Journal of Pioneering Medical Sciences
Received: January 03, 2026 | Accepted: March 22, 2026 | Published: April 15, 2026
Volume 15, Issue S01, Pages 01-06

DOI https://doi.org/10.47310/jpms202515S0101

AI Chatbots in Endodontic Treatment Planning: A Comparative Analysis Using the AIPI Framework

Faisal Alnassar[1*]
[1]Department of Restorative and Prosthetic Dental Sciences, College of Dentistry, Majmaah University, Al-Majmaah, 11952, Saudi Arabia

Author Designation: [1]Associate Professor

*Corresponding author: Faisal Alnassar (e-mail: F.alnassar@mu.edu.sa).

Abstract: More and more research is looking at how healthcare professionals may benefit from artificial intelligence (AI) chatbots powered by LLMs while making judgments. However, not much is known about how effective they are in structured endodontic treatment planning. The degree of agreement between the treatment plans generated by chatbots and the reference plans given in published endodontic cases was evaluated in this cross-sectional study using the Accuracy of Identified Prescriptive Instructions (AIPI) framework. The four chatbot systems that were analyzed were Glass Health, MedGebra GPT-4o, Gemini 2.5 Pro, and ChatGPT 4.5. We used the AIPI scale, which ranges from 0 to 3, to evaluate the platforms. There were a total of 192 conversations as each chatbot received sixteen published endodontic case reports at three different times. Although MedGebra GPT-4o had an average AIPI score of 1.69 ($p < 0.05$), the much higher scores were attained by ChatGPT 4.5, Glass Health, and Gemini 2.5 Pro, totaling 2.69. Although all of the platforms were reliable in terms of time, ChatGPT 4.5 demonstrated the greatest level of consistency at the case level (75%). As a conclusion, although MedGebra had poorer consistency and accuracy, ChatGPT4.5, Glass Health, and Gemini 2.5 Pro all worked dependably and excellently. While these findings provide support for AI chatbots as supplemental educational and decision-support tools, they also show how important it is to validate these systems in certain domains before integrating them into clinical practice and how dangerous it is to have inconsistent or incomplete outcomes in the clinic.

Key Words: Artificial Intelligence, Chatbots, Clinical Decision Support, Endodontics, Large Language Models, Treatment Planning Accuracy

INTRODUCTION

The healthcare industry is seeing a shift due to the proliferation of AI-powered diagnostic, clinical decision-support, and patient education technologies [1]. The ability to engage in meaningful user interaction, synthesize medical knowledge, and provide recommendations based on unique instances is what makes conversational chatbots powered by LLMs so promising among these solutions [2]. A meta-analysis of recent studies indicated that LLMs may achieve expert-level results on medical license examinations and structured clinical reasoning tests [3]. Their potential as supplementary tools for clinical decision-making has been sparked by this.

Up till now, the primary domains where artificial intelligence has been used in dentistry include diagnostic imaging, the diagnosis of dental decay, and radiographic analysis [4-6]. Recent meta-analyses have shown that deep learning and convolutional neural networks are useful AI approaches for detecting, classifying, and forecasting the outcomes of oral health disorders [7,8]. Endodontic studies have shown that AI can detect periapical radiolucent lesions, correctly identify pulp chamber segments, and distinguish root canal designs [9–11]. Studies have shown that ChatGPT can accurately detect pulpal and periapical problems, sometimes even surpassing dental students [6,12].

Despite the proliferation of websites claiming to be clinical decision support tools, research on the potential use of chatbots in structured treatment planning is few. Endodontics is a promising area for research since treatment planning involves a wide variety of skills and knowledge, including radiographic interpretation, patient-specific modifiers, prognosis, procedural sequencing, pulpal and periapical diagnostics, and more. Patient safety may be jeopardized if AI systems made recommendations that seemed sensible but were really incomplete or even hallucinogenic. Students, residents, and practicing dentists

might all benefit greatly from a reliable AI tool that could enhance learning and provide decision assistance [13–15].

To objectively evaluate chatbot performance, formal scoring techniques have been proposed, such as the Accuracy of Identified Prescriptive Instructions (AIPI) [16]. On a scale from 0 to 3, AIPI determines how well AI-generated management advice matches up with a reference treatment plan, providing an alternative to only depending on diagnostic recognition. Here is a practical way to test how well prescriptions work. When compared to medical contexts where AIPI has been used to assess therapeutic decision-making, dental data is severely missing [17]. Despite AIPI's encouraging results in medical research, there aren't any endodontic-specific cross-platform comparisons that evaluate chatbot reliability just yet [18]. This technique should be introduced to endodontics for safety reasons, since critical procedural steps may still be neglected, even with partially correct guidance.

To address this knowledge gap, this research evaluated four chatbot systems for endodontic treatment scheduling: ChatGPT 4.5, Glass Health, MedGebra, and Gemini 2.5 Pro. Temporal consistency between sessions and case-level consistency throughout repeated runs were secondary goals, with total AIPI score serving as the main result. We hypothesised that chatbot performance would vary between platforms and that endodontic treatment planning reliability would not be directly proportional to general medical reasoning reliability. The results have real-world implications for developers, educators, and clinicians since they show how chatbots might help with endodontic decisions but also how important it is to validate them thoroughly before using them in real-world situations.

METHODS

Study Design and Data Sources

This cross-sectional study evaluated the performance of several chatbot systems for endodontic treatment planning using pre-existing case scenarios. Fifteen clinical case reports from scholarly journals formed the basis of the review. There were enough specifics about the diagnoses and treatments recommended for each instance to make a direct comparison to the results produced by the chatbot.

Case Selection

A diverse spectrum of clinical circumstances was represented by sixteen published case studies in endodontics. Case reports describing actual endodontic therapy in the real world, as opposed to hypothetical or unfinished cases, were needed to provide a clear diagnosis and thorough treatment instructions in order to be included. Cases that did not have clear treatment procedures were not included.
The selected cases covered:

- Pulp and periradicular tissue diagnosis
- Nonodontogenic toothache
- Invasive cervical resorption
- MTA repair of supracrestal perforation
- Palatogingival groove management
- Revitalization of a tooth with necrotic pulp and an open apex
- Acute apical abscess treatment
- Periapical cemento-osseous dysplasia
- Avulsion of maxillary incisors with complicated crown fractures
- Microsurgical endodontic retreatment
- Lateral periodontal cyst
- Antibiotic prophylaxis for medically compromised patients
- Vertical root fracture
- Retrieval of separated intracanal instruments
- Internal bleaching
- Apexogenesis of an immature permanent molar

Inclusion and Exclusion Criteria

Cases were considered for inclusion if they had a definitive diagnosis backed by clinical, radiographic, laboratory, and histological evidence and had documented thorough treatment. We did not include patients whose records were missing necessary diagnostic information or whose chatbot answers consisted only of basic disclaimers.

AI Language Models

Our study utilized four AI chatbots: two general-purpose models—Google's Gemini 2.5 Pro and OpenAI's ChatGPT 4.5—both released on February 27, 2025; and two medically oriented models—Glas Health from Glas Health in San Francisco, CA, USA, and MedGebra GPT-4o from Glas Health in Breukelen, Netherlands, respectively.

Every model comes with its own unique collection of features that are determined by the platform's requirements. Glass Health is optimized for structured clinical reasoning, MedGebra GPT-4o for retrieval-supported medical responses, Gemini 2.5 Pro for advanced reasoning tasks, and ChatGPT 4.5 for broad reasoning and conversational aid. Considering these differences, it was necessary to conduct comparative benchmarking rather than assuming that systems were equally helpful from a clinical standpoint.

Data Collection

Every one of the sixteen chatbots were asked the identical question: "What is the endodontic treatment plan based on the provided clinical findings?" The possibility of carryover bias was mitigated by having distinct researchers present each case into a chat session. During the course of the talk, no notes, questions, or clarifications were saved for future use. We were able to generate 192 chatbot conversations by feeding all four of them 16 instances at various times throughout the day: morning (8:00-10:00), afternoon (14:00-16:00), and evening (20:00-22:00). We used the AIPI framework to objectively assess the responses and compared them to the published case report's reference

treatment plan. Word by word, we received the responses. We were able to provide a general idea of how well each chatbot did by finding their median and mean AIPI scores. For the case to be considered consistent, the chatbot's AIPI score must be constant throughout all three time points.

Evaluation Framework

Chatbot outputs were assessed using the Accuracy of Identified Prescriptive Instructions (AIPI) score, a structured framework that evaluates whether AI-generated treatment plans align with published case recommendations [16]. AIPI scores range from 0 to 3, with higher scores reflecting closer agreement with the reference standard:

- 0 = No relevant instruction identified
- 1 = Partially relevant or incomplete instruction
- 2 = Mostly correct instruction with minor omissions or inaccuracies
- 3 = Fully correct instruction consistent with the reference case

This paradigm enables consistent comparisons across platforms and has been utilized in earlier research to evaluate AI systems' clinical reasoning skills [17]. While AIPI performs a good job of ensuring that recommendations are in line with reference standards, it does not take into account factors such as the quality of the narrative, the explainability of the proposal, or the severity of potentially dangerous omissions. Consequently, these factors were taken into account while interpreting the results in a qualitative manner.

Statistical Analysis

For each chatbot's AIPI score and diagnosis accuracy, descriptive statistics were calculated using the median, range, standard deviation (SD), and mean. A Kruskal-Wallis H test was used to evaluate the overall performance of the four chatbots—Gemini 2.5 Pro, MedGebra GPT-4o, Glasshealth, and ChatGPT 4.5—because the ordinal data was not normal. Following statistical significance from the Kruskal-Wallis test, post hoc pairwise comparisons were conducted using the Mann-Whitney U test. Bonferroni was used to account for multiple comparisons.

Three times a day, we ran the Friedman test for related samples on the chatbots to see how stable their temporal behavior was. We used a novel Friedman test to evaluate the chatbots' temporal consistency.

Post hoc Mann-Whitney U tests with Bonferroni correction also allowed us to find out whether the chatbot's performance was consistent or not across instances. A Kruskal-Wallis H test was also used to achieve this. When the p-value was lower than 0.05, we knew that the test had reached statistical significance.

RESULTS

In a total of 122 talks, four chatbot systems evaluated sixteen endodontic cases over three time periods. In comparison to ChatGPT4.5, Glass Health, and Gemini 2.5 Pro, which all achieved 2.69, MedGebra's average AIPI score was 1.69. The Kruskal-Wallis test confirmed that the platforms ran substantially differently in terms of overall performance with $H = 13.675$, $df = 3$, and $p = 0.003$. Post hoc pairwise testing revealed that MedGebra performed significantly worse than ChatGPT 4.5 ($p = 0.008$), Glass Health ($p = 0.014$), and Gemini 2.5 Pro ($p = 0.038$). There were no discernible shifts on the other three stations.

The grading system and platform-level performance may be shown visually in Tables 1 and 2.

There was no significant variation inside the chatbot across various time periods ($p > 0.05$ for all), according to the Friedman tests, which meant that the chatbot's performance was stable all day long.

The results from every occurrence are in agreement. The most stable version of ChatGPT was 4.5, which consistently produced the best outcomes in 12 out of 16 cases. Glass Health was second only to Gemini 2.5 Pro in reliability at 68.8% of the time, while 56.3% were dependable. Returning to last place, MedGebra was only consistent in 37.5% of cases. Even though there were differences, the Kruskal-Wallis test did not find any statistical significance in the consistency rates across the four chatbots ($H = 5.356$, $df = 3$, $p = 0.147$).

Finally, ChatGPT 4.5, Glass Health, and Gemini 2.5 Pro all did well in endodontic scenario treatment planning, delivering outcomes that were comparable to one another and earning good AIPI ratings. However, when it came to aiding in clinical decision-making, MedGebra's accuracy and reliability were below average, indicating that it could have limits.

Table 1: AIPI Scoring Framework Used in the Study

Score	Interpretation
0	No relevant instruction identified
1	Partially relevant or incomplete instruction
2	Mostly correct instruction with minor omissions or inaccuracies
3	Fully correct instruction consistent with the reference case

Table 2: Summary Comparison of Chatbot Performance Based on Reported Study Outcomes

Platform	Mean AIPI score	Case-level consistency	Temporal stability across sessions
ChatGPT 4.5	2.69	75.0%	Stable; Friedman test not significant
Glass Health	2.69	68.8%	Stable; Friedman test not significant
Gemini 2.5 Pro	2.69	56.3%	Stable; Friedman test not significant
MedGebra GPT-4o	1.69	37.5%	Stable; Friedman test not significant

DISCUSSION

In this research, four chatbots were tested utilizing endodontic case scenarios using the AIPI framework: Glass Health, Gemini 2.5 Pro, MedGebra, and ChatGPT 4.5. The results demonstrate that there is a significant platform-specific variation in treatment planning accuracy and repeatability. In addition to focusing on average accuracy, they highlight the importance of thinking about the therapeutic consequences of advice that are partial or unstable when assessing chatbots for dentistry.

The study's findings show that a few of LLMs do a fantastic job of designing endodontic treatments. On a predetermined scale, ChatGPT4.5, Glass Health, and Gemini 2.5 Pro all g*et al*most perfect scores (mean AIPI = 2.69). Numerous research investigating the use of LLMs in dentistry and other medical specialties have reached similar conclusions. Özbay *et al.* [19] shown, for instance, that when it came to pediatric dentistry, both ChatGPT and Gemini reached similarly high rates of accuracy while handling questions about diagnosis and treatment planning. Thus, it can be inferred that the scope of clinical reasoning abilities possessed by both models is more extensive. Another study that evaluated LLMs in fixed prosthodontics found that ChatGPT was quite accurate in basic cases, but it did admit that it was not up to the task of handling more complex cases [20]. Although Gemini's remarks were simpler to comprehend, ChatGPT was shown to be more accurate and reliable in treating endodontic pain in a study that examined the dependability and accuracy of LLMs [21].

We further confirmed previous studies [12] by finding that ChatGPT regularly and correctly outperforms dental students in endodontic scenarios. Similarly, throughout the period of four days, ChatGPT-4.0 maintained its position as the most consistent model in endodontic decision support testing, providing more evidence of its reliability [22]. While Danesh *et al.* [24] discovered that ChatGPT had only intermediate accuracy when it came to complex diagnostic concerns, Tokgöz Kaplan & Cankar [23] discovered that Gemini had superior accuracy when it came to oral trauma. When considered together, these findings demonstrate the impact of various domains, question kinds, and model architectures on the performance of chatbots.

Due to its consistently low ratings and inconsistent performance, MedGebra consistently lagged behind these top models. This discovery proves that clinical AI relies on certain platforms. It is difficult to draw direct analogies to MedGebra. Medgebra achieved better accuracy than ChatGPT-4o in a similar study, even though ChatGPT-4o had consistency problems [25]. The training data, underlying architecture, and intended aim of any model could significantly affect its clinical usefulness [26]. This study's results show that LLM implementation should not be done using a "one-size-fits-all" strategy. Prior to integrating any technology into clinical workflow, thorough validation is necessary.

The discovery that all four platforms remain stable over time is very noteworthy. Although their precision could change depending on the job at hand, our findings are in line with other studies on LLM consistency, which have shown generally acceptable repeatability [20]. The accuracy of ChatGPT-4 and 4o in endodontic education was shown to be unaffected by time of day, according to another study by Öztürk *et al.* [27], which validated our result of constant temporal performance. Answer types did affect consistency, however. Because they are immune to variables that could influence clinical judgment, LLMs prove to be trustworthy resources.

Even though ChatGPT 4.5 showed better numerical consistency in every example, there was no noticeable performance difference between platforms. This could be because the sample size of just 16 instances was too small to draw any firm conclusions. Since consumers may get various recommendations for similar circumstances and treatment guidance may become less reliable as it changes throughout several runs, even a little level of irregularity could affect clinical safety. Therefore, it is important for future studies to quantify not just overall agreement but also the nature and potential impact of disagreement outputs.

Only by expanding our understanding of AI's function in clinical decision-making can we make sense of the findings of this study. While humans may still be involved, chatbots that score well on the AIPI may provide treatment recommendations that are in line with what doctors suggest. Artificial intelligence (AI) has the potential to provide ideas that are reasonable but either incorrect or unsuitable for the given environment, a behavior that has been compared to hallucination in previous research [29–31]. Endodontic errors may manifest in a variety of ways, including the neglect of essential diagnostic steps, the inappropriate sequencing of procedures, or the inability to determine if further imaging or referral is necessary. The low results achieved by MedGebra in this investigation provide credence to the existence of this danger. The same holds true for chatbots; previous studies have warned against placing too much trust in them due to the many unsolved issues surrounding their variable outputs and the possibility of hallucinations [32,33]. Therefore, AI technologies should not be seen as alternatives to human judgment but as supplemental resources or ways to get a second opinion. Validation of these tools should be continuous, the models' limitations should be well understood, and patient safety should be the main priority while using them.

Keep in mind the limitations of the research as you analyze the results. The first is that there may have been insufficient power to identify clinically relevant differences due to the small sample size (16 people), especially when it came to consistency results. Secondly, the study used published case reports as a standard against which to measure. Be advised that these reports may not fully reflect the typical clinical diversity, hazy presentations, or missing history that doctors really encounter in their practice. Thirdly, including multimodal inputs like radiographic photos, CBCT findings, etc. might have improved the chatbot's performance, however the research just employed textual data. Finally, despite the independent grading of the outputs, inter-rater

reliability statistics should be included in future validation studies. Unfortunately, these statistics are not currently available in the dataset. Finally, this study did not clearly categorize error types, output structure, readability, or the seriousness of potentially dangerous recommendations; doing so could improve future safety-oriented evaluations.

Several areas of future research, development, and dentistry practice applications using AI-powered chatbots are affected by the findings. We want larger and more diverse case datasets, in addition to the capability to directly compare outcomes to those of practicing physicians and training cohorts. It would be beneficial for future studies to categorize mistakes and distinguish between harmless information gaps and dangerous suggestions. The effectiveness of these technologies in improving efficiency, learning, and treatment planning quality in real-world contexts such as schools and hospitals can only be determined via user-centered research. Responsibility in implementation requires domain-specific tailoring, transparent reporting, and safety monitoring.

CONCLUSION

Chatbots behave differently when endodontic treatment planning is included, as shown in this research. When compared to ChatGPT4.5, Glass Health, and Gemini 2.5 Pro, MedGebra's AIPI performance was poor in terms of accuracy and consistency. While AI chatbots show great potential for enhancing dental education and clinical reasoning, doctors should thoroughly monitor and evaluate their results, with patient safety as a top priority, before integrating them into daily clinical practice.

Declaration of Interest

This research does not include any conflicts of interest on the part of the author.

Data Availability

Please contact the relevant author if you would like a copy of the data that back up the study's conclusions.

Funding

It was said that no financing was used for this piece.

Ethics Declaration

Not applicable

REFERENCES

[1] Kothinti, R.R. "Deep learning in healthcare: transforming disease diagnosis, personalized treatment, and clinical decision-making through AI-driven innovations." *World Journal of Advanced Research and Reviews*, vol. 24, 2024, pp. 2841–2856.

[2] Chow, J.C. and Li, K. "Large language models in medical chatbots: opportunities, challenges, and the need to address AI risks." *Information*, vol. 16, 2025, pp. 549.

[3] Liu, M. *et al.* "Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis." *Journal of Medical Internet Research*, vol. 26, 2024, pp. e60807.

[4] Alam, M.K. *et al.* "Applications of artificial intelligence in the utilisation of imaging modalities in dentistry: a systematic review and meta-analysis of in-vitro studies." *Heliyon*, vol. 10, 2024, pp. e24221.

[5] Ismail, M.I.B. *et al.* "Diagnostic applications of artificial intelligence in dental care for medically compromised patients: a scoping review." *Digital Dental Journal*, vol. 2, 2025, pp. 100036.

[6] de Moura, J.D.M. *et al.* "Comparative accuracy of artificial intelligence chatbots in pulpal and periradicular diagnosis: a cross-sectional study." *Computers in Biology and Medicine*, vol. 183, 2024, pp. 109332.

[7] Ahmed, N. *et al.* "Artificial intelligence techniques: analysis, application, and outcome in dentistry—a systematic review." *BioMed Research International*, vol. 2021, 2021, pp. 9751564.

[8] Hung, M. *et al.* "Artificial intelligence in dentistry: harnessing big data to predict oral cancer survival." *World Journal of Clinical Oncology*, vol. 11, 2020, pp. 918.

[9] Hiraiwa, T. *et al.* "A deep-learning artificial intelligence system for assessment of root morphology of the mandibular first molar on panoramic radiography." *Dentomaxillofacial Radiology*, vol. 48, 2019, pp. 20180218.

[10] Lin, X. *et al.* "Micro–computed tomography–guided artificial intelligence for pulp cavity and tooth segmentation on cone-beam computed tomography." *Journal of Endodontics*, vol. 47, 2021, pp. 1933–1941.

[11] Sadr, S. *et al.* "Deep learning for detection of periapical radiolucent lesions: a systematic review and meta-analysis of diagnostic test accuracy." *Journal of Endodontics*, vol. 49, 2023, pp. 248–261.

[12] Qutieshat, A. *et al.* "Comparative analysis of diagnostic accuracy in endodontic assessments: dental students vs. artificial intelligence." *Diagnosis*, vol. 11, 2024, pp. 259–265.

[13] Patil, S.R. and Karobari, M.I. "Exploring artificial intelligence for enhanced endodontic practice: applications, challenges, and future directions." *Advances in Public Health*, vol. 2024, 2024, pp. 8075515.

[14] Eggmann, F. and Blatz, M.B. "ChatGPT: chances and challenges for dentistry." *Compendium of Continuing Education in Dentistry*, vol. 44, 2023.

[15] Glick, A. *et al.* "Impact of explainable artificial intelligence assistance on clinical decision-making of novice dental clinicians." *JAMIA Open*, vol. 5, 2022, pp. ooac031.

[16] Lechien, J.R. *et al.* "Validity and reliability of an instrument evaluating the performance of intelligent chatbot: the artificial intelligence performance instrument (AIPI)." *European Archives of Oto-Rhino-Laryngology*, vol. 281, 2024, pp. 2063–2079.

[17] Dronkers, E.A. *et al.* "Evaluating the potential of AI chatbots in treatment decision-making for acquired bilateral vocal fold paralysis in adults." *Journal of Voice*, 2024.

[18] Maniaci, A. *et al.* "AI in clinical decision-making: ChatGPT-4 vs. Llama2 for otolaryngology cases." *European Archives of Oto-Rhino-Laryngology*, 2025.

[19] Özbay, Y. *et al.* "Evaluation of the performance of large language models in clinical decision-making in endodontics." *BMC Oral Health*, vol. 25, 2025, pp. 648.

[20] İşisağ, Ö. and Karakaya, K. "Assessing the accuracy, repeatability, and consistency of ChatGPT 4o in treatment planning for tooth-supported fixed prostheses: a comparative analysis of simple and complex clinical cases." *Clinical Oral Investigations*, vol. 29, 2025, pp. 433.

[21] Aljamani, S. *et al.* "Evaluating large language models in addressing patient questions on endodontic pain: a comparative analysis of accessible chatbots." *Journal of Endodontics*, 2025.

[22] Büker, M. *et al.* "Comparative performance of chatbots in endodontic clinical decision support: a 4-day accuracy and consistency study." *International Dental Journal*, vol. 75, 2025, pp. 100920.

[23] Tokgöz Kaplan, T. and Cankar, M. "Evidence-based potential of generative artificial intelligence large language models on dental avulsion: ChatGPT versus Gemini." *Dental Traumatology*, vol. 41, 2025, pp. 178–186.

[24] Danesh, A. *et al.* "Innovating dental diagnostics: ChatGPT's accuracy on diagnostic challenges." *Oral Diseases*, vol. 31, 2025, pp. 911–917.

[25] Abdulrab, S. *et al.* "Performance of 4 artificial intelligence chatbots in answering endodontic questions." *Journal of Endodontics*, vol. 51, 2025, pp. 602–608.

[26] Yang, X. *et al.* "Application of large language models in disease diagnosis and treatment." *Chinese Medical Journal*, vol. 138, 2025, pp. 130–142.

[27] Arılı Öztürk, E. *et al.* "Evaluation of the performance of ChatGPT-4 and ChatGPT-4o as a learning tool in endodontics." *International Endodontic Journal*, 2025.

[28] Liu, Z. *et al.* "The performance of large language models in dentomaxillofacial radiology: a systematic review." *Dentomaxillofacial Radiology*, 2025.

[29] Aditya, G. "Understanding and addressing AI hallucinations in healthcare and life sciences." *International Journal of Health Sciences*, 2024.

[30] Gondode, P. *et al.* "Artificial intelligence hallucinations in anaesthesia: causes, consequences and countermeasures." *Indian Journal of Anaesthesia*, vol. 68, 2024, pp. 658–661.

[31] Hamid, O.H. "Beyond probabilities: unveiling the delicate dance of large language models (LLMs) and AI-hallucination." In: *Proceedings of the IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, 2024, pp. 85–90.

[32] Reader, A. and Drum, M. "A review of ChatGPT as a reliable source of scientific information regarding endodontic local anesthesia." *Journal of Endodontics*, 2025.

[33] Rokhshad, R. *et al.* "Accuracy and consistency of chatbots versus clinicians for answering pediatric dentistry questions: a pilot study." *Journal of Dentistry*, vol. 144, 2024, pp. 104938.