Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics

Running title: ***AI chatbots in Endodontics***

H. Mohammad-Rahimi DDS [1*]

S.A.H. Ourang DDS [2*]

M.A. Pourhoseingholi PhD [3]

O. Dianat DDS, MS, MDS [4,5]

P.M.H. Dummer BDS, MScD, PhD, DDSc [6]

A. Nosrat DDS, MS, MDS [4,5]

1. Topic Group Dental Diagnostics and Digital Dentistry, ITU/WHO Focus Group AI on Health, Berlin, Germany

2. Dentofacial Deformities Research Center, Research Institute of Dental Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran

3. Gastroenterology and Liver Diseases Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran

4. Division of Endodontics, Department of Advanced Oral Sciences and Therapeutics, University of Maryland, School of Dentistry, Baltimore, Maryland, USA

5. Private Practice, Centreville Endodontics, Centreville, Virginia, USA

6. School of Dentistry, College of Biomedical and Life Sciences, Cardiff University, Cardiff, UK

*Authors contributed equally to this project.

ORCID IDs

H. Mohammad-Rahimi: **0000-0002-4971-5926**

S.A.H. Ourang DDS: **0009-0009-8521-2718**

M.A. Pourhoseingholi: **0000-0002-0121-8031**

O. Dianat: **0000-0001-8768-0456**

P.M.H. Dummer: ***0000-0002-0726-7467***

A. Nosrat: **0000-0003-4768-9717**

Corresponding Author

Dr. Ali Nosrat

Division of Endodontics, Department of Advanced Oral Sciences and Therapeutics, University of Maryland, School of Dentistry

650 West Baltimore St, 4th floor

Baltimore, MD 21201

USA

Email: Nosrat@umaryland.edu

Key words: *Artificial intelligence; Bing; Chatbot; ChatGPT; Endodontics; Google Bard; Large Language Models; Public health.*

Funding

The authors confirm that the study received no funding.

Conflict of interest

The authors confirm they have no conflict of interest.

Author Contribution

H. Mohammad-Rahimi: Conceptualization, Methodology, Data curation, Writing –original draft, Writing – review & editing. **S.A.H. Ourang:** Conceptualization, Data curation, Writing – review & editing. **M.A. Pourhoseingholi:** Conceptualization, Methodology, Statistical analyses**;** **O. Dianat:** Conceptualization, Methodology, Data curation, Writing – review & editing. **P.M.H. Dummer:** Methodology, Writing – review & editing **A. Nosrat:** Conceptualization, Methodology, Data curation, Writing – review & editing.

Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics

Abstract

Aim This study aimed to evaluate and compare the validity and reliability of responses provided by GPT-3.5, Google Bard, and Bing to frequently asked questions (FAQs) in the field of endodontics.

Methodology FAQs were formulated by expert endodontists (n=10) and collected through GPT-3.5 queries (n=10), with every question posed to each chatbot three times. Responses (N=180) were independently evaluated by two board-certified endodontists using a modified Global Quality Score (GQS) on a 5-point Likert Scale (5: strongly agree; 4: agree; 3: neutral; 2: disagree; 1: strongly disagree). Disagreements on scoring were resolved through evidence-based discussions. The validity of responses was analyzed by categorizing scores into valid or invalid at two thresholds: The low threshold was set at score ≥ 4 for all three responses whilst the high threshold was set at score 5 for all three responses. Fisher's Exact test was conducted to compare the validity of responses between chatbots. Cronbach's alpha was calculated to assess the reliability by assessing the consistency of repeated responses for each chatbot.

Results All three chatbots provided answers to all questions. Using the low threshold validity test (GPT-3.5: 95%; Google Bard: 85%; Bing: 75%) there was no significant difference between the platforms (p>0.05). When using the high threshold validity test, the chatbots scores were substantially lower (GPT-3.5: 60%; Google Bard: 15%; Bing: 15%). The validity of GPT-3.5

responses was significantly higher than Google Bard and Bing (p=0.008). All three chatbots achieved an acceptable level of reliability (Cronbach's alpha >0.7).

Conclusions GPT-3.5 provided more credible information on topics related to endodontics compared to Google Bard and Bing.

Introduction

Artificial intelligence (AI) chatbots have revolutionized digital communication by disseminating information and providing opportunities for individuals to ask specific and personalized questions and receive unique and focused responses. With deep learning algorithms, chatbots can now learn from large amounts of data that has provided the potential for them to improve their responses over time. This is performed by mimicking the neural networks of the human brain (LeCun *et al.* 2015; Schmidhuber 2015). GPT-3.5 (OpenAI Inc., San Francisco, CA, USA), Google Bard (Google LLC, Mountain View, CA, USA), and Bing (Microsoft Corporation, Redmond, WA, USA) are platforms that are at the forefront of this evolution.

GPT-3.5, powered by OpenAI Inc. is based on a transformer-based model, named generative pretrained transformer (GPT) trained on diverse internet text. It employs unsupervised learning, improving its responses through the understanding of patterns in its training data (Liu *et al.* 2023). GPT-3.5 reached 100 million active monthly users in January 2023, only two months after it was launched (Milmo 2023). On the other hand, Google Bard leverages the expertise of Google in search and language understanding. It uses bidirectional encoder representations from transformers (BERT), another transformer-based model, which has been specifically developed to understand the context of words in a sentence for a more accurate representation of language (Devlin *et al.* 2018; Pichai 2023). Bing is a product of the vast AI research and development resources of Microsoft. While less is known about its specific architecture, it benefits from Microsoft's comprehensive web-scale data, GPT-3.5, and significant advancements in machine learning and AI techniques (Mehdi 2023). All these chatbots function interactively with users. First, they receive an input prompt from the user. This prompt is processed and encoded into a mathematical representation using a deep learning model. Then, the chatbot attempts to predict the most plausible next response or sequence of responses based on the patterns it has learned from its training data (Sutskever *et al.* 2014).

Chatbots are designed to provide information and facilitate conversations on a wide range of topics, including healthcare. These tools can improve patients' understanding of treatments, side effects/complications, prognoses, and outcomes by providing responses to patient queries (Ayers *et al.* 2023a; Safi *et al.* 2020). Validity and reliability of health information provided to patients are critical for improving patients' awareness of their health condition and avoiding harm. When patients have access to accurate, evidence-based information from reliable sources, they are empowered to make informed decisions about their health (Edwards *et al.* 2001). Valid information allows patients to correctly understand their condition, treatment options, and prognosis. This enables patients to provide informed consent, adhere to recommended treatments, and have realistic expectations about results (Austvoll-Dahlgren & Helseth 2010). Suarez et al. (2023), examined the performance of GPT-4.0 in answering dichotomous (yes/no) endodontic questions formulated based on Position Statements by the European Society of Endodontology. GPT-4.0 reached an overall accuracy of 57.33%. Despite their usage in other areas of healthcare, the performance of AI chatbots as public sources of information within the field of endodontics has not yet been assessed. The aim of this study was to evaluate and compare the validity and reliability of responses provided by three chatbot systems, namely GPT-3.5 by OpenAI, Google Bard by Google, and Bing by Microsoft, to a series of frequently asked questions (FAQs) in the field of endodontics.

Materials and Methods

Data collection

Twenty FAQs about endodontics and root canal treatments were formulated/selected, representing a broad spectrum of patient queries. These questions were collected through two sources:

A- Ten questions were selected based on those commonly asked or encountered during patient consultations with two board certified endodontists (A.N. and O.D.) who work in a full-time endodontic practice.

B- A list of the top 30 FAQs in the field of endodontics that were provided by GPT-3.5 upon request. Questions with similar content to those formulated by expert endodontists were removed and then the top 10 FAQs were selected.

The list of 20 questions is presented in Table 1. These questions cover several fields such as terminology, diagnosis, procedural/technical details of treatments, post-operative care, outcome and prognosis, risks and side effects of treatments, prevention, and alternative treatment options.

Each of the 20 questions was then posed to the three AI chatbots. Each question was repeated three times to assess their reliability. To simulate real-world interaction, the main application programming interface (API) of each chatbot was used to facilitate this process. The following approaches were used:

GPT-3.5: The API was accessed using https://chat.openai.com/. The free version of the API (i.e., GPT-3.5) was used. All questions were asked on the same day (May 24, 2023). A new chat was created for each question.

Google Bard: The API was accessed using https://bard.google.com/. All questions were asked on the same day (May 24, 2023). The chat conversation was reset before asking each new question.

Bing: The API was accessed using Microsoft Edge using https://www.bing.com/chat. All questions were asked on the same day (May 24, 2023). A new chat was created for each question. The "More balanced" conversation style for the responses was chosen.

Scoring

All responses were evaluated independently by two board-certified endodontists (A.N. and O.D.) using a 5-point Likert Scale. A modified version of the Global Quality Score (GQS)(Bernard *et al.* 2007) was used to assign scores based on the 'context' and 'content' of the responses:

Score 5 (Strongly Agree): The answer is correct, and the content is comprehensive.

Score 4 (Agree): The answer is correct and most of the content is correct, but it lacks information, or contains incorrect information.

Score 3 (Neutral): The answer is somewhat correct, but details are primarily incorrect, missing, or irrelevant.

Score 2 (Disagree): The answer is incorrect, but the content includes some correct elements.

Score 1 (Strongly Disagree): The answer and the entire content are incorrect or irrelevant.

By applying these criteria, the chatbot responses were evaluated in a nuanced manner, considering both the correctness of the answers (context) and the completeness and accuracy of the information provided (content). This allowed the validity of responses to be assessed. By repeating the questions three times, the consistency of each chatbot was analyzed, to assess their reliability in providing responses.

Once the scoring was completed independently, the two reviewers shared the score sheets (total of 180 scores each), reviewed, and discussed those answers that were scored differently. Disagreements were resolved through evidence-based discussions over the context and content. Finally, a single scoresheet was prepared for all 180 answers for the purpose of statistical analyses.

Statistical analysis

Analysis of validity: To define the validity of each response (i.e., the degree to which an answer to a question from each chatbot accurately captured what it was intended to answer) the scores were divided into two categories of ‘valid’ and ‘invalid’. Two tests of validity were employed: a Low threshold validity test and a high threshold validity test.

For the low threshold test, the score threshold was 4. If all three responses to a question scored ≥4, the chatbot's answer was deemed valid. If any response scored less than 4, the chatbot's answer was considered invalid. For the high threshold test, the score threshold was 5. Here, the chatbot's answer was considered valid only if all three responses scored 5. If any response scored less than 5, the chatbot's answer was considered invalid.

Fisher’s Exact test was conducted to compare the validity of responses between chatbots. The significance level was set at < 0.05.

Analysis of reliability: Reliability (the degree to which the chatbot produced similar answers when used repeatedly under consistent conditions) was defined by analyzing the 5-point Likert scale scores assigned to the answers to questions when they were repeated three times. To evaluate the consistency of responses (i.e., the reliability) Cronbach’s alpha was calculated for all three responses for all 20 questions together. Cronbach's alpha quantifies the level of consistency on a standardized scale of 0 to 1: 0 indicating no consistency and 1 indicating perfect consistency in responses. A high alpha coefficient suggests that the chatbot provided the same construct consistently and the scale could be considered reliable. A lower alpha coefficient suggests that the chatbot did not provide the same construct consistently and the scale is considered less reliable. A Cronbach's alpha of ≥ 0.70 indicates acceptable reliability in medical and health-related research (Bland *et al.* 1997).

All analyses were performed using R Programming language version 4.1.0.

Results

Descriptive analyses

The three chatbots responded to all 20 questions repeated three times, and a total of 180 answers were collected. The list of answers is provided in supplemental materials. Overall, Bing provided relatively shorter answers (mean word count: 89; range: 50-146) with fewer details. The answers from GPT-3.5 (mean word count: 216; range: 147-340) and Google Bard (mean word count: 233; range: 111-322) were longer and the responses contained more details. When questions were repeated, Bing remained consistent by repeating the same answer with minor or no changes in wording. On the other hand, GPT-3.5 and Google Bard provided answers with different wording, and sometimes with different structures and details to the same questions. All three chatbots frequently encouraged their users to see a dentist or an endodontist for further details about their queries: GPT-3.5 52/60 times; Google Bard 57/60 times; and Bing 13/60 times. The three chatbots provided several wrong or irrelevant statements in their responses. A list of these statements is provided in Table 2.

Statistical analyses

Details of all scores, after agreement between the assessors, for the 180 answers are provided in supplemental Table 1. The average score for the three answers provided to each question is presented in Figure 1. Scores ranged from 2 to 5. Only Bing had a score of 2 in one question (Q6). Otherwise, the scores ranged from 3 to 5.

Low threshold validity: All responses from the three chatbots reached a relatively high level of validity. GPT-3.5 had the highest overall validity, with 19 out of 20 (95%) answers categorized as valid. Google Bard and Bing had lower validities overall, with 17 out of 20 (85%) and 15 out of 20 (75%) valid answers, respectively. There were no significant differences between the three chatbots when analyzed using the low threshold validity test ($p > 0.05$) (Fig. 2).

High threshold validity: All responses from the three chatbots reached a relatively low level of validity when tested against the high threshold test. GPT-3.5 had the highest overall validity, with 12 out of 20 (60%) answers categorized as valid. Google Bard and Bing had lower overall validities, with 3 out of 20 valid responses (15%) for both. The responses of GPT-3.5 reached a significantly higher validity compared to both Google Bard and Bing ($p = 0.008$). Google Bard and Bing were not significantly different ($p = 1.00$) (Fig. 3).

Reliability: All three chatbots had an acceptable level of reliability. Out of the three chatbots, Bing had the highest overall consistency with a Cronbach’s alpha of 0.955, followed by GPT-3.5 with a Cronbach's alpha of 0.746, and Google Bard with a Cronbach's alpha of 0.703.

Discussion

AI chatbots have emerged as a new, powerful, and easy-to-access source of information with the potential to change the way individuals receive and process information, including healthcare information (Walker *et al.* 2023). It is extremely important for scientific bodies such as medical and dental associations/societies to critically appraise the information provided by chatbots and to inform their members as well as the public about the inherent benefits, threats, and deficiencies of these sources of data. Even though it is impossible to police AI chatbots and to eliminate the spread of misinformation, it is the duty of scientists and researchers to fact-check the information provided by them. Use of AI chatbots in healthcare has increased in recent years, with the potential to improve patient satisfaction and reduce healthcare costs (Xu *et al.* 2021). In dentistry, especially in the field of endodontics, using chatbots to answer FAQs posed by patients could provide a valuable resource for those seeking information about their dental condition, potential care and treatment outcomes. The present study is the first to assess the validity and reliability of artificial intelligence chatbots in responding to questions in the field of endodontics.

When testing the validity of responses provided by chatbots the source of questions and the language of questions are important. The questions provided in this paper were aimed to reflect the public's questions/concerns about endodontics and endodontic treatments. Ten questions were formulated by endodontists who practice full time and have daily interactions with patients. An alternative to this approach is to gather questions from a real group of patients. But this approach can be challenging because it cannot include questions/concerns of different groups of patients from different socio-economic/cultural/racial backgrounds worldwide. The other 10 questions were provided by GPT-3.5 as its top FAQs in the field of endodontics. As a large language model trained on massive text datasets, GPT-3.5 can synthesize diverse conversational content across a myriad of subjects. By querying GPT-3.5 to provide its top FAQs in endodontics, a representative sample of texts that GPT-3.5 trained on was obtained, encompassing a vast array of online content as well as academic literature (Cascella *et al.* 2023). Therefore, the questions are likely to contain the most frequently voiced concerns of patients reflected in these sources. All in all, it is complicated to generate representative and inclusive questions. Future studies should provide questions prepared by larger panels of clinicians and patients from different geographical/socio-economic/cultural/racial backgrounds on a global scale to be more representative of the "questions likely to be asked by the public". Another limitation of the present study that is relevant to this debate is that the validity of responses was assessed by only two board-certified endodontists. It would have been desirable to have a larger panel of experts to examine the validity of answers. We recommend that endodontic societies/associations should take over this critical responsibility and form task forces and expert panels to continually publish their assessments on the quality of endodontic information provided by these chatbots and to disseminate their findings to all key stakeholders, including the general public.

The validity of responses provided by chatbots were assessed using two different (low and high) thresholds. At the low threshold test all three chatbots' responses had an overall high validity. When assessed under a high threshold, the validity of responses by all three chatbots dropped considerably. In some circumstances, endodontic diseases can become life threatening with serious morbidities and mortalities at the tooth and patient levels. For instance, if left untreated or treated improperly, an acute apical abscess can result in the death of a patient (Rampa *et al.* 2019). Therefore, misinforming the public on the topic of tooth infections can have serious health consequences. A similar situation in the present study occurred when Google Bard recommended pregnant patients with allergies to local anaesthetics to go through a general anaesthetic for a root canal treatment to avoid complications (Q11, Tables 1 and 2). This is a non-evidence-based recommendation with serious consequences. Misinforming the public combined with lack of easy access to valid sources of information can create widespread chaos in healthcare systems. Recent examples of this statement are public hesitation in visiting dentists throughout the corona-virus disease 2019 (COVID-19) outbreak and pandemic due to fear of contracting the SARS-CoV-2 virus (Nosrat *et al.* 2022a; Nosrat *et al.* 2022b), or misinformation about a link between COVID-19 vaccines and infertility that resulted in vaccine hesitancy in certain social groups (Abbasi 2022). Therefore, since there is free and easy access to these chatbots worldwide, it is necessary for researchers and assessor entities (i.e., endodontic societies/associations) to insist on high standards when assessing their performances in the field of endodontics. The spread of false information remains an ongoing threat by these chatbots. A careful and detailed examination of endodontic information provided by chatbots is needed before referring to them as "reliable" sources of information to the public.

From a technical standpoint, the overall better performance of GPT-3.5 can be attributed to its unique underlying technology. The outcome of the present study is in line with the study by Doshi et al (2023) which compared the same three chatbots for simplifying radiology reports, in which GPT-3.5 had the best performance. Another report concluded that GPT-3.5 provided evidence-based responses to public health questions in the fields of addiction, mental health, and physical health (Ayers *et al.* 2023b). The data used for training these AI models are different, which can explain the variations in their responses. These variations can lead to differences in the narratives of the chatbots when providing human-like responses (Bhardwaz *et al.* 2023; Cascella et al. 2023). GPT-3.5 is based on GPT, which is designed to generate human-like text by predicting the likelihood of a word given the preceding words in a sentence. Consequently, it can generate varying responses depending on the context and its previously learned patterns from extensive datasets (Sanmarchi *et al.* 2023). On the other hand, while it is known that Google Bard is based on BERT (Pichai 2023) and Bing is based on GPT (Mehdi 2023) technologies, the exact architectures and technical details about the models are not available publicly, which limits the ability to understand why the chatbots perform so differently when endodontic questions were asked. The differences in chatbot responses can also be attributed to the varying design philosophies of the AI companies, the specific algorithms employed, the datasets used for training, and the objectives that the AI is designed to achieve (Bhardwaz & Kumar 2023; Jianfeng *et al.* 2019).

The subject of reliability, as a measure of consistency, is also a crucial aspect of the performance of these chatbots. There are no previous peer-reviewed medical/dental studies published on the reliability of chatbots responses to questions that may be posed by the public. All three chatbots had an acceptable level of reliability. The designs of these chatbots are based on deep learning models that have some level of inherent randomness. Therefore, the responses given by the chatbot are not deterministic (Lu *et al.* 2023). In the present study, Bing repeated similar answers consistently. In contrast, Google Bard and GPT-3.5 provided different narratives each time the question was repeated. Bing's consistency suggests that this model utilizes deterministic processes, or a narrower range of responses, to ensure the information's reliability, which are not necessarily valid. On the other hand, the varied responses of Google Bard and GPT-3.5 have the potential to provide users with a more comprehensive understanding of the topic, as they will gain from the range of perspectives provided by each response.

It is worth noting that despite overall high validity scores, these chatbots made critical errors in some of their responses. Several of these responses have the potential to mislead the public on certain topics. One of the questions chatbots consistently struggled with was, "Is it possible to have a root canal on a tooth with a dental crown?". All chatbots started their answers by stating that “the existing crown must be removed or grinded in order for the dentist to gain access to the pulp”. This is incorrect in most clinical scenarios, especially when the crown was recently placed on a tooth with a vital pulp. The complexity of this question may be rooted in the multiple facets of dental knowledge it encompasses - the procedure of a root canal, the structure and function of a dental crown, and the interaction of the two. A correct and comprehensive response would require an understanding that while a dental crown is designed to cover and protect a damaged tooth, it does not exclude the possibility of underlying issues that may necessitate a root canal procedure. Furthermore, the root canal procedure can be more challenging due to the presence of a crown, but it certainly can be performed by a skilled clinician. The responses of AI chatbots often lacked this depth of understanding, demonstrating their limitations in handling complex queries.

A significant limitation of chatbots such as GPT-3.5 in responding to health-related questions is their inability to provide references or citations for the information they generate. Without citing sources, chatbots present claims and data that cannot be verified. This is problematic as the chatbot may synthesize information or pull it from low-quality sources during training, resulting in responses that are biased, outdated, incomplete, or factually incorrect. Furthermore, the sources the chatbot use to provide information might not be public health organizations or high impact scientific journals. This lack of transparency regarding the quality of the training data used by chatbots raises concerns about relying on them for sensitive health information. More work is needed to develop the capabilities of chatbots to cite reputable sources of data and to provide audit trails when answering health related questions.

Conclusion

This study highlights the potential of AI chatbots as public sources for endodontic information. While GPT-3.5 had promising results regarding the validity of its responses, there are areas for future improvements for all three chatbots. As chatbot technologies advance, it is crucial for researchers and endodontic societies/associations to evaluate their performance and publicise their limitations to keep the public informed of valid information and possible misinformation provided by AI chatbots.

Legends

Figure 1. Mean scores for the responses of chatbots to 20 frequently asked questions. Each question was asked 3 times.

Figure 2. Low threshold validity test: results of the Fisher’s Exact test on two-by-two comparisons of the three chatbots on validity of responses provided to 20 questions, each question repeated 3 times. The threshold of validity is the minimum score of 4 in the 5-point Likert Scale obtained in all three responses. The significance level was set at < 0.05. ns: not significant.

Figure 3. High threshold validity test: results of the Fisher’s Exact test on two-by-two comparisons of the three chatbots on validity of responses provided to 20 questions, each question repeated 3 times. The threshold of validity is the minimum score of 5 in the 5-point Likert Scale obtained in all three responses. The significance level was set at < 0.05. The asterisk shows a significant difference. ns: not significant.

Table 1. List of frequently asked questions. A- questions formulated by expert endodontists through their daily interactions with patients. B- questions provided by GPT-3.5 as the top “frequently asked questions” related to endodontics.

Table 2. List of wrong/irrelevant statements by three chatbots in response to 20 FAQs. Note that this table does not provide a full picture on why a chatbot did not receive a score of 5 in each question because it does not show the missing details/information in answers. See Table-1 for the list of questions.

Supplemental Table 1. List of all 180 scores after agreement between the two reviewers.

Table 1. List of frequently asked questions. A- questions formulated by expert endodontists through their daily interactions with patients. B- questions provided by GPT-3.5 as the top “frequently asked questions” related to endodontics.

		QUESTION
A	1	What is root canal treatment?
	2	What are the signs that someone needs a root canal treatment?
	3	Is root canal treatment painful?
	4	How is the recovery after the root canal treatment?
	5	Is it worth doing root canal treatment or is it better to extract the tooth?
	6	Is root canal treatment a permanent solution for tooth pain?
	7	Can a tooth that has root canal treatment still get infected?
	8	What is the difference between root canal treatment and dental implant?
	9	What is the success rate of root canal treatment?
	10	What are the risks associated with root canal treatment?
B	11	Can I get a root canal during pregnancy?
	12	Can a cracked tooth be treated with endodontics?
	13	Is it possible to have a root canal on a tooth with a dental crown?
	14	What is endodontic abscess?
	15	What happens if I don't get a root canal?
	16	Can endodontic treatment help with tooth sensitivity?
	17	Is it necessary to place a dental crown after a root canal?
	18	Can I eat, drink and drive after a root canal?
	19	What can I do to prevent the need for endodontic treatment in the future?
	20	How long does it take to complete a root canal treatment?

Table 2. List of wrong/irrelevant statements by three chatbots in response to 20 FAQs. Note that this table does not provide a full picture on why a chatbot did not receive a score of 5 in each question because it does not show the missing details/information in answers. See Table-1 for the list of questions.

Chatbot	Question #	Wrong/irrelevant statement(s)
GPT-3.5		
	Q4	most people are able to return to their normal activities within a few days after a root canal treatment
	Q13	The dentist will need to remove the crown before accessing the underlying tooth structure and performing the root canal treatment
	Q15	If an infected tooth is left untreated, the infection can potentially spread to the neighboring teeth, leading to new areas of decay or infection. This can create a domino effect, requiring additional dental treatments and potentially affecting multiple teeth.
	Q17	If the treated tooth is a premolar or a molar located in the back of the mouth that does not experience heavy biting forces, a crown may not be required.
Google Bard		
	Q7	Eating a healthy diet will help keep your immune system strong, which can help fight off infection.
	Q8	Root canals are typically less expensive than dental implants. However, dental implants are more durable and can last for many years.
	Q10	Allergic reaction. There is a small risk of an allergic reaction to the local anesthetic or other medications used during the procedure.
	Q10	Some people experience sensitivity to hot or cold after a root canal
	Q11	Schedule your root canal for the second trimester. This is the time when you are most likely to feel your best and have the least amount of morning sickness.
	Q11	However, if you are allergic to local anesthetics, you may need to have a general anesthetic. General anesthetics are generally considered safe during pregnancy, but they do carry a small risk of side effects.
	Q13	If a tooth has a crown, the dentist will need to remove the crown in order to access the root canal system. When a tooth needs a root canal, the dentist will first remove the crown.
	Q15	If the infection in your tooth is not treated, it can spread to other teeth in your mouth. This can lead to more root canals or even tooth extractions.
	Q15	The tooth can die. A dead tooth can become loose and fall out.
	Q15	Sepsis is a life-threatening condition that occurs when the body's response to an infection damages its own tissues and organs.
	Q18	You can drink after a root canal, but avoid hot or cold drinks. These can irritate the tooth that was treated.
	Q18	The anesthetic can make you drowsy, so it is important to wait until you feel alert before driving. If you are taking pain medication, you should also wait until you feel the effects of the medication before driving.
Bing		
	Q3	The pain you feel is because of an infection in the tooth not because of root canal treatment.
	Q6	root canal is not a permanent solution for tooth pain but it can still fix the problem for many years to come
	Q7	Technically, a pulp-less root canal cannot be infected anymore because the infection can only happen on living tissue
	Q8	However, root canal therapy may have a shorter life expectancy, health risks, and jaw bone infections.
	Q10	Inadequate fillings: One of the major risks related to root canals is the inadequate fillings.
	Q11	By the second trimester, nausea should have subsided—the last thing you want to deal with when you have an overactive gag reflex is dental work!

Q13	During a root canal treatment, the dentist will grind down the original crown of your tooth to treat the infected pulp in the roots of your tooth.
Q20	The duration of a root canal treatment depends on the amount of infection in your tooth

References

Abbasi J (2022) Widespread Misinformation About Infertility Continues to Create COVID-19 Vaccine Hesitancy. *Journa of America Medical Association* **327**, 1013-5.

Austvoll-Dahlgren A, Helseth S (2010) What informs parents' decision-making about childhood vaccinations? *Journal of Advanced Nursing* **66**, 2421-30.

Ayers JW, Poliak A, Dredze M et al. (2023a) Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine* **183**, 589-96.

Ayers JW, Zhu Z, Poliak A et al. (2023b) Evaluating Artificial Intelligence Responses to Public Health Questions. *JAMA Network Open* **6**, e2317517.

Bernard A, Langille M, Hughes S, Rose C, Leddin D, van Zanten SV (2007) A Systematic Review of Patient Inflammatory Bowel Disease Information Resources on the World Wide Web. *American College of Gastroenterology* **102**, 2070-77

Bhardwaz S, Kumar J (2023) An Extensive Comparative Analysis of Chatbot Technologies - ChatGPT, Google BARD and Microsoft Bing. In *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pp. 673-9.

Bland JM, Altman DG (1997) Cronbach's alpha. *British Medical Journal* **314**, 572.

Cascella M, Montomoli J, Bellini V, Bignami E (2023) Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *Journal of Medical Systems* **47**, 33.

Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Doshi RH, Amin K, Khosla P, Bajaj S, Chheang S, Forman H (2023) Utilizing Large Language Models to Simplify Radiology Reports: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, Google Bard, and Microsoft Bing. *medRxiv*, 2023.06. 04.23290786.

Edwards A, Elwyn G, Smith C, Williams S, Thornton H (2001) Consumers' views of quality in the consultation and their relevance to 'shared decision-making' approaches. *Health Expectations* **4**, 151-61.

Jianfeng G, Michel G, Lihong L (2019) *Neural Approaches to Conversational AI: Question Answering, Task-oriented Dialogues and Social Chatbots*: now.

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* **521**, 436-44.

Liu Y, Han T, Ma S et al. (2023) Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*.

Lu Q, Qiu B, Ding L, Xie L, Tao D (2023) Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *arXiv preprint arXiv:2303.13809*.

Mehdi Y (2023) Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web. (https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/)

Milmo D (2023) ChatGPT reaches 100 million users two months after launch. The Guardian. The Guardian. (https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app)

Nosrat A, Dianat O, Verma P, Yu P, Wu D, Fouad AF (2022a) Endodontics Specialists' Practice during the Initial Outbreak of Coronavirus Disease 2019. *Journal of Endodontics* **48**, 102-8.

Nosrat A, Yu P, Dianat O et al. (2022b) Endodontic Specialists' Practice During the Coronavirus Disease 2019 Pandemic: 1 Year after the Initial Outbreak. *Journal of Endodontics* **48**, 699-706.

Pichai S (2023) An important next step on our AI journey. (https://blog.google/intl/en-africa/products/explore-get-answers/an-important-next-step-on-our-ai-journey/)

Rampa S, Veeratrishul A, Raimondo M, Connolly C, Allareddy V, Nalliah RP (2019) Hospital-based Emergency Department Visits with Periapical Abscess: Updated Estimates from 7 Years. *Journal of Endodontics* **45**, 250-6.

Safi Z, Abd-Alrazaq A, Khalifa M, Househ M (2020) Technical aspects of developing chatbots for medical applications: scoping review. *Journal of medical Internet research* **22**, e19127.

Sanmarchi F, Bucci A, Nuzzolese AG et al. (2023) A step-by-step researcher's guide to the use of an AI-based transformer in epidemiology: an exploratory analysis of ChatGPT using the STROBE checklist for observational studies. *Journal of Public Health (Berl)*. 1-36

Schmidhuber J (2015) Deep learning in neural networks: An overview. *Neural Networks* **61**, 85-117.

Suárez A, Díaz-Flores García V, Algar J, Gómez Sánchez M, Llorente de Pedro M, Freire Y (2023) Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers. *International Endodontic Journal*.

Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems* **27**.

Walker HL, Ghani S, Kuemmerli C et al. (2023) Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. *Journal of Medical Internet Research* **25**, e47479.

Xu L, Sanders L, Li K, Chow JCL (2021) Chatbot for Health Care and Oncology Applications Using Artificial Intelligence and Machine Learning: Systematic Review. *JMIR Cancer* **7**, e27850.

Supplemental Table 1. *List of all 180 scores after agreement between the two reviewers.*

Questions	GPT-3.5			Google Bard			Bing		
	A	B	C	A	B	C	A	B	C
1	5	5	5	5	5	4	5	5	5
2	5	5	5	5	4	5	5	5	4
3	5	5	5	5	5	5	4	4	5
4	5	4	5	5	4	4	5	5	5
5	5	4	5	4	4	4	5	4	5
6	5	5	5	5	5	5	2	2	2
7	5	5	5	5	5	4	4	4	4
8	5	5	5	4	5	4	5	4	4
9	4	5	5	4	5	4	5	5	5
10	4	5	4	4	4	4	4	4	4
11	5	4	5	4	5	3	3	3	3
12	5	5	5	4	4	4	4	4	4
13	3	3	4	3	3	4	3	3	3
14	5	5	5	5	4	5	4	3	3
15	5	4	4	4	4	4	3	3	3
16	5	5	5	4	4	4	4	4	4
17	5	4	4	5	5	5	4	4	4
18	5	5	5	3	4	3	4	4	4
19	5	5	5	4	4	4	4	4	4
20	5	5	5	4	5	4	4	4	4

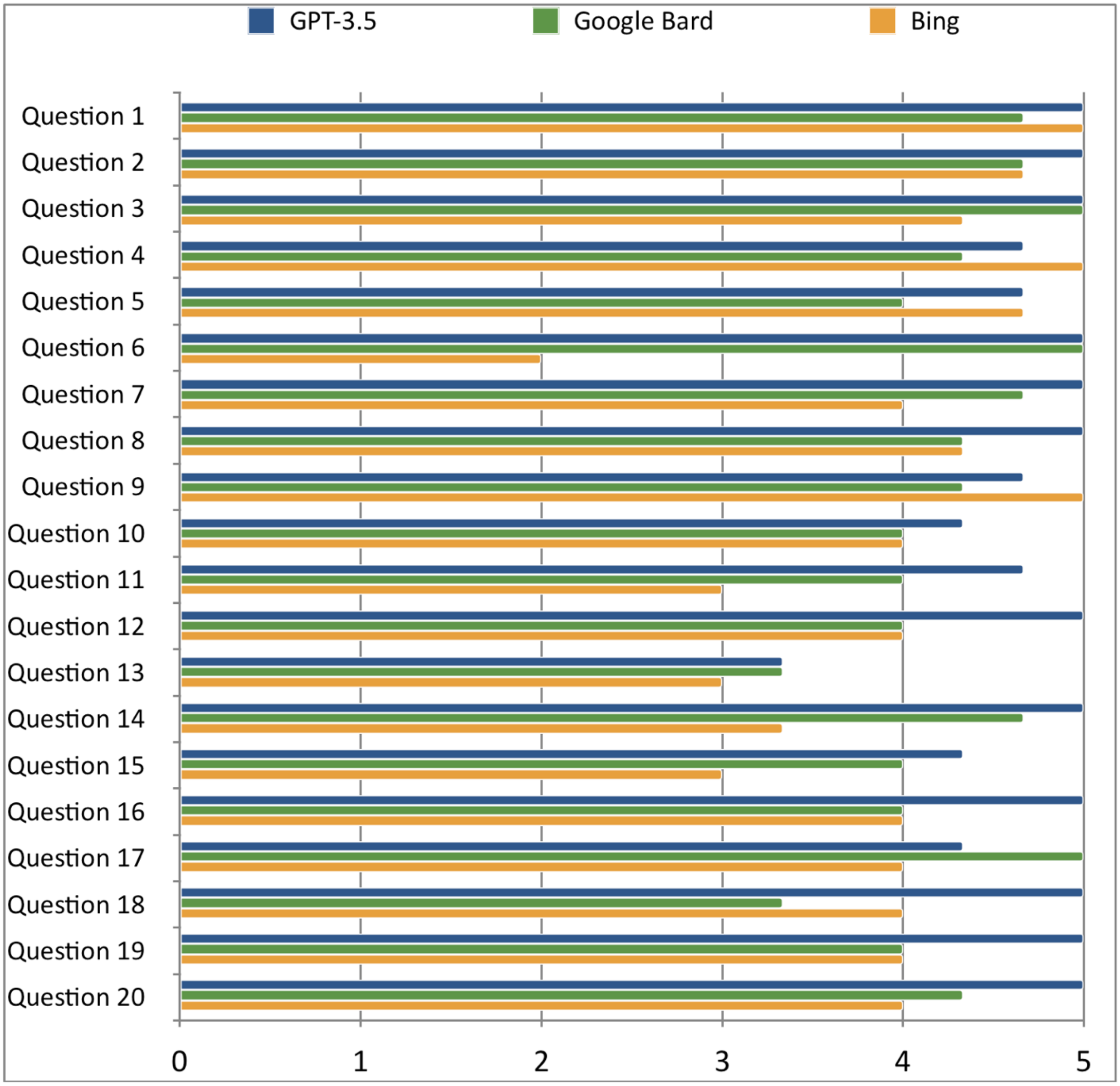

GPT-3.5
Google Bard
Bing
Question 1
Question 2
Question 3
Question 4
Question 5
Question 6
Question 7
Question 8
Question 9
Question 10
Question 11
Question 12
Question 13
Question 14
Question 15
Question 16
Question 17
Question 18
Question 19
Question 20
0
1
2
3
4
5

Figure 1

Figure 2

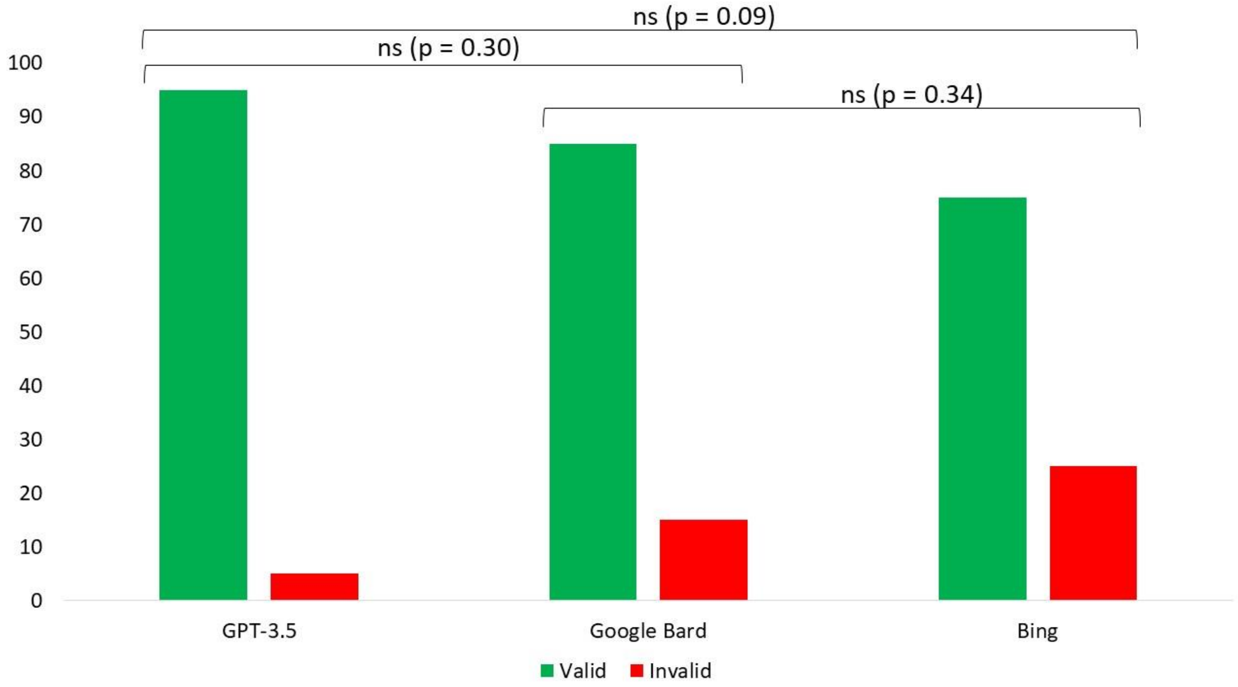
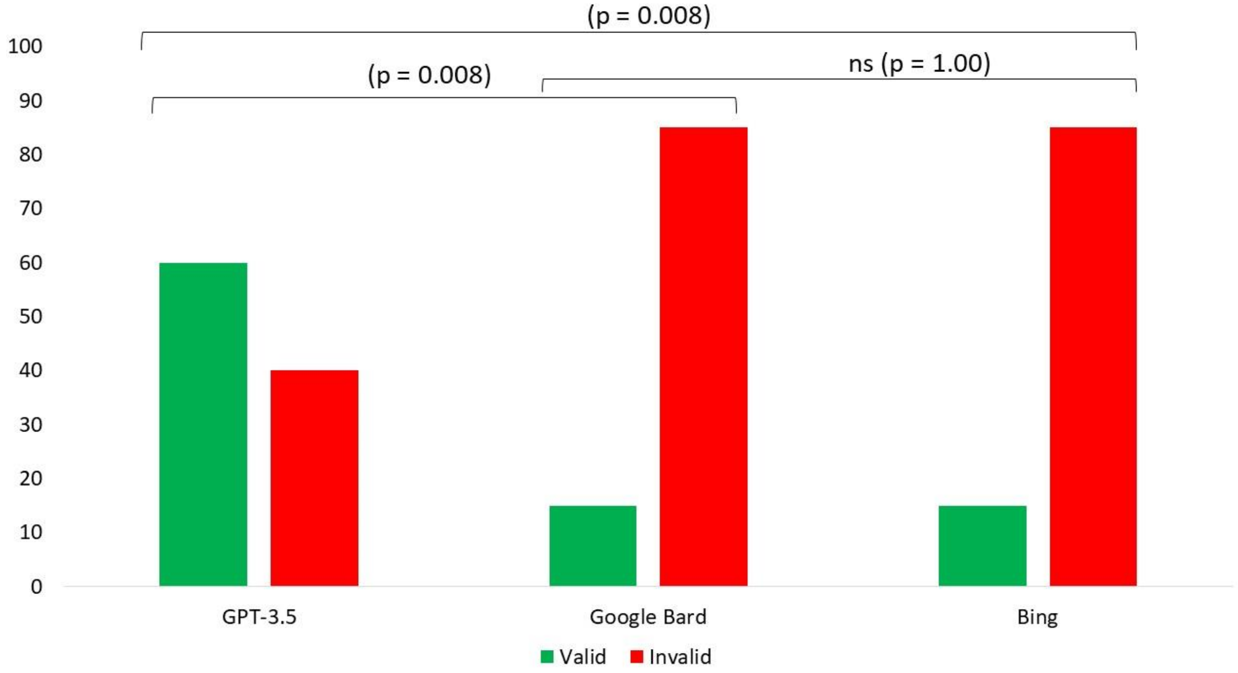

ns (p = 0.09)
ns (p = 0.30)
ns (p = 0.34)
100
90
80
70
60
50
40
30
20
10
0
GPT-3.5
Google Bard
Bing
🟩 Valid
🟥 Invalid

Figure 3


(p = 0.008)
(p = 0.008)
ns (p = 1.00)
100
90
80
70
60
50
40
30
20
10
0
GPT-3.5
Google Bard
Bing
🟩 Valid
🟥 Invalid