CED/NOF-IADR
2024 Oral Health
Research Congress
12 — 14 Sept 2024
Geneva, Switzerland

0129

Assessing Language Models in Dental Education for Accuracy and Consistency

R. Richert[1], C. Lafourcade[2], B. Ballester[3], O. Kérourédan[2]

[1]Lyon Dental Hospital, Lyon, France, [2]UFR des Sciences Odontologiques, Université de Bordeaux, Bordeaux, France, [3]Assistance Publique Des Hôpitaux de Marseille, Marseille, France

Objectives The aim of this study was to evaluate and compare the performance of various language models in the context of dental education.

Methods Four language models were evaluated: ChatGPT 3.5 (OpenAI), ChatGPT 4 (OpenAI), Claude 3 (Anthropic) and the open model Mistral AI. The procedure sheets of the CNEOC (Collège National des Enseignants en Odontologie Conservatrice) were used to design a first set of 428 multiple choice questions (MCQ). The questions were answered and classified by two experts depending on their category (emergencies, restorative procedures, disinfection, etc.). Accuracy was assessed by comparing the differences between the expert's responses and the response given by each language model. Consistency was assessed using two metrics: robustness (the ability to provide identical responses to paraphrased questions). The context understanding was also evaluated based on the model's response to the appropriate category for each question. Finally, the lexicon of endodontic terms of the AAE (Association of American Endodontists) was used to create a secondary set of 539 questions and the accuracy was assessed as the ability to predict the correct term when given its definition.

Results The more advanced models (Claude 3, ChatGPT4) demonstrated significantly higher accuracy in answering the MCQs compared to the simpler models (ChatGPT3.5, Mistral AI), but the robustness was low for all models. All performances for defining terms and for context understanding were high, except for Mistral AI.

Conclusions While advanced language models demonstrate high accuracy and potential for dental education, their limited robustness necessitates caution in educational use. To improve performance, future studies should explore the integration of additional resources.