









**BRIEF CONTRIBUTION**

# ChatG-PD? Comparing large language model artificial intelligence and faculty rankings of the competitiveness of standardized letters of evaluation

Benjamin Schnapp MD, MEd<sup>1</sup>  | Morgan Sehdev MD<sup>2</sup>  | Caitlin Schrepel MD<sup>3</sup>  | Sharon Bord MD<sup>4</sup>  | Alexis Pelletier-Bui MD<sup>5</sup>  | Al'ai Alvarez MD<sup>6</sup>  | Nicole M. Dubosh MD<sup>7</sup>  | Yoon Soo Park PhD<sup>8</sup> | Eric Shappell MD, MHPE<sup>9</sup> 

<sup>1</sup>BerbeeWalsh Department of Emergency Medicine, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, USA

<sup>2</sup>Harvard-Affiliated Emergency Medicine Residency, Brigham and Women's Hospital/Massachusetts General Hospital, Boston, Massachusetts, USA

<sup>3</sup>Department of Emergency Medicine, University of Washington, Seattle, Washington, USA

<sup>4</sup>Department of Emergency Medicine, The Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

<sup>5</sup>Department of Emergency Medicine, Cooper Medical School of Rowan University, Camden, New Jersey, USA

<sup>6</sup>Department of Emergency Medicine, Stanford University, Palo Alto, California, USA

<sup>7</sup>Department of Emergency Medicine, Beth Israel Deaconess Medical Center/Harvard Medical School, Boston, Massachusetts, USA

<sup>8</sup>Department of Medical Education, University of Illinois at Chicago, Chicago, Illinois, USA

<sup>9</sup>Department of Emergency Medicine, Massachusetts General Hospital/Harvard Medical School, Boston, Massachusetts, USA

**Correspondence**

Benjamin Schnapp, BerbeeWalsh  
Department of Emergency Medicine,  
University of Wisconsin School of  
Medicine and Public Health, Madison, WI  
53705, USA.

Email: [bschnapp@medicine.wisc.edu](mailto:bschnapp@medicine.wisc.edu)

**Abstract**

**Background:** While faculty have previously been shown to have high levels of agreement about the competitiveness of emergency medicine (EM) standardized letters of evaluation (SLOEs), reviewing SLOEs remains a highly time-intensive process for faculty. Artificial intelligence large language models (LLMs) have shown promise for effectively analyzing large volumes of data across a variety of contexts, but their ability to interpret SLOEs is unknown.

**Objective:** The objective was to evaluate the ability of LLMs to rate EM SLOEs on competitiveness compared to faculty consensus and previously developed algorithms.

**Methods:** Fifty mock SLOE letters were drafted and analyzed seven times by a data-focused LLM with instructions to rank them based on desirability for residency. The LLM was also asked to use its own criteria to decide which characteristics are most important for residency and revise its ranking of the SLOEs. LLM-generated rank lists were compared with faculty consensus rankings.

**Results:** There was a high degree of correlation ( $r = 0.96$ ) between the rank list initially generated by LLM consensus and the rank list generated by trained faculty. The

Supervising Editor: Michael Gottlieb

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Author(s). *AEM Education and Training* published by Wiley Periodicals LLC on behalf of Society for Academic Emergency Medicine.

correlation between the revised list generated by the LLM and the faculty consensus was lower ( $r=0.86$ ).

**Conclusions:** The LLM generated rankings showed strong correlation with expert faculty consensus rankings with minimal input of faculty time and effort.

## INTRODUCTION

Emergency medicine (EM) continues to see a high volume of applicants to the specialty, with the total number of applicants in 2024 exceeding prepandemic numbers from 2019.<sup>1</sup> Reviewing residency applications continues to be extremely time-consuming, with one study estimating that each application requires between 10 and 30 minutes for review, adding up to hundreds of faculty hours for most programs.<sup>2</sup> There is continued interest among stakeholders in reforming the application process to allow more time for best practices such as holistic review,<sup>3</sup> but many proposed reforms, such as application caps, are felt to be unacceptable.<sup>4</sup> Significant challenges remain for programs attempting to accomplish a thoughtful holistic review of applicants,<sup>2</sup> and there exists a strong need for additional tools to help programs review applicants more efficiently and effectively.

Recently, artificial intelligence large language models (LLMs) like ChatGPT have begun to be used to solve complex data analysis problems in fields as diverse as software development and customer service.<sup>5</sup> ChatGPT has also been shown to be effective in the medical realm, achieving high levels of performance on the USMLE medical licensing examinations<sup>6</sup> as well as passing scores on specialty board examination tests.<sup>7</sup> LLMs have also been shown to have useful applications in medical research, including literature review and drug development.<sup>8</sup> Within residency education, natural language processing has been used to predict whether applicants were invited for an interview with moderate precision.<sup>9</sup>

Previous work has shown that faculty show high levels of consensus when evaluating the competitiveness of standardized letters of evaluation (SLOEs)<sup>10</sup> and that this consensus holds for the new competency-based SLOE 2.0.<sup>11</sup> An automated process for evaluating SLOEs offers the promise of greatly increasing the efficiency of residency applicant files; however, it is unknown whether LLMs are capable of analyzing SLOE data for competitiveness in the same way as human faculty. This study aimed to compare SLOE competitiveness rankings generated by an LLM to SLOE competitiveness rankings generated by the consensus of trained faculty.

## METHODS

### Content generation and faculty consensus ranking process

We utilized SLOE content developed for a previous study by a panel of expert faculty using previously described methods.<sup>11</sup> Faculty consensus rankings were generated by seven academic EM faculty

with significant experience with SLOEs, also described in a previous study.<sup>10</sup> These SLOEs were created to match the distribution of all SLOEs submitted nationally. No changes were made to the previously created SLOEs.

### LLM ranking process

A free, data-focused LLM frontend was utilized to accomplish the analysis: Julius (Julius.ai), which leverages ChatGPT 4o. To accomplish its analysis, Julius generated code in Python.

To generate a rank list using the LLM interface, an Excel spreadsheet containing the data from 50 mock SLOEs based on the new revised 2022–2023 EM SLOE template previously generated by faculty were uploaded into the Julius system for analysis. The LLM was opened to a fresh session and then given the prompts seen in Table 1.

As the initial response to this prompt often did not contain the entire list of 50 SLOEs or contained ties, it was then subsequently prompted to correct any errors. Analyses that were not completed due to errors or that generated output other than a full rank list were discarded. As LLMs often perform better when asked to iterate on an idea or topic and can sometimes take suboptimal shortcuts<sup>12</sup> it was then given the second prompt to generate a revised list with more in-depth analysis.

Because LLMs are stochastic and may generate different responses to the same queries executed multiple times based on weights,<sup>13</sup> the prompts above were presented to Julius seven different times (under a new thread each time to ensure that previous data was not being incorporated into the analysis) paralleling the seven faculty used to generate the initial consensus rankings; these rankings were then averaged to obtain the LLM consensus rankings.

### Data analysis

The mean of the seven LLM ratings for each SLOE was dubbed the “Initial LLM Ranking.” The mean of the seven LLM responses to the second prompt was deemed to be the “Revised LLM Ranking.” The mean faculty rating for each SLOE was termed the “Faculty Consensus Ranking.” Agreement was then calculated between the LLM consensus scores and the faculty rankings. Exact agreement meant the two consensus rankings were the same, tight meant the rankings were within two positions of each other, and close meant the rankings were within four positions, while loose meant the rankings were within six positions. Pearson's correlation coefficients

were also calculated using Excel. This study was deemed exempt by the Mass General Brigham Institutional Review Board.

## RESULTS

The ranking agreements for each level (exact, tight, close, and loose) between the LLM rankings and the faculty consensus rankings are shown in Table 2. Graphical representations of the initial and revised LLM rankings versus faculty consensus are depicted in Figure 1. Correlation between the initial LLM ranking and the faculty consensus ranking was 0.96 ( $p < 0.01$ ). Correlation between the revised LLM ranking and the faculty consensus ranking was 0.89 ( $p < 0.01$ ). A transcript of each of the seven LLM interactions is available as Appendix S1.

**TABLE 1** Prompts provided to the LLM on how to analyze the list of SLOEs.

### Initial LLM prompt:

*Each of the rows in this spreadsheet represents an individual applicant to Emergency Medicine residency. The columns generally represent how highly each applicant was rated in each of these categories. For the "rank list" category, 'Top 10' is the highest score, 'Top 1/3' is the next best, followed by 'Middle 1/3', 'Lower 1/3', then 'Unlikely'. For the "Guidance" category, "Minimal" is the best rating, followed by 'Standard', followed by 'Moderate', then 'Most'. For the categories scored by numbers, '5' is the highest possible score, '1' is the lowest score. For the 'Ability' categories, 'Fully' is most desirable, then 'Mostly', then 'Pre'. The narrative column contains descriptive information about each applicant. You are the Program Director of an Emergency Medicine residency. Using all of the information provided, including the columns about how highly the candidate will reside on your rank list, the column about how much guidance the applicant will need, and the narrative information, think about which applicants are most desirable for a spot in your Emergency Medicine Residency. Please reorder the entire list of applicants from most desirable to least desirable with no ties.*

### Revised LLM prompt:

*That analysis is superficial. Please be thoughtful in considering which characteristics are MOST important in selecting applicants for training to be an Emergency Medicine physician and reconsider the weights for each of the categories that make up the Composite Score.*

Abbreviations: LLM, large language model; SLOE, standardized letters of evaluation.

**TABLE 2** Ranking agreement for the LLM compared with faculty consensus rankings.

	Consensus: faculty rankings	Consensus: LLM initial rankings	Consensus: LLM revised rankings
Exact	22%	16%	18%
Tight	84%	56%	38%
Close	92%	76%	58%
Loose	97%	86%	70%
Correlation with consensus ratings	N/A	0.96	0.89

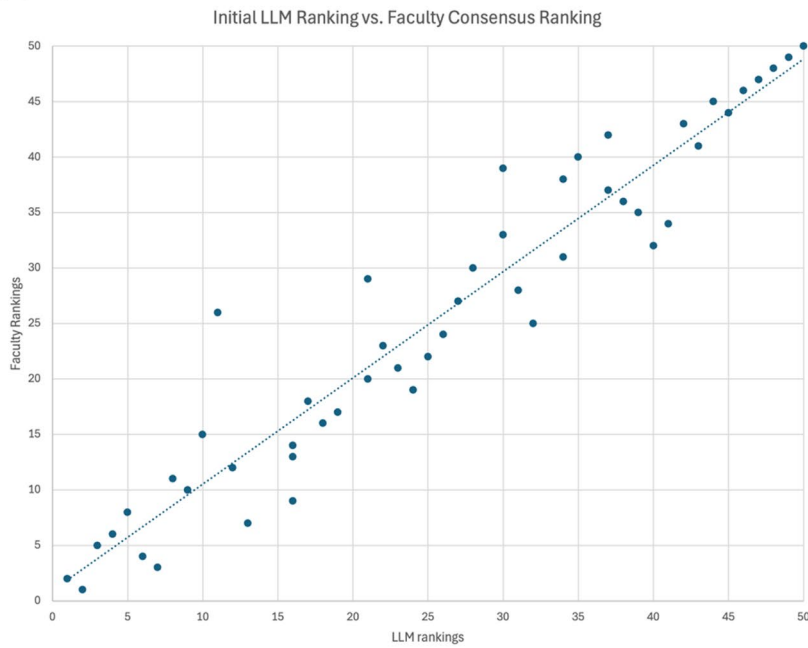
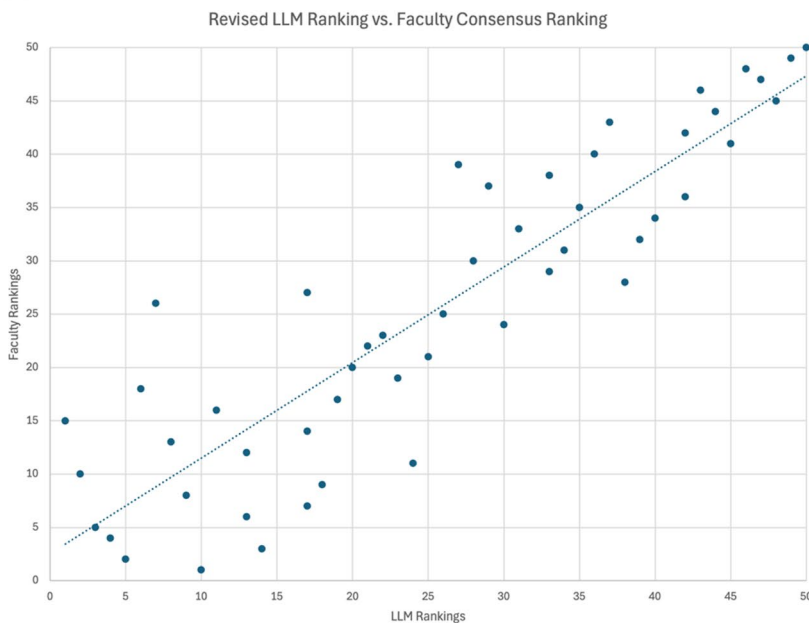
Abbreviation: LLM, large language model.

## DISCUSSION

An artificial intelligence LLM appears broadly effective at determining the competitiveness of mock SLOEs and their rankings appear similar to faculty consensus. This approach, especially if developed further, offers a potentially useful time-saving tool to program leaders tasked with analyzing large numbers of applicant letters, as the LLM was able to generate nearly identical rankings as trained faculty within seconds.

One interesting phenomenon noted in this study is that the model generated for scoring SLOEs mostly ignored the narrative data in favor of the rating scale data, despite explicit instructions to include it in the prompt (see Appendix S1 for full transcripts). It is unclear what may be driving this phenomenon. LLMs are clearly capable of parsing and analyzing narrative data. When asked explicitly to focus on the narrative data and create a system for scoring it, it can do so; however, it may be unclear on how to integrate and weight this sensibly with quantitative data, as suggested by the revised LLM consensus results, which were less reliable than the initial LLM consensus. There are also anecdotal reports that artificial limits are being imposed on the true capabilities of the LLM,<sup>12</sup> perhaps to limit the use of the computational power needed to generate responses. It is interesting to note, however, that when faced with a complex analysis task, humans may take similar shortcuts to what the LLM uses.<sup>14</sup> A qualitative study examining how faculty evaluate SLOE competitiveness similarly shows that the narrative is one of the least important factors evaluated.<sup>15</sup> However, the overall process faculty describe for evaluating SLOEs is complex and includes multiple factors, in contrast to the somewhat simplistic scoring systems generated by the LLM in response to the initial prompt. It is interesting to note that despite the complexity described by faculty in their rankings, their consensus rankings are overall similar to that derived from the LLM's relatively simple scoring.

Notably, the competitiveness rankings generated by the LLM changed markedly when it was asked to consider what factors are most important when selecting applicants. In contrast to program directors, who lean most on the global assessment question,<sup>16</sup> the LLM mostly ignored this, instead heavily prioritizing the SLOE's ratings of clinical skills or other factors, which resulted in a consensus list that was less consistent with faculty consensus. However, LLMs

**(A) Initial****FIGURE 1** Comparative competitiveness rankings.**(B) Revised**

offer the opportunity to substitute our own judgment instead: residency programs interested in prioritizing specific types of applicants (e.g., high receptivity to feedback) could easily ask it to reanalyze based on these criteria instead with more specific instructions included in the prompt. This approach, however, should be undertaken with caution. LLMs are known to be vulnerable to bias<sup>17,18</sup> and specifying increasingly fixed criteria may encourage programs to rank applicants with problematically similar “fit.”<sup>19</sup> Periodic bias audits of LLM analyses may be necessary to ensure applicants are not being systematically disadvantaged.

Currently, practical limitations may limit widespread use of LLMs for SLOE analysis and residency selection. It is necessary for SLOE content to be reformatted and placed in an organized database to allow LLMs to parse the data and perform their analysis; this is currently impractical to accomplish at scale for hundreds of real-life applicants. However, specialty-specific educational organizations like the Council of Emergency Medicine Residency Directors (CORD) could be useful for moving toward a future state where SLOE data are accessible in this way; centralization of critical resources for residency application has previously been proposed

as a useful reform for increasing efficiency.<sup>20</sup> Additionally, there may be questions around acceptability of LLM analysis, including whether residency program directors would feel comfortable outsourcing their SLOE review to an algorithm that currently devalues narrative data. Also, because SLOEs represent student data, the Family Educational Rights and Privacy Act (FERPA) dictates that they must be kept private;<sup>21</sup> commercial LLMs without data protection built in would almost certainly be unacceptable for real SLOEs.

## LIMITATIONS

This study is limited by the use of a single LLM to perform the analysis; it is possible that other LLMs would analyze the data differently or that the LLM would have improved performance with additional training. Additionally, it is possible that the prompts were suboptimal and that different wording would have been interpreted by the LLM differently to generate different results. It was also limited by the use of mock SLOEs instead of real SLOEs and by the use of a small group of faculty raters to generate the “gold standard” consensus; it is possible that different raters using real SLOEs would create different results.

## CONCLUSIONS

Large language models have the potential to generate useful rank lists of SLOE competitiveness that are highly correlated with expert faculty ratings of competitiveness with significantly less input of time and energy from faculty. Large language models performed worse when asked to use their own criteria to select residency applicants.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## ORCID

Benjamin Schnapp  <https://orcid.org/0000-0001-5031-8269>

Morgan Sehdev  <https://orcid.org/0000-0002-2981-0512>

Caitlin Schrepel  <https://orcid.org/0000-0003-0204-7657>

Sharon Bord  <https://orcid.org/0009-0008-5987-4139>

Alexis Pelletier-Bui  <https://orcid.org/0000-0002-9969-4123>

Al'ai Alvarez  <https://orcid.org/0000-0002-5438-2476>

Nicole M. Dubosh  <https://orcid.org/0000-0001-8674-5334>

Eric Shappell  <https://orcid.org/0000-0003-0281-3219>

## REFERENCES

- Cook TP. IMGs dramatically shift the 2024 match. *Emerg Med News*. 2024;46(2):1,21. doi:10.1097/01.EEM.0001006928.52936.e0
- Golden BP, Holland R, Zakowski L, Smith J. Using a consensus-driven approach to incorporate holistic review into an internal medicine residency program. *J Grad Med Educ*. 2023;15(4):469-474. doi:10.4300/JGME-D-22-00637.1
- Garrick JF, Perez B, Anaebere TC, Craine P, Lyons C, Lee T. The diversity snowball effect: the quest to increase diversity in emergency medicine: a case study of Highland's emergency medicine residency program. *Ann Emerg Med*. 2019;73(6):639-647. doi:10.1016/j.annemergmed.2019.01.039
- Dacre M, Branzetti J, Hopson LR, Regan L, Gisondi MA. Rejecting reforms, yet calling for change: a qualitative analysis of proposed reforms to the residency application process. *Acad Med*. 2023;98(2):219-227. doi:10.1097/ACM.0000000000005100
- Kalla D, Smith N, Samaah F, Kuraku S. Study and analysis of ChatGPT and its impact on different fields of study. *Int J Innov Sci Res Technol*. 2023;8(3):827-833.
- Knoedler L, Alfertshofer M, Knoedler S, et al. Pure wisdom or Potemkin Villages? A comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE Step 3 style questions: quantitative analysis. *JMIR Med Educ*. 2024;10:e51148. doi:10.2196/51148
- Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery*. 2023;93(6):1353-1365. doi:10.1227/neu.0000000000002632
- Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in medical research: current status and future directions. *J Multidiscip Healthc*. 2023;16:1513-1520. doi:10.2147/JMDH.S413470
- Mahtani AU, Reinstein I, Marin M, Burk-Rafel J. A new tool for holistic residency application review: using natural language processing of applicant experiences to predict interview invitation. *Acad Med*. 2023;98(9):1018-1021. doi:10.1097/ACM.0000000000005210
- Sehdev M, Schnapp B, Dubosh NM, et al. Measuring and predicting faculty consensus rankings of standardized letters of evaluation. *J Grad Med Educ*. 2024;16(1):51-58. doi:10.4300/JGME-D-22-00901.1
- Schnapp B, Sehdev M, Schrepel C, et al. Faculty consensus on competitiveness for the new competency-based emergency medicine standardized letter of evaluation. *AEM Educ Train*. 2024;8(5):e11024. doi:10.1002/aet2.11024
- Edwards B. As ChatGPT gets “lazy,” people test “winter break hypothesis” as the cause. *ARS Technica*. 2023 Accessed December 5, 2024. <https://arstechnica.com/information-technology/2023/12/is-chatgpt-becoming-lazier-because-its-december-people-run-tests-to-find-out/>
- Saba WS. *Stochastic LLMs Do Not Understand Language: Towards Symbolic, Explainable and Ontologically Based LLMs*. Springer; 2023:3-19. doi:10.48550/arXiv.2309.05918
- Dale S. Heuristics and biases: the science of decision-making. *Bus Inf Rev*. 2015;32(2):93-99. doi:10.1177/0266382115592536
- Schrepel C, Sehdev M, Dubosh NM, et al. Decoding competitiveness: exploring how emergency medicine faculty interpret standardized letters of evaluation. *AEM Educ Train*. 2024;8(4):e11019. doi:10.1002/aet2.11019
- Katzung KG, Ankel F, Clark M, et al. What do program directors look for in an applicant? *J Emerg Med*. 2019;56(5):e95-e101. doi:10.1016/j.jemermed.2019.01.010
- Navigli R, Conia S, Ross B. Biases in large language models: origins, inventory, and discussion. *J Data Inf Qual*. 2023;15(2):1-21. doi:10.1145/3597307
- Kotek H, Dockum R, Sun D. Gender bias and stereotypes in large language models. *Proceedings of the ACM Collective Intelligence Conference*. ACM; 2023:12-24. doi:10.1145/3582269.3615599
- Shappell E, Schnapp B. The F word: how “fit” threatens the validity of resident recruitment. *J Grad Med Educ*. 2019;11(6):635-636. doi:10.4300/JGME-D-19-00400.1
- Williams C, Kwan B, Pereira A, Moody E, Angus S, El-Bayoumi J. A call to improve conditions for conducting holistic review in

graduate medical education recruitment. *MedEdPublish*. 2019;8:76. doi:[10.15694/mep.2019.000076.1](https://doi.org/10.15694/mep.2019.000076.1)

21. Pfeifer CM. Privacy, trainee rights, and accountability in radiology education. *Acad Radiol*. 2017;24(6):717-720.

### SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Schnapp B, Sehdev M, Schrepel C, et al. ChatG-PD? Comparing large language model artificial intelligence and faculty rankings of the competitiveness of standardized letters of evaluation. *AEM Educ Train*. 2024;8:e11052. doi:[10.1002/aet2.11052](https://doi.org/10.1002/aet2.11052)