Original Research Paper

Development and evaluation of an artificial intelligence-based model for assessment improvement in Brazil's national online Continuing Medical Education program

Alisson Oliveira dos Santos[1], Tales Mota Machado[2], Josué de Lacerda Silva[3], João Paulo Valadares Vilaça[4], Henrique Pereira Alves[5], Moreno Magalhães de Souza Rodrigues[6], Leonardo Cançado Monteiro Savassi[7], Adelson Guaraci Jantsch[8], and Alysson Feliciano Lemos[9]

[1]MD, PhD, Professor, Universidade Federal de Mato Grosso do Sul, Campus Três Lagoas, Três Lagoas, Brazil
[2]MSc, IT Technician, Universidade Federal de Ouro Preto, Ouro Preto, Brazil
[3]BS, Technology Coordinator, Open University of SUS, Brasilia, Brazil
[4]BS, Systems Analyst, Open University of SUS, Brasília, Brazil
[5]BS, Data Analysis Specialist, Open University of SUS, Brasilia, Brazil
[6]PhD, Public Health Researcher, Center for Data and Knowledge Integration in Health, Fiocruz-Rondônia, Porto Velho, Brazil
[7]MD, PhD, Associate Professor, Federal University of Ouro Preto, Ouro Preto, Brazil
[8]MD, PhD, Researcher, Open University of SUS, Brasila, Brazil
[9]MSc, Coordinator of Program and Project Evaluation and Monitoring, Executive Secretariat, Open University of SUS, Brasilia, Brazil

Abstract

Background: The challenge of providing timely, high-quality feedback to thousands of healthcare professionals enrolled in distance-based Family Medicine specialization programs in Brazil creates an opportunity for the implementation of artificial intelligence. This study evaluates the effectiveness of Large Language Models (LLMs) in assisting tutors with student assessment in these programs. Methods: We implemented GPT-4o to analyze student responses to practical challenges in a Family Medicine distance education course. The system was structured through dataset preparation (518 student responses), prompt engineering, fine-tuning, and Retrieval-Augmented Generation. Evaluation included: human expert assessment using a 5-item Likert questionnaire (n=26 responses); metrics-based analysis comparing text length between LLM and tutor feedback (n=104); semantic similarity analysis between tutor- and LLM-generated texts (n=11); and comparison of scores assigned by tutors versus LLM (n=104). Results: Expert assessment showed high ratings for clarity (100% scoring "strongly agree") but lower scores regarding LLM's ability to replace tutors. LLM-generated feedback was significantly longer than tutors' (mean 190.11 vs. 109.69 words, p<.001). Semantic similarity between LLM and tutor responses was high (mean 85.92%). LLM-assigned scores differed slightly but significantly from tutor scores (mean 8.31 vs. 8.80, p<.001). Discussion: LLMs can generate clear, semantically aligned feedback and assign grades that approximate tutor scoring, offering a scalable enhancement to assessment in distance-based medical education. Nevertheless, they should be seen as a complement to human tutors rather than a replacement, especially where nuanced, contextualized guidance is required. Careful attention to regional language variation and domain-specific content will be essential for the safe, equitable integration of AI into continuing professional development. These findings are, however, limited by the use of data from a single Brazilian region and one course module, indicating the need for broader validation.

Keywords: Large Language Models, Continuing education, Familly Medicine, mentoring

Date submitted: 27-August-2025

Email: Alisson Oliveira dos Santos (alisson.o.santos@ufms.br)

Citation: Oliveira dos Santos A, Mota Machado T, de Lacerda Silva J, Paulo Valadares Vilaça J, Pereira Alves H, Magalhães de Souza Rodrigues M, Cançado Monteiro Savassi L, Guaraci Jantsch A, and Feliciano Lemos A. Development and evaluation of an artificial intelligence-based model for assessment improvement in Brazil's national online Continuing Medical Education program. Educ Health 2026;39:12-20

Online access: www.educationforhealthjournal.org
DOI: 10.62694/efh.2026.441

Published by The Network: Towards Unity for Health

BACKGROUND

Recent advances in Artificial Intelligence (AI), particularly in Natural Language Processing (NLP)[1] have introduced tools for process optimization across multiple sectors, including education and healthcare.[2] Among the most significant of these advances is the emergence of Large Language Models (LLMs)[2,] AI systems capable of understanding and generating language in a manner closely resembling human use.[3]

In educational settings, these models offer transformative potential by enabling personalized feedback,[4] adaptive content generation,[5] and scalable assessment support, especially in distance learning environments.[6]

The Open University of the Unified Health System (UNA-SUS) plays a key role in continuing education for Brazilian health professionals, offering specialization and training courses through digital platforms tailored to workforce needs.[7]

Two government programs were created to expand the supply of physicians in Brazilian Primary Health Care (PHC): the Doctors for Brazil Program (Programa Médicos pelo Brasil, PMB) and the More Doctors for Brazil Program (Programa Mais Médicos para o Brasil, PMMB). Under these programs, physicians receive a federal stipend to work in PHC and are also required to complete a distance-learning specialization course in Family Medicine. This course is developed in partnership with the UNA-SUS.

In 2024, the first year of PMMB's specialization course, more than 15,000 professionals enrolled, reinforcing the commitment to improving healthcare quality.[8] The PMMB Family Medicine specialization includes summative assessments in each of its 32 modules, forums, and a practical challenge, designed to promote application of knowledge to real-life scenarios.

In this model, tutors are responsible for monitoring activities and providing personalized feedback, a role that preferably requires a specialization in Family Medicine. However, the limited number of physicians with this title (only 11,255 in 2022[9]) creates a significant scalability problem for delivering prompt, rigorous evaluations to thousands of learners. LLMs can directly address this issue by providing such guidance at scale, while also enabling more efficient knowledge management.

This project aims to implement LLM-based AI tools to assist tutors in evaluating student performance in the PMB and PMMB Family Medicine specialization programs.

METHODS

Planning the Application

The "Practical Challenge" was the evaluative process chosen for implementing the language model application, as it represents a stage in the assessment process that involves text writing by the student professional, followed by feedback and the assignment of grades (from 0 to 10) by the tutor. For each module, practical challenges are proposed to the student professionals, requiring them to apply the knowledge acquired to real-life situations from their experience.

The challenges are implemented and follow the same logic in both specialization courses; however, since PMB was initiated earlier, it features a larger database of responses.

Selecting the LLM

There are several LLMs available in the community, including both open-source and commercial options. Ideally, one should test and compare all of them to select the one that best fits the specific problem.

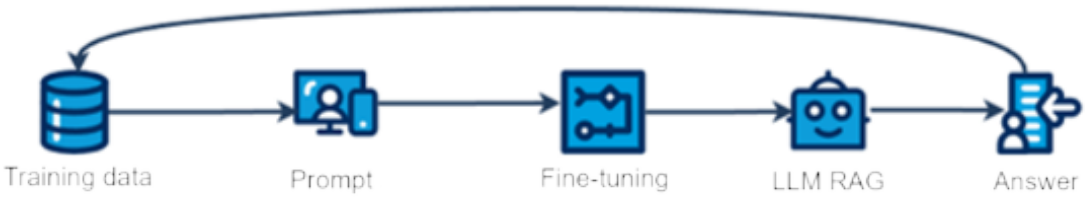
For this project, the selection was based on LiveBench, a ranking system that provides a robust and objective evaluation of language models, mitigating common challenges such as test set contamination and human bias.[10] Following this criterion, the LLM selected for the analyses was GPT-4o, which was among the reference models at the start of testing (August 2024). It is important to note that these rankings are dynamic, and model performance can change rapidly with new development cycles.

System structuring

Figure 1 illustrates the methodological flow employed in structuring the AI system. Adapting the model to the Family Medicine specialization required a multi-stage process involving model training (via dataset preparation and fine-tuning), prompt engineering, and context enrichment (via Retrieval-Augmented Generation).

Preparing the Training Dataset

Figure 1: Overall Structure of the Method to be Employed for System Implementation



At this stage of the project, a training dataset was compiled from a large, domain-specific corpus. It consisted of anonymized student submissions and corresponding tutor feedback from previous iterations of the PMB course.

The dataset was divided in an 80:20 ratio, meaning that 80% of the data was allocated for training the model and 20% for testing. It was randomized to ensure sample heterogeneity. Data quality was paramount for the model's training, necessitating rigorous data selection and preparation, which involved labeling the data to indicate what constituted a response, a grade, and feedback. This preparation was intended to guarantee the relevance and accuracy of the subsequent training.

Developing the Prompt
The creation of prompts involves developing specific instructions or commands that guide the LLM in generating responses or carrying out tasks.[11] These prompts were designed to reflect the learning objectives of the course modules and the assessment requirements. This step was instrumental in aligning the model's output with the project's expectations, ensuring that the responses generated were relevant and consistent with the evaluation criteria.

Fine-tuning
Fine-tuning is the process by which the pretrained LLM is adapted to the specific context of the project.[12] Drawing on the developed training set, the model was exposed to real-world examples from the specialization course, allowing it to learn the specialized language, patterns, and nuances of Family Medicine. This process involved configuring parameters and optimizing the model to enhance its performance in assessment and feedback tasks, ensuring that the generated responses were both accurate and contextually appropriate.

Retrieval-Augmented Generation
After the fine-tuning phase, a Retrieval-Augmented Generation (RAG) process was implemented to further enhance the model's ability to generate contextually rich and accurate responses. In this approach, the system integrates a retrieval component that searches for and obtains relevant documents from a reference corpus,[13] specifically, using the module's textbook as the sole source for retrieval. This retrieved document is then incorporated into the generation process, enabling the model to ground its outputs in material that is both vetted for accuracy and pertinent to the subject matter. This step ensures that the final responses are not only informed by the fine-tuning data but also enriched with the comprehensive context provided by the module's materials.

Evaluation of the Responses

In this phase, the language model was employed to evaluate the generated responses, comparing them with those provided by human tutors. This evaluation occurred in two stages: review by a second human evaluator, and a metrics-based assessment.

A second human evaluator (a family physician experienced in AI and educational tutoring) assessed the quality of the LLM-generated responses by comparing them to those produced by the original PMB tutor. This assessment was conducted using a five-item Likert-scale questionnaire covering clarity, relevance, adherence to established criteria, similarity to the tutor's feedback, and potential to replace the human response. Total scores ranged from 5 to 25, with higher values indicating better LLM performance. This questionnaire was applied to 5% of the total responses.

To explore the potential association between the questionnaire scores and the grades assigned by tutors, a correlation analysis was planned. The choice of coefficient (Pearson, Spearman, or Kendall) would be contingent on the data distribution, which was to be assessed using the Shapiro-Wilk test. If a statistical correlation was not feasible due to the data's distribution, a visual analysis using a scatter plot with a fitted regression line was planned as an alternative approach.

In parallel, a metrics-based evaluation was conducted on 20% of the dataset, consistent with literature recommending a minimum of 100 responses in LLM evaluation studies.[14] This included three components: (1) comparison of response lengths between tutors and the LLM; (2) semantic similarity analysis using BERT-based embeddings. Embeddings are numerical vectors that encode each sentence's meaning, produced by a language model (BERT). This approach enables comparison of the content of texts generated by tutors and by the LLM, capturing similarity in meaning beyond exact word overlap. Results are reported as percentages; higher values indicate greater textual similarity between the excerpts;[15] and (3) comparison of scores assigned by the tutors and those generated by the LLM.

Before applying inferential statistics, the normality of the paired differences was again assessed using the Shapiro-Wilk test. Given the violation of normality assumptions, the Wilcoxon signed-rank test was used to compare both response length and scores. Effect size (r) and 95% confidence intervals for means were also calculated to complement p-values. Given the limited number of comparisons, no adjustment for multiple testing was applied.

Statistical analyses were performed in Python (https://www.python.org/) using the SciPy stats module (https://docs.scipy.org/doc/scipy/tutorial/stats.html). Plots were generated using the Matplotlib (https://matplotlib.org/) and Seaborn (https://seaborn.pydata.org/) libraries.

Ethical Aspects and Data Protection

Ethical considerations and personal data protection for both student professionals and tutors were upheld throughout the system's development and implementation. All interactions complied with applicable privacy and data protection regulations through anonymization of sensitive information and implementation of access controls.

The project was approved by the Oswaldo Cruz Foundation (FIOCRUZ-DF) Research Ethics Committee under the number 30887420.7.0000.8027.

RESULTS

Application Structure

The application was developed and tested based on the practical challenge from Module 11 (Older Adult Health). The database consisted of 518 responses from student professionals in the Midwest region of Brazil, along with their corresponding tutor feedback.

Tagging and fine-tuning were performed on 414 (80%) of these responses and feedback entries. Prompts to the LLM were applied in accordance with best practices in prompt engineering.

Model Evaluation

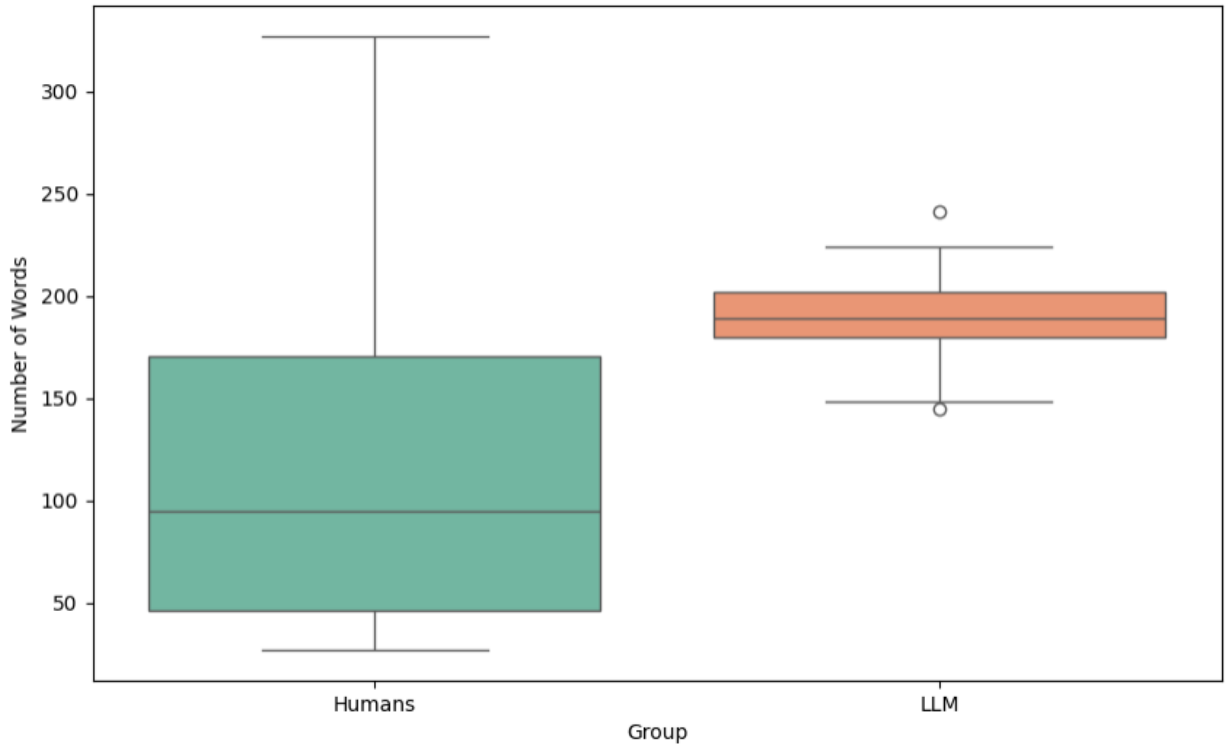
The analysis of LLM-generated responses by a second human evaluator covered 26 evaluations (5% of all responses). The scores ranged from 14 to 22 points, with a mean of 19.00 (SD = 1.98) and a median of 19.50. When examining individual questionnaire items, all responses to the first item ("Is the text generated by the LLM clear and properly written?") received the highest possible rating of 5 ("strongly agree"). In contrast, scores for the fifth item ("Can the LLM-generated text replace the tutor's response?") ranged from 1 to 3, indicating predominant disagreement or neutrality.

Text length was evaluated in 104 paired responses (20% of the dataset). The LLM-generated texts were significantly longer than those written by human tutors. The mean word count for tutor feedback was 109.69 (SD = 71.80; 95% CI: 95.73-123.66), whereas the LLM responses averaged 190.11 words (SD = 16.97; 95% CI: 186.81-193.41). A Wilcoxon signed-rank test confirmed this difference was statistically significant ($W = 430.0$, $z = -8.52$, $p < .001$, $r = 0.84$) (Figure 2).

Regarding the comparison of texts using semantic embeddings, 11 paired responses were evaluated. The similarity scores ranged from 79.10% to 93.54%, with a mean of 85.92% (SD = 4.94%; 95% CI: 82.60%-89.24%) and a median of 86.18%. These results indicate a relatively high degree of semantic similarity between LLM-generated and tutor-generated feedback.

The comparison of scores assigned by tutors and those generated by the LLM also involved 104

Figure 2: Box plot comparing the text length produced by human evaluators and tutors' feedback



paired responses. The tutor-assigned scores had a mean of 8.80 (SD = 1.33; 95% CI: 8.55-9.06), while the LLM-generated scores averaged 8.31 (SD = 1.32; 95% CI: 8.05-8.56). The Wilcoxon signed-rank test revealed a statistically significant difference between these scores (W = 104.0, z = -8.52, p < .001, r = 0.84) (Figure 3).

DISCUSSION

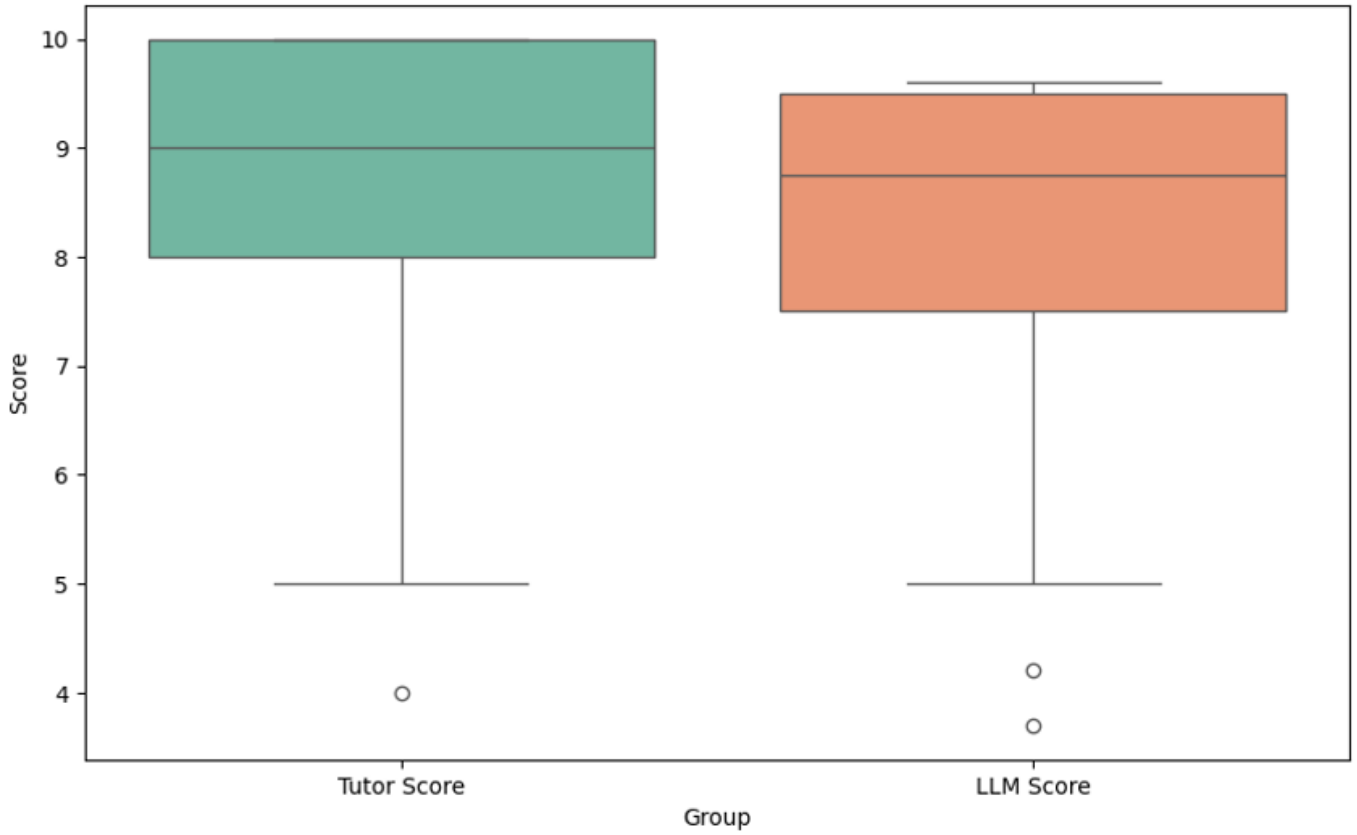
The results of this study indicate that implementing LLMs in the educational assessment process, specifically in the distance-based Family Medicine specialization program of PMB and PMMB, presents both potential and actual challenges. Although still limited in number, existing studies in the literature show promising outcomes regarding the use of AI for tutoring in distance education courses. Chen et al. (2024)[16] developed an LLM-based tutoring system with personalized learning support, capable of automatically adjusting course plans, providing informative instructions, and creating adaptive assessments. Meanwhile, Modran et al. (2024)[17] investigated the creation of an intelligent tutoring system featuring instant feedback, thereby offering a personalized and effective learning experience for students.

Structuring the application highlighted that even in automated tasks, human effort remains crucial. Although fine-tuning is essential for model training, it requires considerable time and resources, which can be costly when dealing with large datasets.[18] Furthermore, meeting the costs associated with maintaining and deploying LLMs in real time, consistent with the findings in the literature,[19] remains a challenge for their implementation.

In the context of the second human evaluator's analysis, based on the constructed questionnaire, the data indicate that LLM-generated text performs well in terms of clarity and organization. This aligns with studies suggesting that adequately trained LLMs can imitate the logic and fluency of human language, producing coherent and comprehensible text.[20]

Although statistical correlation could not be formally calculated due to the limited variability in the questionnaire scores, a visual inspection of the data suggested a positive relationship between the scores assigned by the tutor to the student professional, and the second human evaluator's assessments of the LLM outputs. This trend may indicate that in cases of lower student performance, the LLM-generated responses may not meet the

Figure 3: Box plot comparing the scores assigned by tutors and by the LLM



same expectations as those provided by human tutors. These situations require the tutor not only to correct the student's text but also to introduce new elements and more detailed explanations. This finding is consistent with the low scores on questionnaire item 5 ("Can the LLM-generated text replace the tutor's response?") and reinforces the literature consensus that LLMs should be seen as complementary to human work, with the latter being more capable of generating novel ideas.[21]

Another important finding was the statistically significant difference in the length of the texts generated by LLMs compared to those produced by human tutors. Although the LLM-generated texts were, on average, longer, this does not necessarily translate into higher quality or practical applicability. The lower standard deviation, relative to the variability observed in human-generated texts, indicates a tendency for the AI to be repetitive when dealing with the same task, even under different inputs.[22] This verbosity has practical implications. While clarity was rated highly, excessively long feedback may reduce tutor efficiency and student engagement. It is plausible that the prompt design, which explicitly requested "detailed and constructive feedback" and "practical suggestions", contributed to this length. In contrast, the greater variability found in the tutors' texts indicates a tendency toward feedback more adapted to the student professional's context and challenges.

Despite the average difference in text length between the LLM and tutors, the semantic similarity results are promising. It can be inferred that, despite some variability, most responses demonstrate satisfactory alignment.

Finally, it is worth highlighting the comparison between the grades assigned by tutors and those generated by the LLM. Although the absolute difference in mean scores was small, the difference was statistically significant. This finding suggests that even for tasks like grading, the model deviates from human standards. This differs from what is found in the literature, which indicates that AI performs better on more objective tasks,[23] and may indicate the need for model refinement.

Limitations and future directions

The study has several limitations that should be noted. The fact that the analyses were conducted using responses from only one region of Brazil introduces a significant source of bias. By not including responses from other regions, the model may struggle with regional language variations, specific clinical situations, and cultural adaptations. To mitigate this limitation, new training data should be incorporated as the application is deployed in production.

Beyond regional bias, broader ethical implications of integrating LLMs into assessment must be considered. These include the risk of algorithmic bias propagation, wherein the models reproduce or amplify systemic inequities present in their training data,[24] the potential for over-standardization of feedback, which risks homogenizing evaluative dialogue and stifling student creativity;[25] and the deskilling, the reduction in educators' judgment skills, as assessment tasks are increasingly outsourced to technology.[26]

Another relevant issue is that the analyses were carried out using only a single module (Older Adult Health). For a more comprehensive assessment, it is necessary to understand the LLM's performance across different areas of the course.

Furthermore, the sample sizes for specific analyses pose significant limitations. The human evaluator assessment (n=26) limits the reliability of the questionnaire findings. More critically, the semantic similarity analysis was conducted on a small subset (n=11). While the high mean similarity is promising, this finding must be interpreted with caution. These results should be considered preliminary and require validation in a larger, more representative sample.

Looking ahead, it will be important to analyze how the application behaves in a production environment. From there, conducting an evaluation that includes tutors' viewpoints and perceptions of its impact will be fundamental. In addition, a pivotal next step will be to collect feedback from the student professionals on the LLM-generated responses and to examine their effects on course performance and practical experiences. Future research should also explore the use of different LLMs, including open-source models, to compare performance, cost, and controllability.

Conclusions

This study underscores the potential of using LLMs in the assessment process for distance-based specialization courses in Family Medicine. LLMs have demonstrated the ability to provide clear and relevant feedback, as well as assign grades that are comparable to those of human tutors, although not statistically significant. Nonetheless, the findings also suggest that LLMs still cannot fully replace the role of tutors, particularly in situations requiring contextual or personalized judgment. This highlights the importance of a complementary approach, wherein AI supports rather than replaces human evaluation.

Moreover, the challenges encountered during the application's implementation, such as repetition, limited contextual adaptation, and training on a small dataset, indicate that improvements are necessary before this technology can be deployed on a large scale. In the future, it will be essential to expand the dataset to include other regions of the country and to conduct tests in additional course modules. Doing so is expected to enhance the efficiency and quality of the educational process without undermining the human-centered interaction and guidance that are essential in training healthcare professionals.

References

1. Arfi S, Srivastava N, Sharma N. Artificial Intelligence: An Emerging Intellectual Sword for Battling Carcinomas. *Current Pharmaceutical Biotechnology*. 2023;24(14):1784-1794. doi:10.2174/1389201024666230411091057

2. Lopes AA. Artificial intelligence in the education of health professions: a descriptive analysis through bibliometrics. *EDULEARN23 Proc*. Published online 2023:4000-4005. https://doi.org/10.21125/edulearn.2023.1077

3. Alsawas M, Alahdab F, Asi N, Li DC, Wang Z, Murad MH. Natural language processing: use in EBM and a guide for appraisal. *Evidence Based Medicine*. 2016;21(4):136-138. https://doi.org/10.1136/ebmed-2016-110437

4. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Published online August 1, 2023. https://doi.org/10.48550/arXiv.1706.03762

5. Suri G, Slater LR, Ziaee A, Nguyen M. Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *Journal of Experimental Psychology: General*. Published online February 8, 2024. https://doi.org/10.1037/xge0001547

6. Phung T, Cambronero J, Gulwani S, et al. Generating high-precision feedback for programming syntax errors using Large Language Models. Published online April 28, 2023. https://doi.org/10.48550/arXiv.2302.04662

7. Leiker D, Finnigan S, Gyllen AR, Cukurova M. Prototyping the use of Large Language Models (LLMs) for adult learning content creation at scale. Published online June 2, 2023. https://doi.org/10.48550/arXiv.2306.01815

8. Tu X, Zou J, Su WJ, Zhang L. What should data science education do with Large Language Models? Published online July 7, 2023. https://doi.org/10.48550/arXiv.2307.02792

9. Savassi LCM, dosSantos AO, Gasque KCS, et al. Continuing online education to health workforce: elderly's health care training experience. *European Journal of Public Health*. 2020;30(Supplement_5): ckaa165.163. https://doi.org/10.1093/eurpub/ckaa165.163

10. UNA-SUS. Começa o Curso de Especialização em Medicina de Família e Comunidade do Programa Mais Médicos. March 26, 2024. Accessed March 27, 2024. https://www.unasus.gov.br/noticia/comeca-o-curso-de-especializacao-em-medicina-de-familia-e-comunidade-do-programa-mais-medicos

11. Scheffer M. *Demografia Médica no Brasil 2023*. 6th ed. FMUSP, AMB; 2023. Accessed March 27, 2024. https://amb.org.br/wp-content/uploads/2023/02/DemografiaMedica2023_8fev-1.pdf

12. White C, Dooley S, Roberts M, et al. LiveBench: A challenging, contamination-free LLM Benchmark. Published online June 27, 2024. https://doi.org/10.48550/arXiv.2406.19314

13. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*. 2023;55(9):195:1-195:35. https://doi.org/10.1145/3560815

14. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. Published online May 24, 2019. https://doi.org/10.48550/arXiv.1810.04805

15. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Published online April 12, 2021. https://doi.org/10.48550/arXiv.2005.11401

16. Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review. *Npj Digital Medicine*. 2024;7(1):1-20. https://doi.org/10.1038/s41746-024-01258-7

17. Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. Published online August 27, 2019. https://doi.org/10.48550/arXiv.1908.10084

18. Chen Y, Ding N, Zheng HT, Liu Z, Sun M, Zhou B. Empowering private tutoring by chaining large language models. Published online August 4, 2024. https://doi.org/10.48550/arXiv.2309.08112

19. Modran PL, Bogdan IC, Ursuțiu D, Samoila C, Modran PL. LLM Intelligent agent tutoring in higher education courses using a RAG Approach. Published online July 5, 2024. https://doi.org/10.20944/preprints202407.0519.v1

20. Parthasarathy VB, Zafar A, Khan A, Shahid A. The ultimate guide to fine-tuning LLMs from basics to breakthroughs: an exhaustive review of technologies, research, best practices, applied research challenges and opportunities. Published online August 23, 2024. https://doi.org/10.48550/arXiv.2408.13296

21. Doshi AR, Hauser OP. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*. 10(28):eadn5290. https://doi.org/10.1126/sciadv.adn5290

22. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. Published online July 22, 2020. https://doi.org/10.48550/arXiv.2005.14165

23. Bano M, Zowghi D, Whittle J. Exploring qualitative research using LLMs. Published online June 23, 2023. https://doi.org/10.48550/arXiv.2306.13298

24. de Wynter A, Wang X, Sokolov A, Gu Q, Chen SQ. An evaluation on large language model outputs: Discourse and memorization. *Natural Languages Processing Journal*. 2023;4:100024. https://doi.org/10.1016/j.nlp.2023.100024

25. Lee Y, Kim S, Rossi RA, Yu T, Chen X. Learning to reduce: towards improving performance of large language models on structured data. Published online July 3, 2024. https://doi.org/10.48550/arXiv.2407.02750