# Assessment of Chatbots's knowledge in Special Needs Dentistry

**Mouada Fadul, DMD; Rata Rokhshad, DDS; Kimberly Carr, DMD; Janice Jackson, DMD; Guihua Zhai, MS; Ping Zhang, DDS, PhD**

Department of Pediatric Dentistry, University of Alabama at Birmingham

Center for Clinical and Translational Science, University of Alabama at Birmingham

## ABSTRACT

**Purpose:** Chatbots, powered by big language models, use artificial intelligence to mimic real human chats; however, their ability for decision-making in dentistry hasn't been investigated much. This study aimed to evaluate the accuracy and consistency of chatbots in answering questions related to special needs dentistry.

**Methods:** Nine publicly accessible chatbots, including Google Bard, ChatGPT 4, ChatGPT 3.5, Llama, Sage, Claude 2 100k, Claude-instant, Claude-instant-100k, and Google Palm, were evaluated on their ability to answer a set of 25 true/false questions related to special needs dentistry and 15 questions for syndrome diagnosis based on their oral manifestations. Each chatbot was asked independently 3 times at 3-week intervals from November to December 2023 and the responses were evaluated by pediatric dentistry residents. Using SAS software, the comparison of accuracy rates among the chatbots was conducted employing the exact Wilcoxon test, a well-established statistical procedure renowned for its robustness in scientific analyses. Cronbach's alpha was utilized to measure the consistency of the chatbots' responses.

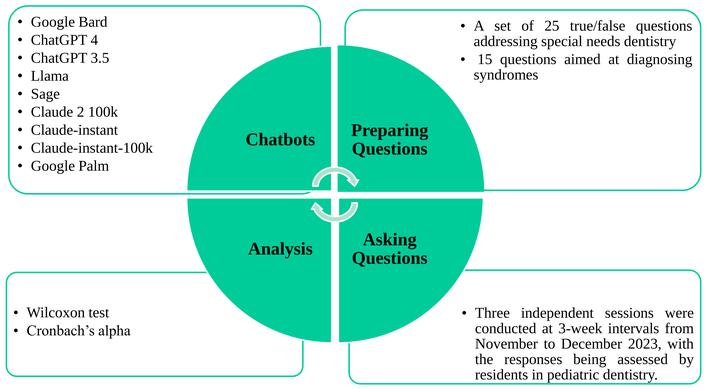
**Results:** Chatbots had an average accuracy of 55%±4% in answering all questions, 37%±6% in diagnosis, and 67%±8% in answering true/false questions. No significant difference ($p>0.05$) of the accuracy rates was detected between any pairwise chatbot comparison. All chatbots demonstrated acceptable reliability (Cronbach alpha >0.7), with Claude instant having the highest reliability of 0.93.

**Conclusion:** Chatbots exhibit high consistency in responding to all questions, and a higher accuracy in responding to true/false questions than diagnostic questions.

## INTRODUCTION

A chatbot is a form of artificial intelligence software that has been developed by processing extensive quantities of text data, often reaching into the hundreds of terabytes. These models are designed to map out the statistical frequencies and relationships of various textual elements, including words, graphemes, characters, and punctuation, found within a wide array of text that has been generated by humans and made publicly accessible. Chatbots are programmed to disseminate information and facilitate discourse across a multitude of subjects, with applications extending to the healthcare domain. A scant number of investigations have appraised chatbot proficiency in fielding questions pertinent to dental examinations, encompassing areas such as endodontics and oral oncology.

## MATERIALS AND METHODS




## RESULTS

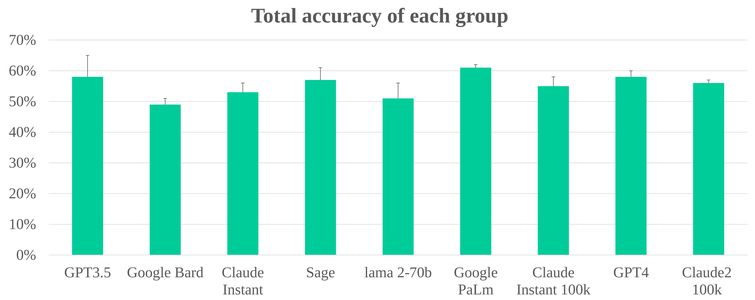



**Figure 1.** Total accuracy for chatbots to answering all the questions. Chatbots had an average accuracy of 55%±4% in answering all questions. There were no statistically significant differences between groups.

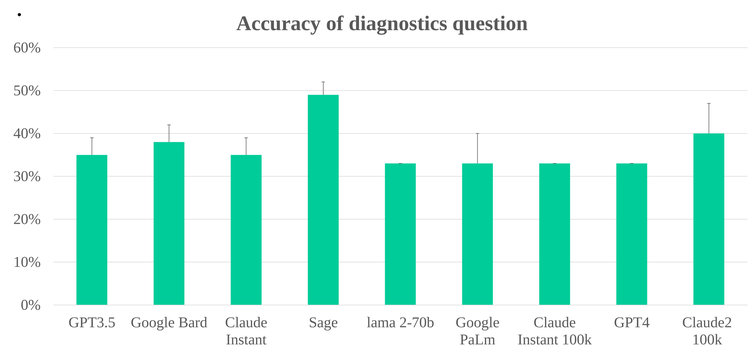



**Figure 2.** Accuracy of Chatbots in answering the diagnostic questions. Sage had the highest accuracy (49%±3%) in diagnosis questions. There were no statistically significant differences between groups.

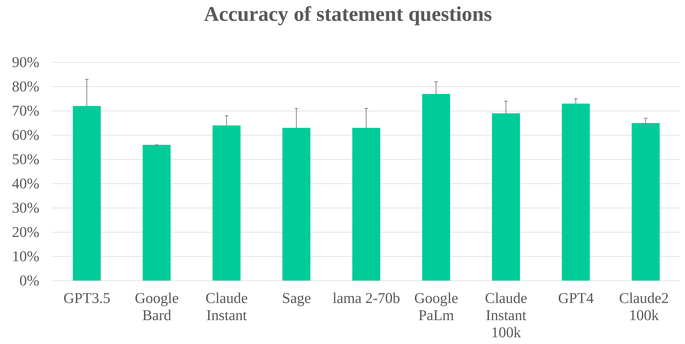



**Figure 3.** Accuracy of chatbots in answering the statement questions. Google Palm had the highest accuracy (77%±4%) in true/false questions. There were no statistically significant differences between groups.

| Group | Cronbach Alpha |
|---|---|
| GPT3.5 | 0.842 |
| Google Bard | 0.839 |
| Claude instant | 0.875 |
| Sage | 0.875 |
| Llama 2-70b | 0.802 |
| Google PaLm | 0.783 |
| Claude instant 100k | 0.923 |
| GPT4 | 0.905 |
| Claude 2 100k | 0.890 |

**Table 1.** The reliability of each chatbots. Cronbach alpha higher than 0.7 means they have acceptable reliability.

## CONCLUSIONS

- There is no significant difference in response accuracy among these nine Chatbots.
- Chatbots exhibit a higher accuracy in responding to true/false questions than diagnostic questions.
- All chatbots demonstrated acceptable reliability.

## REFERENCES

1. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, Gigante A, Valencia A, Rementeria MJ, Chadha AS, Mavridis N. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ Digit Med 2020;3:81. doi:10.1038/s41746-020-0288-5.
2. Pithpornchaiyakul S, Naorungroj S, Pupong K, Hunsrisakhun J. Using a Chatbot as an Alternative Approach for In-Person Toothbrushing Training During the COVID-19 Pandemic: Comparative Study. J Med Internet Res 2022;24:e39218. doi:10.2196/39218.
3. Tiwari A, Kumar A, Jain S, Dhull KS, Sajjanar A, Puthenkandathil R, Paiwal K, Singh R. Implications of ChatGPT in Public Health Dentistry: A Systematic Review. Cureus 2023. doi:10.7759/cureus.40367.
4. Li R, Kumar A, Chen JH. How Chatbots and Large Language Model Artificial Intelligence Systems Will Reshape Modern Medicine. JAMA Intern Med 2023;183:596. doi:10.1001/jamainternmed.2023.1835.
5. Huh AJH, Chen J-W, Bakland L, Goodacre C. Comparison of Different Clinical Decision Support Tools in Aiding Dental and Medical Professionals in Managing Primary Dentition Traumatic Injuries. Pediatr Emerg Care 2022;38:e534–9. doi:10.1097/PEC.0000000000002409.
6. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M, Smith DM. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med 2023;183:589. doi:10.1001/jamainternmed.2023.1838.