Does AI Utilization Influence Residency Selection Committees? Analysis of Letters of Recommendation written by Real Mentors and Generative AI: A Randomized, Single-Blind, Multi-Center Study.

Samuel K. Simister[1], Eric Huish[2], Andrea Halim[3], Eugene Tsai[4], Hai Le[1], Dominick Tuason[3], John Meehan[1], Augustine Saiz[1], Holly Leshikar[1], Zachary Lum[1]

[1]*University of California, Davis, Sacramento, CA*
[2] *San Joaquin General Hospital, French Camp, CA*
[3]*Yale University, New Haven, CT*
[4]*Cedars Sinai, Los Angeles, CA*

PRESENTING AUTHOR: sksimister@ucdavis.edu

DISCLOSURES: All authors have no AI-related or otherwise relevant disclosures to include.

INTRODUCTION: The residency selection process is constantly evolving from the introduction of standard Letters of Recommendation (LOR) to preference signaling. LORs are considered an important aspect of applications. The potential capabilities of generative artificial intelligence (AI) tools have been relatively unexplored, specifically for authoring LOR for medical students applying to residencies. This study aimed to investigate the ability of generative AI to author LOR for residency applicants. We hypothesize that faculty on residency selection committees cannot determine differences between real and AI statements.

METHODS: Fifteen real LORs generated 15 unique and distinct personal statements from ChatGPT and BARD each, resulting in 45 statements. Statements were then randomized, blinded, and presented to faculty reviewers on residency selection committees. Seven reviewers assessed the statements by eleven metrics (following descriptors listed on the AOA's standardized LOR), specifying whether the personal statement was real or AI-generated. Descriptive statistics and the appropriate significance tests were calculated according to (1) actual and (2) perceived author in SPSS.

RESULTS SECTION: When evaluating LOR, faculty correctly identified authorship 48.3% of the time, with real, cGPT, and BARD percentages equaling 35.3%, 53.3%, and 56.7%, respectively ($p < 0.001$). The accuracy of identifying authorship did not increase over time (AUC 0.45, $p=0.102$). When comparing quality metrics by the actual author, there were no differences in means or significance. However, when comparing quality metrics by perceived author, all metrics were significantly higher for letters categorized as "real," including higher applicant ranking (7.32 vs. 5.08, $p < 0.001$) and desire for applicant (6.91 vs. 4.42, $p < 0.001$). See Table 1.

DISCUSSION: Faculty members were unsuccessful in determining the difference between human and AI-generated LOR more than half the time, with no significant difference between their scorings. However, letters that were perceived as real by evaluators had significantly higher quality metrics and ranking considerations for the respective applicant, showing a relationship between the quality of the letter and the perceived author.

SIGNIFICANCE/CLINICAL RELEVANCE: The results from our study suggest that generative AI can author LORs that perform similarly to real authors. With the increasing competitiveness of orthopaedic surgery residency, this highlights the importance of selection committees considering the role of LOR and their potential influence on residency applications.

Table 1: Quality metrics organized by author (Actual or Perceived)

	Actual			Perceived		
SLOR Quality Metrics*	**Real†**	**AI†**	**p**	**Real†**	**AI†**	**P**
Patient care	5.55 ± 2.6	5.78 ± 2.4	0.441	6.44 ± 2.4	5.25 ± 2.5	**<0.001**
Interpersonal skills	6.05 ± 2.3	6.04 ± 2.4	0.987	7.1 ± 1.9	5.39 ± 2.4	**<0.001**
Teamwork	6.06 ± 2.5	6.07 ± 2.4	0.96	7.3 ± 1.7	5.31 ± 2.5	**<0.001**
Procedural/Technical skills	5.33 ± 2.9	5.09 ± 3.1	0.502	6.28 ± 2.7	4.49 ± 3.0	**<0.001**
Adaptability	5.22 ± 2.9	5.4 ± 2.8	0.589	6.46 ± 2.3	4.65 ± 2.9	**<0.001**
Work Ethic	6.21 ± 2.3	6.15 ± 2.3	0.835	7.34 ± 1.6	5.45 ± 2.3	**<0.001**
Professionalism	6.1 ± 2.4	6.01 ± 2.5	0.769	7.29 ± 1.8	5.27 ± 2.5	**<0.001**
Research	2.38 ± 2.9	2.66 ± 3.1	0.452	3.11 ± 3.3	2.23 ± 2.8	**0.017**
Commitment to specialty	6.00 ± 2.4	6.03 ± 2.5	0.91	7.36 ± 1.6	5.2 ± 2.5	**<0.001**
Perceived Rank	5.96 ± 3.4	5.92 ± 2.4	0.882	7.32 ± 1.6	5.08 ± 2.4	**<0.001**
Desire to have in your program	5.35 ± 2.7	5.37 ± 2.7	0.951	6.91 ± 1.8	4.42 ± 2.5	**<0.001**

*SLOR, Standardized Letter of Recommendation
†Presented as mean ± standard deviation