incongruent with the intentions of the prevailing admission policies.[30,31] With this context, and given the potential disruption AI poses to authenticity, we recommend medical schools consider the removal of the essay as a selection tool in their admissions process.

Our study is not without limitations. First and foremost, it is a small study, investigating a small sample of raters, essays, and prompts. This limits the generalizability of the results and highlights the need for the work to be reproduced with a larger group of writers, raters, and essay directions. Furthermore, the inter-rater reliability presented here was not strong. This limitation, however, is not uncommon in admissions research focusing on writing samples.[6,7,32]

Conclusion

This study offers evidence that suggests that AI technology has the potential to disrupt the utility of current systems. Education researchers are encouraged to continue to develop evidence that offer clarity in this space. Moreover, medical educators are encouraged to consider deeply how new technology may upend candidate appraisal candidates and spur fundamental shifts how and what attributes we seek in medical students.

Conflicts of Interest: Lawrence Grierson is an editor for CMEJ. He adhered to the CMEJ policy for editor-as-authors. The authors declare that they have no other competing interests.
Funding: Not applicable.
Edited by: Christina St-Onge (senior section editor); Marcel D'Eon (editor-in-chief)
Acknowledgements: The authors would like to Janice Sancan, Dr. Karen Bailey, Aimun Shah, and Rod Parsa for their assistance and input on the study.

References

1. Lee J, Wu AS, Li D, Kulasegaram K (Mahan). Artificial Intelligence in undergraduate medical education: a scoping review. *Acad Med*. 2021;96:S62-70. https://doi.org/10.1097/ACM.0000000000004291
2. Darcy AM, Louie AK, Roberts LW. Machine learning and the profession of medicine. *JAMA*. 2016;315:551. https://doi.org/10.1001/jama.2015.18421
3. Hashimoto DA, Johnson KB. The use of Artificial Intelligence tools to prepare medical school applications. *Acad Med.* 2023;98:978-82. https://doi.org/10.1097/ACM.0000000000005309
4. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative Artificial Intelligence for medical education: potential impact and opportunity. *Acad Med*. 2023. https://doi.org/10.1097/ACM.0000000000005439
5. Salvatori P. Reliability and Validity of Admissions Tools Used to Select Students for the Health Professions. *Adv Health Sci Educ.* 2001;6:159-75. https://doi.org/10.1023/A:1011489618208
6. Kulatunga Moruzi C, Norman GR. Validity of admissions measures in predicting performance outcomes: the contribution of cognitive and non-cognitive dimensions. *Teach Learn Med*. 2002;14:34-42. https://doi.org/10.1207/S15328015TLM1401_9
7. Youdas JW, Hallman HO, Carey JR, Bogard CL, Garrett TR. Reliability and Validity of Judgments of Applicant Essays as a Predictor of Academic Success in an Entry-Level Physical Therapy Education Program. *J Phys Ther Educ*. 1992;6:15. https://doi.org/10.1097/00001416-199201000-00005
8. Peeters MJ, Schmude KA, Steinmiller CL. Inter-rater reliability and false confidence in precision: Using standard error of measurement within PharmD admissions essay rubric development. *Curr Pharm Teach Learn*. 2014;6:298-303. https://doi.org/10.1016/j.cptl.2013.11.014
9. MD Program University of Toronto. *MD Program University of Toronto non-academic requirements*. Available at: https://applymd.utoronto.ca/non-academic-requirements [Accessed on Nov 3, 2024].
10. Schulich Medicine & Dentistry Admissions. *Admission requirements*. Available at: https://www.schulich.uwo.ca/med_dent_admissions/medicine/admission_requirements.html [Accessed on Nov 3, 2024].
11. Toronto Metropolitan University (TMU) School of Medicine. *Admissions requirements*. Available at: https://www.torontomu.ca/school-of-medicine/programs/md/ [Accessed on Nov 3, 2024].
12. Dalhousie University Medical School Admissions. *Admission requirements.* Available at: https://medicine.dal.ca/departments/core-units/admissions/admissions.html [Accessed on Nov 3, 2024].
13. Queen's University School of Medicine. *Queen's-Lakeridge Health MD Family Medicine program eligibility & application process*. Available at: https://meds.queensu.ca/academics/mdprogram/queens-lakeridge-health-md-family-medicine-program/eligibility_application_process [Accessed on Nov 3, 2024].
14. Albanese MA, Snow MH, Skochelak SE, Huggett KN, Farrell PM. Assessing personal qualities in medical school admissions: *Acad Med*. 2003;78:313-21. https://doi.org/10.1097/00001888-200303000-00016
15. Marzuki, Widiati U, Rusdin D, Darwin, Indrawati I. The impact of AI writing tools on the content and organization of students' writing: EFL teachers' perspective. *Cogent Educ.* 2023;10:2236469. https://doi.org/10.1080/2331186X.2023.2236469
16. Yeadon W, Inyang O-O, Mizouri A, Peach A, Testrow CP. The death of the short-form physics essay in the coming AI revolution. *Phys Educ*. 2023;58:035027. https://doi.org/10.1088/1361-6552/acc5cf
17. 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. *Sci Comm Educ;* 2022. https://doi.org/10.1101/2022.12.23.521610
18. Herbold S, Hautli-Janisz A, Heuer U, Kikteva Z, Trautsch A. A large-scale comparison of human-written versus ChatGPT-generated essays. *Sci Rep*. 2023;13:18617. https://doi.org/10.1038/s41598-023-45644-9
19. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med.* 2023;183:589. https://doi.org/10.1001/jamainternmed.2023.1838
20. Waltzer T, Cox RL, Heyman GD. Testing the ability of teachers and students to differentiate between essays generated by ChatGPT and high school students. *Hum Behav Emerg Technol.* 2023;2023:1923981. https://doi.org/10.1155/2023/1923981
21. Yan D, Fauss M, Hao J, Cui W. Detection of AI-generated essays in writing assessments. *Psychol Test Assess Model.* 2023;65:125-44.

22. Fleckenstein J, Meyer J, Jansen T, Keller SD, Köller O, Möller J. Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Comput Educ Artif Intell*. 2024;6:100209. https://doi.org/10.1016/j.caeai.2024.100209
23. Chen D, Parsa R, Hope A, et al. Physician and artificial intelligence chatbot responses to cancer questions from social media. *JAMA Oncol*. 2024;10:956-60. https://doi.org/10.1001/jamaoncol.2024.0836
24. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)-a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42:377-81. https://doi.org/10.1016/j.jbi.2008.08.010
25. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform*. 2019;95:103208. https://doi.org/10.1016/j.jbi.2019.103208
26. Shaheen MY. Applications of Artificial Intelligence (AI) in healthcare: a review. *Sci Prepr.* 2021. https://doi.org/10.14293/S2199-1006.1.SOR-.PPVRY8K.v1
27. Dias PP, Jayasinghe LB, Waldmann D. Investigation of mycelium-miscanthus composites as building insulation material. *Results Mater*. 2021;10:100189. https://doi.org/10.1016/j.rinma.2021.100189
28. White J, Brownell K, Lemay J-F, Lockyer JM. "What do they want me to say?" The hidden curriculum at work in the medical school selection process: a qualitative study. *BMC Med* Educ. 2012;12:17. https://doi.org/10.1186/1472-6920-12-17
29. Lin KY, Anspach RR, Crawford B, Parnami S, Fuhrel-Forbis A, Vries RGD. What must I do to succeed?: narratives from the US premedical experience. *Soc Sci Med* 1982. 2014;119:98. https://doi.org/10.1016/j.socscimed.2014.08.017
30. Yang L, Chang I, Ritz S, Grierson L. Research experiences for Canadian aspiring physicians: a descriptive analysis of medical school admission policies concerning research involvement in Canada. *BMC Med Educ*. 2022;22:151. https://doi.org/10.1186/s12909-022-03207-y
31. Chang I, Yang L, Elma A, Ritz SA, Grierson L. A brief report of aspiring medical student perceptions and behaviours concerning research experiences for selection into Canadian medical schools. *Can Med Educ J.* 2023;14:77-81. https://doi.org/10.36834/cmej.76255
32. Brown B, Carpio B, Roberts J. The use of an autobiographical letter in the nursing admissions process: initial reliability and validity. *Can J Nurs Res Rev Can Rech En Sci Infirm*. 1991;23:9-20.