To Bot or Not to Bot? The Impact of Generative AI on ERAS Personal Statements

Rahul R Menon BS[1], Donald Solomon MD[1, 2], Mia Berenson BS, VB Kushnir JD, Yekaterina Shapiro MD[1, 2]
Cooper Medical School of Rowan University[1], Cooper University Hospital Department of Otolaryngology[2]

Introduction

Over the past few years, Artificial Intelligence (AI) has opened paths to innovation and improvement in healthcare. The further we step into the era of AI, the more important it has become to explore how generative AI might reshape more traditional processes, particularly in its use, and potential abuse, by medical student applicants in the Electronic Residency Application Service (ERAS).

Generative AI, like OpenAI's ChatGPT 4.0, excels at crafting human-like content based on patterns learned from extensive datasets[1]. When prompted, it can create new content from what it has learned from the datasets almost instantaneously[2]. Like generative AI, physicians grapple with extensive datasets in their everyday lives, but generative AI does not "understand" information as humans do. Because of this, AI often generates information that at first glance seems plausible but is factually incorrect and sometimes borderline nonsensical.

When medical students enter their fourth year of medical school, many will be submitting their applications in ERAS for residency positions. In the residency application process, reviewers consider several items like test scores, reference letters, extracurricular activities, research items, and the personal statement under a holistic approach[3]. The personal statement has been considered a crucial component in reviewing an application[4]. Now, with the introduction of generative AI, it is quite possible the technology has already been used by students who are composing and presenting their narratives[5,6].

We explore the use of AI in crafting personal statements for ENT residency applications, comparing between genuine and AI-written statements while considering AI's impact on the residency selection process. AI detection tools and the judgement of Attorneys against that of ENTs are explored for detection of AI-written work.

Objective

To evaluate the ability of generative artificial intelligence (AI), specifically OpenAI's ChatGPT 4.0, to produce convincing personal statements for otolaryngology residency applications and assess whether expert reviewers can distinguish AI-generated from human-written content.

Methods

Chat-GPT 4.0 was prompted to draft 5 personal statements for otolaryngology residency. Central to the prompting was the goal to sound as human as possible. These 5 statements were compiled into a survey with 5 de-identified applicant-written essays, and the statements were graded on a rubric for originality, readability, how convincing the essay is, and desire to extend an interview to the writer of the personal statement. Each query for a new AI-generated personal statement was executed in its own ChatGPT session rather than in sequence to avoid any bias. Human-written personal statements for Otolaryngology were collected with permission and de-identified from applicants who applied to the Otolaryngology Residency program at our institution in 2019.

The AI-generated essays and human-written statements were individually assessed by four Otolaryngologists from our institution, blinded to the source of each statement. Based on a rubric provided to reviewers, the statements were graded for originality, readability, how convincing the essay is, and desire to meet the writer of the personal statement on a subjective range (Likert scale from 1-5). Four Attorneys were also surveyed to comment and check if they could differentiate between the AI-generated and human-written statements. The study design included Attorneys due to an assumption that they are trained to read with critical thinking and precision. Finally, the personal statements were passed through an AI detection tool (Scribbr) to check its performance against our human reviewers.

Data were analyzed using descriptive statistics, paired t-tests for comparisons, and the intraclass correlation coefficient to calculate inter-rater reliability. Statistical significance was set at $P < 0.05$, with analyses performed in R (Version 4.4.2).

Results

Four Otolaryngologists and four Attorneys reviewed ten personal statements for a total of 80 evaluations. Between the human and AI-generated personal statements, the means are similar, with the highest variance being in how convincing the AI-generated personal statements were to the reviewer (Table 1). Ultimately however, there was no significant difference (all $P > 0.05$) in these criteria between the AI-generated and human-written personal statements (Figure 1, Table 2). The intraclass correlation coefficient was *0.66* for both AI and human written statements, indicating moderate inter-rater reliability.

When stratifying the data by the profession of the reviewer, this revealed no statistically significant difference between Attorneys and Otolaryngologists regarding the assessed personal statement criteria (Figure 2).

In determining whether a personal statement was the product of AI or not, Attorneys were able to correctly distinguish half of the statements while otolaryngologists were right 90% of the time. The AI detection tool detected an average of 93.6% (82-100%) AI-generated text in the personal statements that were the full product of AI, as well as an average of 4% (0-20%) of AI-generated text in the human-written personal statements from 2019 (Figure 3).

Table 1. Summary of Personal Statement Criterion Data
Comparison Between AI and Human Statements

Criterion	Type	Mean	Standard Deviation (SD)
Readability	AI	3.628571	0.2166536
Readability	Human	3.525000	0.3964125
Originality	AI	3.650000	0.1629801
Originality	Human	3.525000	0.3893103
Desire to Meet Writer	AI	3.400000	0.5108204
Desire to Meet Writer	Human	3.300000	0.5768990
Convincing Nature	AI	3.350000	0.7623975
Convincing Nature	Human	3.425000	0.3601215

Table 2. Paired T-Test Comparing AI vs. Human Criterion Scores
Comparison Across Criteria

Criterion	T-Statistic	P-Value
Readability	-0.32638	0.7450
Originality	0.54167	0.5896
Convincing Nature	-0.28390	0.7773
Desire to Meet Writer	0.36992	0.7125

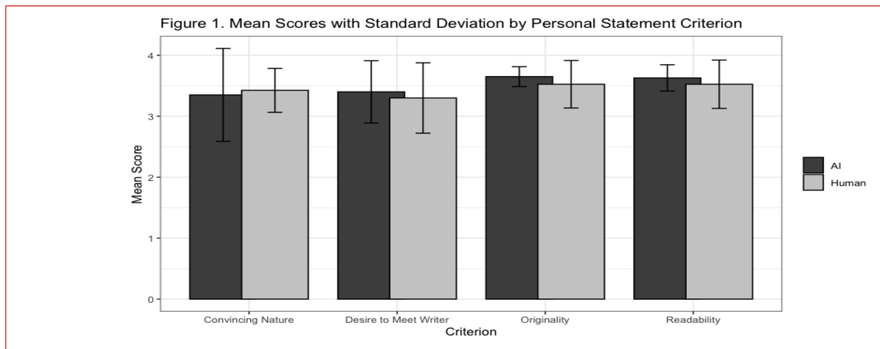

Figure 1. Mean Scores by Criterion for AI vs. Human Personal Statements Mean reviewer scores for each evaluation criterion. Error bars represent standard deviations, indicating variability in ratings

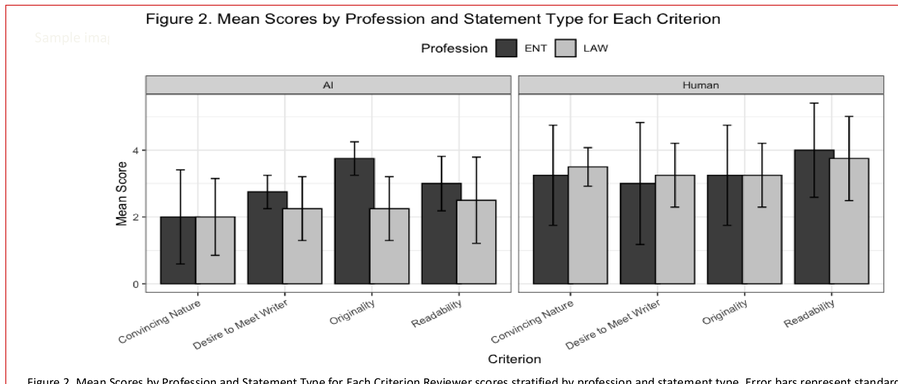

Figure 2. Mean Scores by Profession and Statement Type for Each Criterion Reviewer scores stratified by profession and statement type. Error bars represent standard deviations, reflecting rater variability across evaluation criteria.

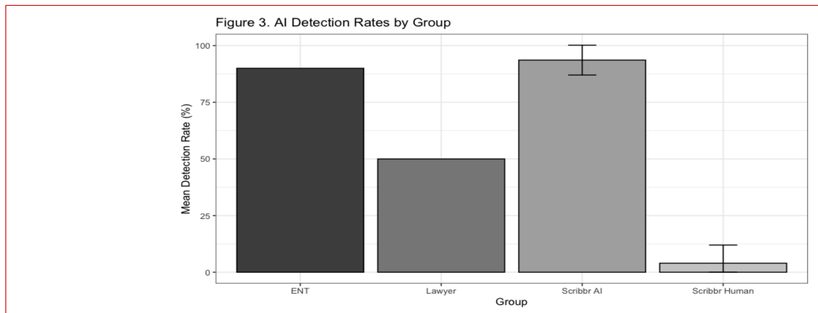

Figure 3. AI Detection Rates by Group Mean detection rates for identifying AI-generated or human-written statements by reviewer group. Error bars represent standard deviations.

Discussion

Generative AI can create high-quality personal statements that can be difficult to distinguish from human voices. While AI-generated personal statements, on paper, compare similarly to applicant-written statements in readability, originality, and persuasiveness, there seems to be something, either unwittingly communicated or lost in translation, in these bot essays that the astute reader can pick up on[7].

These results once again reiterate the fact that while generative AI seems to have amazing capabilities in creating text and synthesizing information, it lacks personality—which, among other things, seems essential to convey in a personal statemen

Despite this, one can make the case that ChatGPT, and other generative models like it, level the playing field. AI models could potentially be utilized because they could help those underrepresented in medicine, those who are first-generation in higher education, or those from lower socioeconomic statuses who may not have the time, connections, or resources in navigating the application process and sometimes assisting with editing or suggesting changes to their work where needed[1,8]. AI language models like ChatGPT and Google Bard are free and easily accessible tools that could in theory be utilized to enhance an application. They can evaluate personal statements for grammar and syntax and improve the clarity and conciseness of whatever the writer inputs. Aside from picking out instances of plagiarism or AI-generated content, they could even be used as a tool to objectively critique personal statements and help prevent potential biases that may arise with human reviewers, who, as we demonstrate, can be fallible. Moreover, in the human-written statements, which were compiled before the advent of generative AI, one statement still returned an AI-generated content percentage of 20%. For this reason, while these AI detection tools can be helpful, their outputs should be taken with a healthy dose of skepticism, at least for the time being.

Conclusion

Generative AI, like ChatGPT among others, could greatly impact the residency application process and complicate conventional approaches used to evaluate personal statements.

Further, the debate on responsibility by individuals for AI-generated content serves to highlight the need for clear guidelines and ethical standards in using such technologies[9]. In the field of medicine that is forging ahead with the introduction of advanced technologies, great caution is called for to ensure compatibility of AI with the guiding principles of equity and transparency, in addition to the overall objective of ensuring that the most qualified and most diverse candidates enter residency programs.

To bot or not to bot? We hope that for the medical students who have written or will write their personal statements soon that the answer is a resounding no. AI may be able to write with better syntax or grammar, but it can't write *YOU*.

How Do You Compare?

References: