

UCLA

UCLA Previously Published Works

Title

Caution Advised When Using Artificial Intelligence Chatbots to Review Psychiatry Residency Applications

Permalink

<https://escholarship.org/uc/item/34z7255z>

Journal

Academic Psychiatry

ISSN

1042-9670

Authors

Heldt, Jonathan

Yang, Yvonne

DeBonis, Katrina

Publication Date

2026

DOI

10.1007/s40596-025-02296-3

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, available at

<https://creativecommons.org/licenses/by-nc-nd/4.0/>

Peer reviewed



Caution Advised When Using Artificial Intelligence Chatbots to Review Psychiatry Residency Applications

Jonathan Heldt^{1,2} · Yvonne Yang^{1,2} · Katrina DeBonis^{1,2}

Received: 13 August 2025 / Accepted: 5 December 2025

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2025

To the Editor:

Artificial intelligence (AI) is increasingly being used in graduate medical education (GME) recruitment [1]. Given its potential to reduce administrative burdens, many programs are likely considering whether to use AI in future recruitment cycles.

There are few existing studies on AI in GME recruitment [2]. The UCLA Psychiatry Residency is conducting a study (IRB-25-0672) comparing humans against AI chatbots for reviewing residency applications. We felt compelled to share preliminary data that could inform decision-making.

Residents in our program ($n = 71$ across five classes from recruitment cycles 2020–2021 through 2024–2025) were invited to participate. Participation was voluntary, and risks of participation were explained. Our data security plan involved enterprise-grade security and did not use or retain uploaded data for public model training. Participants ($n = 22$, 31%) consented to their past applications being uploaded to two common AI chatbots: Google Gemini (GG) 2.5 Pro and OpenAI ChatGPT (OA) GPT-4-turbo.

Between July 25 and 28, 2025, we prompted the AI chatbots to score the applications according to our program's holistic review rubric [3]. This rubric includes both multi-point dimensional domains (Community Service, Leadership, Research, Clinical Performance, and Letters of Recommendation) and binary domains (Humanism/Professionalism, Resilience, Distance Traveled, Yellow Flags, and Red Flags). Each domain is defined algorithmically. For example, the "Exceptional" rating for Research is defined as "Commitment of at least 2 years full-time to research and exceptional scientific contribution demonstrated by ≥ 2 of

the following: PhD, first author peer-reviewed manuscript, superlative principal investigator letter of recommendation, etc." Domains also include concrete examples. For instance, Yellow Flags are defined as "issues that are not immediately disqualifying but may give pause about extending an interview, e.g., academic struggles that are explained by circumstances," while Red Flags are "concerns that should likely disqualify an applicant, e.g., descriptions of unprofessional behavior."

We attached the same rubric text used by human reviewers and prompted the AI chatbots with this text: "Using the rubric provided, provide a point-by-point analysis of the attached application. Each domain should have a discrete score based on the rubric." We did not train either AI on any existing data sets. We then compared the AI-generated ratings to human ratings of the same applications from prior recruitment cycles. Statistical analysis was conducted in IBM SPSS Statistics version 30.0 ($\alpha = 0.05$). Two-way repeated measures ANOVA was used for dimensional domains. A Generalized Linear Mixed Model (GLMM) with a binomial distribution and logit link function was used for binary domains, with reviewer and domain as fixed effects and application as a random effect.

Compared to humans, AI-processed applications were assigned significantly higher scores in dimensional domains ($F(2,42) = 18.086$, $p < 0.001$). Pairwise comparisons with a Bonferroni adjustment revealed that the overall mean score for humans was significantly lower than both GG ($p = 0.02$) and OA ($p < 0.001$). This difference remained statistically significant when looking at individual dimensional domains as well ($F(10,210) = 3.432$, $p = 0.008$). Presented descriptively, the mean scores for each dimensional domain were Community Service (humans = 3.5, GG = 3.8, OA = 3.8, max = 4), Leadership (humans = 3.5, GG = 3.9, OA = 4.0, max = 4), Research (humans = 4.8, GG = 5.5, OA = 5.5, max = 6), Clinical Performance (humans = 3.5, GG = 3.3, OA = 3.8, max = 4), and Letters of Recommendation (humans = 3.9, GG = 4.0, OA = 4.0, max = 4).

✉ Jonathan Heldt
jhheldt@mednet.ucla.edu

¹ University of California, Los Angeles, Los Angeles, CA, USA

² United States Department of Veterans Affairs, Los Angeles, CA, USA

A similar pattern emerged in binary domains. The GLMM revealed significant differences between humans and AI, with AI assigning positive scores more often ($F(2,315) = 74.368$, $p < 0.001$). This remained true when looking at individual domains ($F(8,315) = 189.858$, $p < 0.001$). For example, while humans found Resilience in around half of applications, it was identified by AI in nearly all applications (humans = 47%, GG = 100%, OA = 95%). This pattern was also seen for Distance Traveled (humans = 36%, GG = 73%, OA = 95%) and Humanism/Professionalism (humans = 55%, GG = 86%, OA = 100%). No significant differences were observed for Yellow Flags or Red Flags.

Some technical issues were also observed. Two-thirds of the applications had to be re-analyzed due to errors including rubric non-adherence (e.g., AI producing a rating of “Superior” when the rubric had only “Exceptional,” “Outstanding,” “Good,” or “Minimal/None”). In one application, GG and OA both failed to render the text of the MSPE and produced garbled output in its place. Rather than flagging this, both AI chatbots produced a low Clinical Performance rating that likely would have prevented this applicant from being interviewed if the error went undiscovered.

Even if these technical issues were fixed, we remain concerned by the trends we observed. Both AI chatbots had difficulty interpreting the real-world significance of applicant experiences. For example, our rubric explicitly defined Resilience as “overcoming longitudinal or structural disadvantages including discrimination, marginalization, illness, trauma, or disability.” Nevertheless, AI still coded “participating in sports” and “working as a medical assistant” as evidence of Resilience. While one could argue that AI being overly generous toward applicants is preferable to being punitive, artificially inflated scores still adversely impact the fairness of recruitment. If > 95% of applicants have Resilience, then this domain becomes meaningless for distinguishing applicants who truly have persevered through major disadvantages.

There are limitations to our findings. We only analyzed applications for residents in our program who, by virtue of having been offered an interview, represent the higher end of initial human scoring. Including the full spectrum of applications would enable more robust analyses such as determining relative rankings between applicants. In addition, our protocol was intended to test the use of two general purpose AI chatbots for GME recruitment. Accordingly, our findings may not be generalizable to GME-specific commercial AI platforms (e.g., Thalamus Cortex) or custom-designed machine-learning models optimized for this task [4]. Similarly, we used a standardized non-iterative prompt throughout our process; repeat cycles of prompt engineering could lead to higher concordance between human and AI scores [5]. Our preliminary findings should be viewed

as a pragmatic test of AI chatbots in real-world settings rather than an evaluation of AI’s full potential under ideal conditions.

In conclusion, we found evidence that two common AI chatbots struggle to accurately assess the complex factors considered in GME recruitment and may make critical errors that would unfairly disadvantage certain applicants. While the potential of AI for GME recruitment is vast, we presently advise caution when using AI chatbots to make decisions with potentially career-altering implications and advocate for continued human involvement in this process.

Author contributions J.H., Y.Y., and K.D. jointly conceived of and designed this study. J.H. and K.D. served as principal investigators and recruited participants. J.H. and Y.Y. performed data analysis. J.H. wrote the initial draft of the manuscript. All authors reviewed and approved the final manuscript.

Funding None (self-funded).

Data Availability A summary of data is provided within the manuscript. Original data and complete statistical analysis documents are available upon request.

Declarations

Ethics Approval Our study received IRB approval (IRB-25–0672) on 6/18/2025. Participants were provided with complete consent details at the time of enrollment. We took all appropriate measures to minimize risk to participants, including selecting generative AI tools that specifically are contracted with our institution and meet our privacy standards, following the guidelines of the Association of American Medical Colleges’ Protect Data Privacy policy that any generative AI tools used should have enterprise-grade security and privacy practices, using only institutional accounts with multi-factor authentication to access these tools, manually de-identifying any data that were uploaded to the generative AI tools by redacting any immediately identifiable data, storing all study data on institutionally affiliated services with multi-factor authentication, and deleting all data uploaded to the generative AI services at the conclusion of the study.

Competing interests The authors declare no competing interests.

References

1. Ajmani A, Klufas T, Zhou AE, Sloan B, Grant-Kels JM. Ethical implications of artificial intelligence in residency applicant screening. *J Am Acad Dermatol*. 2025. <https://doi.org/10.1016/j.jaad.2025.05.1416>.
2. Sumner MD, Howell TC, Soto AL, Kaplan S, Tracy ET, Zaas AK, et al. The use of artificial intelligence in residency application evaluation—a scoping review. *J Grad Med Educ*. 2025;17(3):308–19.
3. Barcelo NE, Shadravan S, Wells CR, Goodsmith N, Tarrant B, Shaddox T, et al. Reimagining merit and representation: promoting equity and reducing bias in GME through holistic review. *Acad Psychiatry*. 2021;45(1):34–42.

4. Drum B, Shi J, Peterson B, Lamb S, Hurdle JF, Gradick C. Using natural language processing and machine learning to identify internal medicine-pediatrics residency values in applications. *Acad Med.* 2023;98(11):1278–82.
5. Mesko B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J Med Internet Res.* 2023;4(25):e50638.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.