applied sciences

MDPI

Article

Artificial Intelligence Chatbots and Temporomandibular Disorders: A Comparative Content Analysis over One Year

Serena Incerti Parenti [1,*], Alessandro Maglioni [1], Elia Evangelisti [1], Antonio Luigi Tiberio Gracco [2], Giovanni Badiali [3], Giulio Alessandri-Bonetti [1] and Maria Lavinia Bartolucci [1]

[1] Section of Orthodontics and Dental Sleep Medicine, Department of Biomedical and Neuromotor Sciences (DIBINEM), University of Bologna, 40125 Bologna, Italy; alessandro.maglioni@unibo.it (A.M.); giulio.alessandri@unibo.it (G.A.-B.); maria.bartolucci3@unibo.it (M.L.B.)
[2] Department of Neuroscience, School of Dentistry, University of Padua, 35128 Padua, Italy
[3] Oral and Maxillofacial Surgery Unit, IRCCS Azienda Ospedaliero-Universitaria di Bologna, 40138 Bologna, Italy
* Correspondence: serena.incerti@unibo.it

Abstract

As the use of artificial intelligence (AI) chatbots for medical queries expands, their reliability may vary as models evolve. We longitudinally assessed the quality, reliability, and readability of information on temporomandibular disorders (TMD) generated by three widely used chatbots (ChatGPT, Gemini, and Microsoft Copilot). Ten TMD questions were submitted to each chatbot at two timepoints (T1: February 2024; T2: February 2025). Two blinded evaluators independently assessed all answers using validated tools like the Global Quality Score (GQS), PEMAT, DISCERN, CLEAR, Flesch Reading Ease (FRE), and Flesch–Kincaid Grade Level (FKGL) tools. Analyses followed METRICS guidance. Comparisons between models and across timepoints were conducted using non-parametric tests. At T1, Copilot scored significantly lower in GQS, CLEAR appropriateness, and relevance ($p < 0.01$), while ChatGPT provided less evidence-based content than its counterparts ($p < 0.001$). Reliability was poor across models (mean DISCERN score: 34.73 $\pm$ 9.49), and readability was difficult (mean FRE: 34.64; FKGL: 14.13). At T2, performances improved across chatbots, particularly for Copilot, yet actionability remained limited and citations were inconsistent. This year-long longitudinal analysis shows an overall improvement in chatbot performance, although concerns regarding information reliability persist. These findings underscore the importance of human oversight of AI-mediated patient information, reaffirming that clinicians should remain the primary source of patient education.

Keywords: temporomandibular disorders; patient education; artificial intelligence; orofacial pain; Internet use

Academic Editor: George Drosatos

Received: 25 October 2025
Revised: 19 November 2025
Accepted: 20 November 2025
Published: 24 November 2025

Citation: Incerti Parenti, S.; Maglioni, A.; Evangelisti, E.; Gracco, A.L.T.; Badiali, G.; Alessandri-Bonetti, G.; Bartolucci, M.L. Artificial Intelligence Chatbots and Temporomandibular Disorders: A Comparative Content Analysis over One Year. *Appl. Sci.* **2025**, *15*, 12441. https://doi.org/10.3390/app152312441

1. Introduction

Temporomandibular disorders (TMD) encompass a heterogeneous group of conditions affecting the temporomandibular joint, the masticatory muscles, and associated structures [1]. Mainly characterized by pain and dysfunction in the masticatory system [2], TMD represents the most prevalent source of chronic pain of non-dental origin in the orofacial region [3], with a global prevalence ranging from 5% to 12% [4]. Their etiology is multifactorial, taking into account both organic and psychosocial components [2]. Given the absence of a single causal factor and the natural fluctuation of symptoms over time,

TMD management should adhere to a stepped-care model, initially prioritizing conservative approaches such as cognitive-behavioral therapy, habit reversal, patient education, self-management, and physical therapy [3]. TMD represents the second most prevalent chronic musculoskeletal disorder [5], and individuals with chronic conditions are among the most frequent seekers of online health information [6]. This highlights the importance of evaluating the quality and accuracy of patient education on TMD. The increasing reliance on the Internet for medical information stems from its convenience and accessibility [7]. However, online sources vary widely, ranging from peer-reviewed scientific literature and authoritative medical websites to blogs and social media content, leading to a heterogeneous quality of available information [8–10]. A recent advancement in online patient education is the integration of artificial intelligence (AI)-driven chatbots, which leverage large language models (LLMs) to understand user queries and generate responses in a chat interface [11]. These systems have been trained on extensive multilingual datasets and are capable of producing human-like informational content [12,13]. AI-powered chatbots can be employed for various purposes, including information retrieval on specific topics, the generation of educational or entertainment content, customer support, and healthcare assistance [11,14].

From 2023 to 2025, major AI chatbots underwent multiple updates to underlying LLM architectures, affecting citation behavior, actionability, and hallucination risk. Because many updates are silent to end-users, the "same" chatbot can produce different medical advice across months, an instance of temporal instability commonly referred to as model drift, undermining the stability of patient education materials [15,16].

Previous studies have evaluated the quality of online patient education about TMD on websites and social media, reporting a poor quality of the content provided. The authors pointed out the need for greater control of information provided by digital sources in order to make the most of their educational potential [17–19]. Recent investigations evaluated chatbots' performance on bruxism [20] and TMD [21,22] at one time point, but to date, there are no studies that longitudinally assessed chatbot-generated TMD content across multiple platforms using a comprehensive analysis. The aim of this study was to determine if commonly used AI chatbots (ChatGPT, Gemini, and Copilot) show significant changes in the quality, reliability, and readability of TMD information over one year and whether changes are consistent across models, evaluating common TMD-related questions at two time points (February 2024 and February 2025). The null hypothesis is that there are no differences in performance between models and across timepoints.

2. Materials and Methods

This study followed the METRICS checklist for the design and reporting of AI research in healthcare [23]. The investigation was conducted in accordance with the Declaration of Helsinki and did not require approval from an ethics committee.

2.1. Models of AI Tested, Settings, Testing Time, and Duration

The AI chatbots were chosen based on their accessibility, popularity, and advanced language processing capabilities. The models evaluated included ChatGPT (OpenAI, San Francisco, CA, USA), Gemini (Google, Mountain View, CA, USA), and Microsoft Copilot (Microsoft Corporation, Redmond, WA, USA), with their most recent updates available at the time of testing in February 2024 (ChatGPT-3.5; Gemini Pro 1.0; Copilot GPT-4) and February 2025 (ChatGPT-4o; Gemini 1.5 Flash; Copilot GPT-4o). Testing was performed using each chatbot's default configuration to ensure replicability of the generated content.

The study involved submitting 10 distinct queries to each chatbot, for a total of 30 responses analyzed per time point, a number chosen a priori to allow a balance between

practical feasibility and comprehensive analysis of each query. The broad topic of the study was TMD, encompassing specific subtopics such as general knowledge and risk factors, diagnosis, treatment, and management.

To determine the queries, ChatGPT was directly prompted to generate a list of frequently searched questions about TMD to have a list of lay-level sets of questions. On 26 February 2024, a clear version of ChatGPT accessed via Google Chrome was queried with the following request: "Could you tell me what the ten most frequently asked questions on the internet about temporomandibular disorders are?" Two independent researchers with expertise in TMD and orofacial pain, independent of the raters, reviewed the items for neutrality, scope, and patient relevance before testing, confirming that all questions would be relevant to patient education domains [2]. The ten questions generated by the AI were subsequently used for this investigation (Table 1).

Table 1. Ten questions about TMD submitted to ChatGPT, Gemini, and Microsoft Copilot.

1	What are the symptoms of temporomandibular disorders?
2	How is temporomandibular disorder diagnosed?
3	What causes temporomandibular disorders?
4	Are there effective home remedies for TMD relief?
5	What are the available treatments for temporomandibular disorders?
6	Can stress and anxiety contribute to temporomandibular disorders?
7	What exercises can help with temporomandibular disorders?
8	Are there specific foods to avoid with temporomandibular disorders?
9	How long does it take to recover from temporomandibular disorders?
10	When should I see a doctor for temporomandibular disorders?

On 27 February 2024 (from 2:00 PM to 7:00 PM, Bologna, Italy, Central European Time), a new account was created for each chatbot to ensure an incognito testing environment. Access to chatbots was made through Google Chrome (consumer build) on Windows 11 Home, with dedicated accounts on each platform operating under the free plan and default consumer safety guardrails. Dedicated email addresses, generated specifically for this study, were used for account creation. Fresh instances of ChatGPT, Gemini, and Copilot were then queried with the ten selected questions, followed by the prompt: "Please provide references." To minimize the influence of prior responses, each question was submitted in a new chat window, and the option "regenerate response" was not used. Additionally, after each query, the search history was manually deleted, along with clearing the browser history and cache. If a chatbot generated multiple response options for a given question, requesting the user to select the most appropriate one, the most comprehensive answer was chosen. When applicable, the "more balanced" conversation style was selected. Browsing, retrieval, and plugins were off or not available in the consumer UIs. Temperature and related decoding parameters were not user-exposed at either timepoint; to limit variance attributable to unknown sampling settings, we analyzed the first response verbatim for each model and timepoint. All questions were presented in English, and the resulting responses (T1) and references were recorded and anonymized for subsequent analysis.

After nearly one year, on 11 February 2025 (from 2:00 PM to 7:00 PM, Bologna, Italy, Central European Time), the same protocol was applied. The resulting responses were anonymized as T2.

The chatbot-generated answers were subsequently analyzed for quality, reliability, and readability. Two independent researchers with expertise in TMD and orofacial pain (A.M. and E.E.), blinded to the information source, conducted the analysis. Any discrepancies were resolved through discussion with a third researcher (M.L.B.) until consensus was reached. As a descriptive measure of inter-rater agreement before consensus, we calculated

the percentage of ratings in which the two examiners assigned identical scores, which was 78% across all instruments and items.

2.2. Quality and Reliability Assessment

The evaluation of the quality of the AI-generated content was based on the modified Global Quality Score (GQS) [24] and on the Patient Education Materials Assessment Tool (PEMAT) [25]. GQS is a five-point Likert scale introduced by Bernard et al. in 2007 to subjectively determine the overall quality, flow, and possible benefit for patients of websites [24], with higher scores meaning better quality. PEMAT is instead an instrument developed with the purpose of evaluating the understandability and actionability of patient education materials [25]. Two versions of PEMAT have been produced, one for printable materials (PEMAT-P) and one for audiovisual materials (PEMAT-A/V); the former was used in this study since AI chatbots did not provide any audiovisual information. PEMAT-P consists of 24 items with answer options "Agree" (1 point), "Disagree" (0 points), and "Not Applicable." Percentage scoring of both understandability and actionability could then be computed, with higher scoring corresponding to higher quality values.

The reliability of the information was assessed using the DISCERN instrument. DISCERN is a validated tool with the purpose of judging the quality of written information on treatment choices for a health problem [26]. DISCERN incorporates three sections with a total of 16 items, to which the researcher can give a score ranging from 1 to 5 according to how much they agree with the questions' statement. The first section aims to determine whether the publication is reliable, the second is about the quality of information on treatment choices, and the last section is only one query that indicates the overall rating of the material. The total DISCERN score could be calculated with the sum of all 16 questions, and its value ranges from 16 to 80, with higher scores meaning a more reliable publication and better indications about treatment options. The DISCERN total score can be better interpreted through the following categories: "very poor" for scores 16–26, "poor" for 27–38, "fair" for 39–50, "good" for 51–62, and "excellent" for 63–80.

2.3. Content Assessment

CLEAR is a tool specifically tailored to assess the quality of health-related content generated by AI-based models [27]. It is composed of five items: (1) Completeness; (2) Lack of false information; (3) Evidence; (4) Appropriateness; and (5) Relevance. Each item is scored based on a five-point Likert scale from excellent to poor. For descriptive interpretation of the CLEAR items as an indication of the quality of the generated content, the scores were classified into the following categories: CLEAR scores of 1–1.79 were classified as "poor"; 1.80–2.59 as "satisfactory"; 2.60–3.39 as "good"; 3.40–4.19 as "very good"; and 4.20–5.00 as "excellent" [27]. The range of CLEAR total scores was 5–25, divided into three categories: 5–11 categorized as "poor" content, 12–18 categorized as "average" content, and 19–25 categorized as "very good" content [27].

2.4. Readability Assessment

To analyze the readability of each answer, the Flesch Reading Ease score (FRE) [28], the Flesch–Kincaid Grade Level score (FKGL) [29], and the word count were used. The FRE rates on a 100-point scale how easy it is to understand a text. Lower scores indicate a higher difficulty reading level. The FKGL assesses the educational level necessary to comprehend the complexity of the text, with a higher grade corresponding to more difficult content [30]. The longer the text, in terms of word count, the less efficient it was considered, since the reader could be discouraged from entirely reading a long answer.

2.5. Statistical and Data Analysis

The normality of data was assessed with the Kolmogorov–Smirnov test. We prespecified four families based on the underlying construct: (1) quality of patient information (GQS, PEMAT Understandability, CLEAR total); (2) reliability of information (DISCERN Reliability, CLEAR Lack of false information, CLEAR Evidence, DISCERN Total); (3) readability and length (FRE, FKGL, word count); and (4) contextual quality of content (CLEAR Appropriateness, CLEAR Completeness, CLEAR Relevance). PEMAT Actionability and DISCERN Treatment Choice were treated as standalone endpoints and were therefore not included in a family. For each family, and separately for each comparison type (between-chatbot comparisons at T1, between-chatbot comparisons at T2, and within-chatbot longitudinal comparisons T1 vs. T2), we controlled the family-wise error rate using a Holm-Bonferroni correction across all endpoints in that family.

To compare the GQS, CLEAR, DISCERN, and readability scores between the groups at T1 and T2, the Kruskal–Wallis test, followed by the Mann–Whitney post-hoc test with Bonferroni correction, was used. The $\chi 2$ test was employed to analyze the PEMAT scores.

To compare the scores between T1 and T2, the Wilcoxon Signed Ranks Test was used. Statistical analyses were performed using the statistical software SPSS for Windows (version 18.0; 2009; SPSS, IBM, Armonk, NY, USA). The limit for statistical significance was set at $p < 0.05$.

3. Results

The dataset comprised 30 responses at each time point: T1 (27 February 2024) and T2 (11 February 2025), with 10 responses per chatbot (ChatGPT, Gemini, and Copilot).

Inter-rater agreement before consensus was high, with the two examiners assigning identical scores in 78% of all ratings across instruments.

At T1, the GQS indicated high content quality for ChatGPT (mean, 95% Confidence Interval, CI, 4.2, 3.9–4.5) and Gemini (4.4, 3.9–4.9), with well-structured responses covering most relevant information. Conversely, Copilot scored significantly lower (3.0, 2.3–3.7), reflecting moderate quality, suboptimal flow, and insufficiently discussed content. This difference was statistically significant ($p < 0.01$) (Figure 1).

Regarding CLEAR criteria, Copilot performed significantly worse than Gemini and ChatGPT in "Appropriateness" (Copilot: 3.4, 2.6–4.2; Gemini: 4.9, 4.5–5.0; ChatGPT: 4.7, 4.4–5.0; $p < 0.001$) and "Relevance" (Copilot: 3.2, 2.3–4.1; Gemini: 4.9, 4.4–5.0; ChatGPT: 4.8, 4.5–5.0; $p < 0.001$). Post-hoc analysis confirmed that Copilot had a significantly lower CLEAR total score than Gemini (Copilot: 17.1, 14.3–19.9; Gemini: 20.5, 19.1–21.9; $p = 0.008$), while ChatGPT showed a non-significant difference compared with Gemini (ChatGPT: 20.2, 18.6–21.8; $p = 0.076$). Notably, ChatGPT had a significantly lower "Evidence" score than both Gemini and Copilot (ChatGPT: 2.4, 1.6–3.2; Gemini: 2.8, 2.5–3.1; Copilot: 3.1, 2.3–3.9; $p < 0.001$) (Figure 2).

No significant differences were found among chatbots in PEMAT-P understandability, actionability, reliability (DISCERN 1–8), quality of treatment-related information (DISCERN 9–15), DISCERN total score, or readability metrics (FRE, FKGL, and word count) at T1 (Table 2).

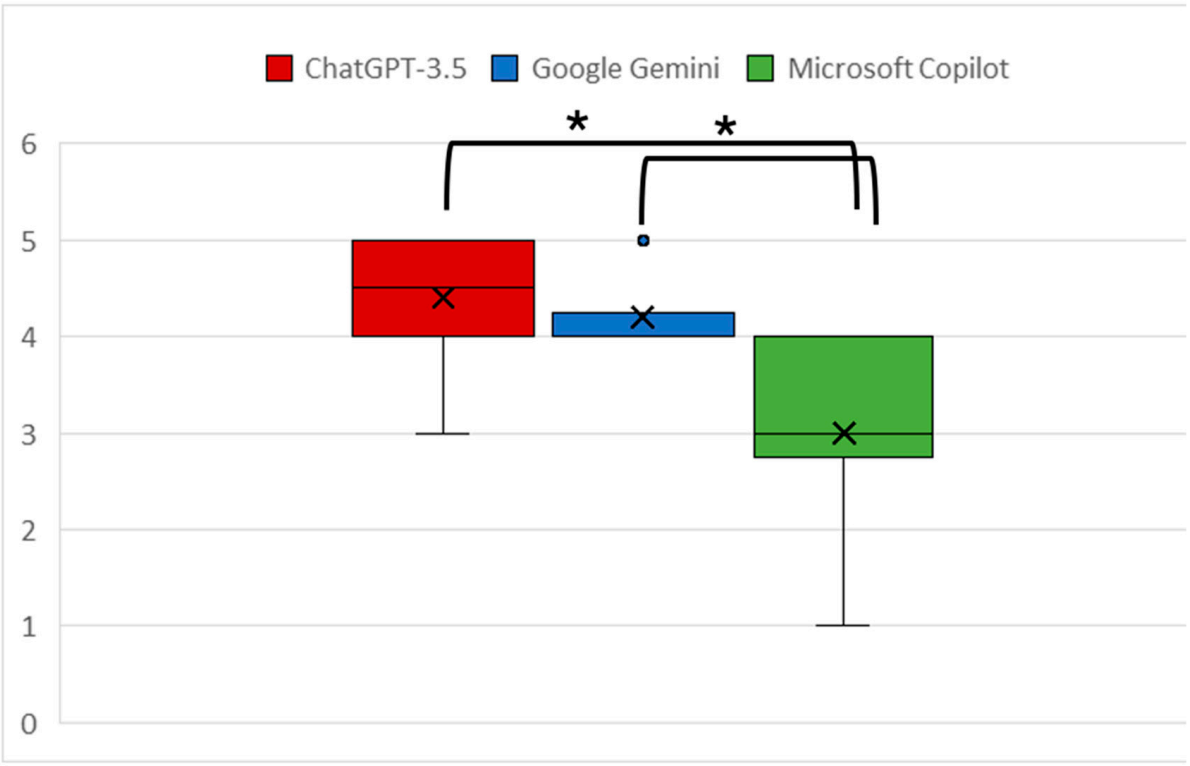

Figure 1. T1 GQS scores for the AI chatbots' answers. Boxes represent the interquartile range, horizontal lines the median, whiskers the minimum and maximum values, circles individual scores, and crosses the mean. *: $p < 0.01$.

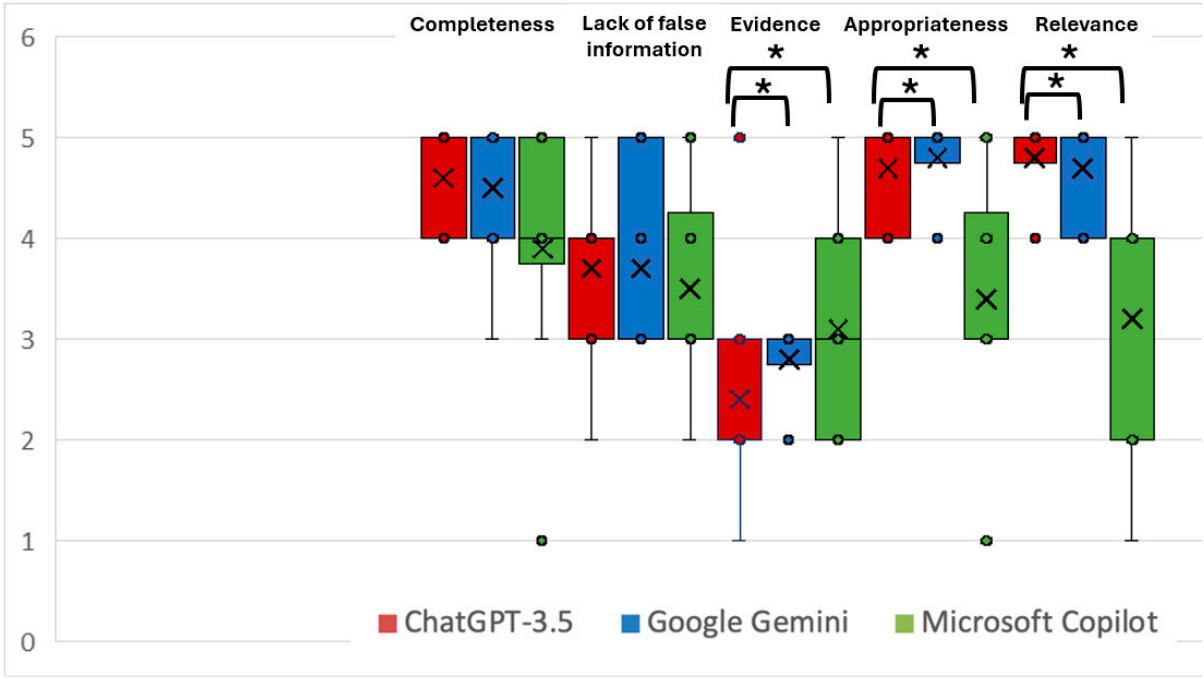

Figure 2. T1 CLEAR scores for the AI chatbots' answers. Boxes represent the interquartile range, horizontal lines the median, whiskers the minimum and maximum values, circles individual scores, and crosses the mean. *: $p < 0.001$.

Table 2. Assessment in T1 of quality (PEMAT), reliability (DISCERN), and readability (FRE, FKGL, and word count) scores for the AI chatbots' answers. *p*-values are Holm–Bonferroni corrected for multiple comparisons within the prespecified family of endpoints.

Assessment Variables	ChatGPT Mean (95% CI)	Google Gemini Mean (95% CI)	Microsoft Copilot Mean (95% CI)	*p*-Value
PEMAT Understandability	60.3 (53.7–66.8)	73.2 (68.5–77.9)	61 (51.7–70.3)	0.070
PEMAT Actionability	48.0 (38.0–58.0)	50.0 (36.1–63.9)	42.0 (31.4–52.6)	0.392
DISCERN Reliability	23.2 (21.7–24.7)	25.2 (24.3–26.1)	26.6 (23.1–30.1)	0.694
DISCERN Treatment choice	15.2 (9.1–21.3)	15.8 (10.3–21.3)	13.6 (11.3–15-9)	0.958
DISCERN Total	32.0 (24.4–39.6)	34.4 (27.1–41.7)	34.3 (28.7–39.9)	0.850
Flesch Reading Ease Score (FRE)	28.5 (16.3–40.8)	41.8 (32.8–50.8)	33.5 (23.9–43.2)	0.210
Flesch–Kincaid Grade Level Score (FKGL)	15.7 (12.6–18.8)	11.9 (10.5–13.4)	14.7 (12.6–16.8)	0.159
word count	320.8 (273.8–367.8)	282.7 (251.0–314.4)	258.6 (194.9–322.3)	0.210

At T2, previously observed differences disappeared, with no statistically significant variations across chatbots in any assessed parameter (Table 3).

Table 3. Assessment in T2 of quality (PEMAT), reliability (DISCERN), and readability (FRE, FKGL, and word count) scores for the AI chatbots' answers. *p*-values are Holm–Bonferroni corrected for multiple comparisons within the prespecified family of endpoints.

Assessment Variables	ChatGPT Mean (95% CI)	Google Gemini Mean (95% CI)	Microsoft Copilot Mean (95% CI)	*p*-Value
Global Quality Scale	4.3 (3.8–4.8)	4.2 (3.7–4.7)	4.1 (3.7–4.5)	0.879
CLEAR Completeness	4.6 (4.2–5.0)	4.3 (3.8–4.8)	4.1 (3.7–4.5)	0.174
CLEAR Lack of false information	3.9 (3.4–4.4)	4.0 (3.5–4.5)	4.2 (3.6–4.8)	0.631
CLEAR Evidence	3.5 (2.6–4.4)	3.0 (2.2–3.8)	3.4 (2.6–4.2)	0.677
CLEAR Appropriateness	4.9 (4.8–5.0)	4.7 (4.4–5.0)	4.8 (4.5–5.0)	0.197
CLEAR Relevance	4.6 (4.1–5.0)	4.2 (3.5–4.9)	4.6 (4.2–5.0)	0.501
CLEAR Total Score	21.6 (19.8–23.4)	20.2 (18.5–21.9)	21.1 (19.8–22.4)	0.879
PEMAT Understandability	68.3 (61.0–75.7)	63.3 (57.6–69.1)	60.8 (53.9–67.7)	0.879
PEMAT Actionability	26.0 (8.1–43.9)	36.0 (19.8–52.2)	38.0 (22.3–53.7)	0.460
DISCERN Reliability	25.1 (22.3–27.9)	24.1 (22.1–26.1)	25.4 (22.8–28.0)	0.835
DISCERN Treatment choice	16.2 (9.7–22.7)	19.2 (14.0–24.4)	15.0 (10.3–19.7)	0.361
DISCERN Total	34.7 (26.2–43.2)	35.4 (26.6–44.2)	34.0 (27.8–40.2)	0.898
Flesch Reading Ease Score (FRE)	46.8 (37.5–56.0)	53.8 (42.4–65.2)	45.9 (34.0–57.7)	0.292
Flesch–Kincaid Grade Level Score (FKGL)	9.9 (8.1–11.6)	9.0 (7.1–11.0)	10.5 (8.4–12.5)	0.336
Word count	276.9 (181.4–372.4)	234.2 (160.6–307.8)	180.3 (120.6–240.0)	0.168

When comparing T1 and T2, Copilot demonstrated significant improvements in 2025 across multiple metrics, including GQS, CLEAR total score, and readability (FRE, FKGL). Specific enhancements were noted in CLEAR subcategories "Lack of false information," "Appropriateness," and "Relevance" (Table 4). ChatGPT improved in "Evidence" and readability measures (FRE, FKGL) but declined in PEMAT actionability (Table 4). Gemini also showed better readability scores (FRE, FKGL) in 2025 but exhibited lower PEMAT understandability and actionability (Table 4).

Table 4. Comparison between T1 and T2 scores for the AI chatbots' answers. *p*-values are Holm–Bonferroni corrected for multiple comparisons within the prespecified family of endpoints. *: $p < 0.05$.

	ChatGPT Mean (95% CI)		*p*-Value	Google Gemini Mean (95% CI)		*p*-Value	Microsoft Copilot Mean (95% CI)		*p*-Value
Assessment Variables	T1	T2		T1	T2		T1	T2	
Global Quality Scale	4.2 (3.9–4.5)	4.3 (3.8–4.8)	0.705	4.4 (3.9–4.9)	4.2 (3.7–4.7)	0.960	3.0 (2.3–3.7)	4.1 (3.7–4.5)	0.030 *
CLEAR Completeness	4.6 (4.2–5.0)	4.6 (4.2–5.0)	1.00	4.5 (4.0–5.0)	4.3 (3.8–4.8)	0.634	3.9 (3.0–4.8)	4.1 (3.7–4.5)	0.739
CLEAR Lack of false information	3.7 (3.1–4.3)	3.9 (3.4–4.4)	0.414	3.7 (3.0–4.4)	4.0 (3.5–4.5)	0.720	3.5 (2.8–4.2)	4.2 (3.6–4.8)	0.152
CLEAR Evidence	2.4 (1.6–3.2)	3.5 (2.6–4.4)	0.172	2.8 (2.5–3.1)	3.0 (2.2–3.8)	0.720	3.1 (2.3–3.9)	3.4 (2.6–4.2)	1.00
CLEAR Appropriateness	4.7 (4.4–5.0)	4.9 (4.8–5.0)	0.249	4.8 (4.5–5.0)	4.7 (4.4–5.0)	0.634	3.4 (2.6–4.2)	4.8 (4.5–5.0)	0.042 *
CLEAR Relevance	4.8 (4.5–5.0)	4.6 (4.1–5.0)	0.828	4.7 (4.4–5.0)	4.2 (3.5–4.9)	0.393	3.2 (2.3–4.1)	4.6 (4.2–5.0)	0.042 *
CLEAR Total Score	20.2 (18.6–21.8)	21.6 (19.8–23.4)	0.342	20.5 (19.1–21.9)	20.2 (18.5–21.9)	1.00	17.1 (14.3–19.9)	21.1 (19.8–22.4)	0.024 *
PEMAT Understandability	60.3 (53.7–66.8)	68.3 (61.0–75.7)	0.342	73.2 (68.5–77.9)	63.3 (57.6–69.1)	0.033 *	61.0 (51.7–70.3)	60.8 (53.9–67.7)	1.00
PEMAT Actionability	48.0 (38.0–58.0)	26.0 (8.1–43.9)	0.035 *	50.0 (36.1–63.9)	36.0 (19.8–52.2)	0.038 *	42.0 (31.4–52.6)	38.0 (22.3–53.7)	0.671
DISCERN Reliability	23.2 (21.7–24.7)	25.1 (22.3–27.9)	0.411	25.2 (24.3–26.1)	24.1 (22.1–26.1)	0.720	26.6 (23.1–30.1)	25.4 (22.8–28.0)	0.918
DISCERN Treatment choice	15.2 (9.1–21.3)	16.2 (9.7–22.7)	0.408	15.8 (10.3–21.3)	19.20 (14.0–24.4)	0.054	13.6 (11.3–15.9)	15.0 (10.3–19.7)	0.279
DISCERN Total	32.0 (24.4–39.6)	34.7 (26.2–43.2)	0.411	34.4 (27.1–41.7)	35.4 (26.6–44.2)	0.720	34.3 (28.7–39.9)	34.0 (27.8–40.2)	1.00
Flesch Reading Ease Score (FRE)	28.5 (16.3–40.8)	46.8 (37.5–56.0)	0.027 *	41.8 (32.8–50.8)	53.8 (42.4–65.2)	0.056	33.6 (23.9–43.2)	45.9 (34.0–57.7)	0.026 *
Flesch–Kincaid Grade Level Score (FKGL)	15.7 (12.6–18.8)	9.9 (8.1–11.6)	0.027 *	11.9 (10.5–13.4)	9.0 (7.1–11.0)	0.051	14.7 (12.6–16.8)	10.5 (8.4–12.5)	0.021 *
Word count	320.8 (274–368)	276.9 (181.4–372.4)	0.241	282.7 (251.0–314.4)	234.2 (160.6–307.8)	0.139	258.6 (120.6–240.0)	180.3 (120.6–240.0)	0.074

4. Discussion

The present investigation evaluated and compared the quality, reliability, and readability of the responses to commonly asked questions on TMD provided by three AI chatbots at two time points (February 2024 and February 2025), aiming to examine possible performance differences over time. The output generated by AI-based chatbots was already evaluated in many medical and dental domains [31,32]. Recent studies evaluated AI chatbots' expertise in providing information on specific TMD subtypes [21,22], but this is the first study to add a longitudinal perspective, identifying significant improvements in chatbots' performance over a year and persistent weaknesses, such as the lack of evidence-based recommendations. Additionally, a comprehensive evaluation framework was employed, including CLEAR and PEMAT scores, offering a broad assessment of content understandability, readability, and actionability, but to our knowledge, this is the first study evaluating the quality, reliability, and readability of information provided on TMD. The use of the METRICS checklist to standardize the reporting of generative AI-based studies in healthcare education and practice is also innovative. To prevent algorithms from generating answers based on the user's background, the questions were asked from newly created accounts for each chatbot, new conversation windows were opened for each question, and the browsing history was erased each time. This is also essential to avoid falling into "rabbit holes" or "echo chamber," situations where the LLM starts providing tangential information that leads away from the original topic or amplifies and reinforces the participant's pre-existing beliefs, thanks to repetition and algorithm biases.

The main outcome that emerges from the present investigation is that AI chatbots could provide a good overview of TMD.

However, after performing an accurate analysis of the message given, a series of critical issues related to reliability and usability emerged.

At T1, despite positive scores in "Completeness" and "Relevance," the items "Lack of false information" and "Evidence" obtained lower scores, suggesting that some answers presented misleading information, potentially harmful for patients, and that most of the content generated by chatbots was poorly evidence-based and with scarce bibliographic references. Moreover, no AI chatbot generated sufficient information about the modality of each treatment or mentioned enough about the possible risks or side effects of the treatments, thus failing to provide patients with the proper information to make an informed and unbiased treatment choice. Regarding readability, mean FRE and FKGL scores indicated that the answers were quite difficult to understand for laypeople, considering that most chatbots' answers were understandable for an individual with a school level above "high school."

The findings at T1 seem to suggest that AI models may not yet have reached a level of sophistication necessary for handling complex queries.

At T2, the assessment indicated an overall good quality of responses, with a more stable answer consistency. Notable improvements were observed in the description of therapeutic choices and the inclusion of scientific references. A comparison between T1 and T2 responses from Copilot revealed a significant improvement across most previously lacking aspects identified in 2024, ultimately achieving comparable performance to the other chatbots in 2025. These patterns may reflect platform-level updates, although this remains a plausible hypothesis rather than a confirmed causal explanation. For example, Microsoft's integration of GPT-4o into Copilot during 2024 could have contributed to changes in output characteristics, including decoding and safety behaviors. Notably, all three chatbots demonstrated a significant trend in improvement in readability, with scores corresponding to a "high school" level of comprehension. Otherwise, a significant reduction in actionability was registered for both ChatGPT and Gemini. These findings are noteworthy and might indicate a possible trend in chatbot responses toward limiting therapeutic indications in favor of enhancing descriptive content. The higher readability scores observed at T2 further support this hypothesis. Technically, several plausible explanations may account for the shifts observed from T1 to T2: model upgrades (architecture and context window) that improve overviews and readability; citation pipelines that could strengthen evidence mentions [33,34]; and safety policies that can reduce actionable "how-to" guidance in favor of risk-averse language [35], consistent with the T2 drop in actionability.

From a methodological standpoint, our findings support the notion—also suggested by prior literature—that point-in-time chatbot evaluations may be fragile, possibly due to the evolving nature of chatbot model identity [15]. These mechanisms have been described in prior literature, but in this study, they remain indirect inferences rather than confirmed drivers of the observed patterns.

Moreover, we observed that, at T1, AI chatbots provided reasonably accurate and helpful responses to the broadest questions concerning TMD symptoms and TMD general knowledge but did not perform as well when it came to more advanced technical topics, such as the TMD diagnosis process and proper treatment procedures. As an example, the order of possible approaches to TMD was subverted by chatbots that, instead of prioritizing evidence-based conservative therapies, first described temporomandibular joint (TMJ) surgery and frequently did not list its side effects and potential complications [36]. An evidence-based approach for the treatment of TMD supports a stepped care model with counselling and behavioral therapy as first-line interventions, aiming to change

maladaptive behaviors. Jaw exercises and a properly designed occlusal splint can also be supportive, combined, for short periods, with anti-inflammatory drugs. These are considered conservative therapies, since they aim at avoiding joint or muscle overload without producing irreversible changes to the masticatory system [37,38]. In February 2025, chatbots have become more cautious in suggesting temporomandibular joint surgery, placing it after conservative therapies, emphasizing its appropriateness only in specific and severe cases. However, there was still a lack of significant detail on potential risks and side effects.

A persistent issue we observed at both T1 and T2 was the indication of orthodontic therapy and minor bite adjustments among the therapeutic options for TMD. Specifically, the chatbots reported that "An improper bite, such as a deep bite, underbite, or crossbite, can put strain on the jaw joint and contribute to pain" and "Conditions such as a deep bite, underbite, or crossbite" could cause TMD onset. Current evidence does not support the presence of a correlation between dental occlusion and TMD: occlusal adjustments showed no to low effect on muscle pain reduction and were even harmful in patients with arthralgia and disocclusion [37,38]. Therefore, irreversible interventions such as selective grinding or building up of teeth in order to improve occlusal stability should be discouraged [37–39].

A sensible improvement in the quality of references was registered over time, which could be related to updates in the models' training datasets, although this remains speculative [40]. Nevertheless, it is crucial to inform patients about LLM limitations, allocating sufficient time to explain the evidence supporting a stepped-care model for TMD [41] and emphasizing the need to critically evaluate the sources of information.

Even in instances where chatbots provide accurate and reliable responses, another concern remains: the lack of easily accessible references in the current versions of freely available chatbots. It is often unclear where the information is retrieved, and it is not uncommon to encounter "artificial hallucinations" [42], which are responses containing false or misleading information presented as a fact, mainly due to reference divergence [12]. The present outcomes are broadly consistent with the findings of Hassan et al., who found that all LLMs displayed moderate reliability and lacked robust evidence-based content [21]. Similarly, Kula et al. analyzed ChatGPT-4's performance on specific TMD subtypes, reporting moderate-to-high reliability and usefulness but finding variable accuracy depending on the disorder [22]. Both authors conclude that the information being assessed does not meet the required standards of quality on an ongoing basis and therefore requires verification. The continuous evolution of LLMs in recent times is remarkable, even within highly specialized fields such as orofacial pain. However, health disinformation remains a significant public health concern, as recognized by the World Health Organization in 2022 [43], and it is imperative to enhance the accuracy and reliability of chatbot responses, strengthening their reliance on evidence-based references and verified medical databases, considering the persistent deficiencies in citation and treatment choice and their ethical implications [44]. Sustaining quality over time also requires attention to informational sustainability [45]. In fact training on AI-generated content can degrade model fidelity, indicating the need for curated, citable, human-authored resources and continuous evaluation pipelines [46,47].

In the present study, the dataset was limited to ten queries that lacked standardization and academic referencing. Moreover, because the prompt set was originally seeded by ChatGPT, there is a possibility that the content was biased toward one model's priors. Future studies should employ expert-generated or consensus-based prompts to minimize this risk and enhance methodological robustness. Even if the evaluation tools are standardized, the subjectivity in assigning the scores might be a potential bias. There is a possibility that the AI chatbots have improved their content since this investigation, so the present results might not be fully representative of their potential.

The widespread availability of health information is a cornerstone of patient-centered healthcare and empowers patients to make informed decisions about their health [48,49]. It is crucial to acknowledge that the medical suggestions provided by chatbots should be considered as general advice or informational content rather than a substitute for professional medical consultation and guidance.

5. Conclusions

Over the last year, chatbot performance on TMD information improved, particularly in readability, yet evidence quality and actionability remained inconsistent. Chatbot outputs should therefore be treated as adjuncts under clinical supervision, with routine factual verification [50]. Clinicians should therefore remain the primary source of patient education, providing evidence-based knowledge, counselling, and guidance, while also guiding patients in critically evaluating online resources. In these terms, chatbot outputs could be valuable as a positive reinforcement of the information obtained from the health professional [51]. Looking ahead, technological and scientific sustainability will depend on source traceability, multilingual inclusion, and equity [45,52]. In this regard, the role of academic institutions in disseminating high-quality content on the web could help to appropriately train the chatbots and provide tools for the conscious use of AI. Under these conditions, chatbots may support medical education and patient counselling as supervised, verifiable aids.

Author Contributions: Conceptualization, S.I.P., G.A.-B., and M.L.B.; methodology, S.I.P., A.L.T.G., G.B., and M.L.B.; formal analysis, S.I.P.; investigation, A.M., E.E., and M.L.B.; writing—original draft preparation, A.M.; writing—review and editing, M.L.B.; supervision, A.L.T.G., G.B., and G.A.-B.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
TMD	Temporomandibular Disorders
GQS	Global Quality Score
PEMAT	Patient Education Materials Assessment Tool
CLEAR	Completeness, Lack of false information, Evidence, Appropriateness, Relevance
FRE	Flesch Reading Ease
FKGL	Flesch–Kincaid Grade Level
LLM	Large Language Model
TMJ	Temporomandibular Joint
DC/TMD	Diagnostic Criteria for Temporomandibular Disorders

References

1. Okeson, J.; De Kanter, R. Temporomandibular disorders in the medical practice. *J. Fam. Pract.* **1996**, *43*, 347–356.
2. Schiffman, E.; Ohrbach, R.; Truelove, E.; Look, J.; Anderson, G.; Goulet, J.-P.; List, T.; Svensson, P.; Gonzalez, Y.; Lobbezoo, F.; et al. Diagnostic Criteria for Temporomandibular Disorders (DC/TMD) for Clinical and Research Applications: Recommendations of the International RDC/TMD Consortium Network and Orofacial Pain Special Interest Group. *J. Oral Facial Pain Headache* **2014**, *28*, 6–27. [CrossRef]

3. List, T.; Jensen, R.H. Temporomandibular disorders: Old ideas and new concepts. *Cephalalgia* **2017**, *37*, 692–704. [CrossRef] [PubMed]
4. Valesan, L.F.; Da-Cas, C.D.; Réus, J.C.; Denardin, A.C.S.; Garanhani, R.R.; Bonotto, D.; Januzzi, E.; Mendes de Souza, B.D. Prevalence of temporomandibular joint disorders: A systematic review and meta-analysis. *Clin. Oral Investig.* **2021**, *25*, 441–453. [CrossRef] [PubMed]
5. Yao, L.; Sadeghirad, B.; Li, M.; Li, J.; Wang, Q.; Crandon, H.N.; Martin, G.; Morgan, R.; Florez, I.D.; Hunskaar, B.S.; et al. Management of chronic pain secondary to temporomandibular disorders: A systematic review and network meta-analysis of randomised trials. *BMJ* **2023**, *383*, e076226. [CrossRef]
6. Bundorf, M.K.; Wagner, T.H.; Singer, S.J.; Baker, L.C. Who searches the internet for health information? *Health Serv. Res.* **2006**, *41*, 819–836. [CrossRef] [PubMed]
7. Marton, C.; Choo, C.W. A review of theoretical models of health information seeking on the web, J. A review of theoretical models of health information seeking on the web. *J. Doc.* **2012**, *68*, 330–352. [CrossRef]
8. Doyle, D.J.; Ruskin, K.J.; Engel, T.P. The Internet and medicine: Past, present, and future. The Internet and medicine: Past, present, and future. *Yale J. Biol. Med.* **1996**, *69*, 429–437.
9. Jia, X.; Pang, Y.; Liu, L.S. Online Health Information Seeking Behavior: A Systematic Review. *Healthcare* **2021**, *9*, 1740. [CrossRef]
10. Tan, S.S.L.; Goonawardene, N. Internet Health Information Seeking and the Patient-Physician Relationship: A Systematic Review. *J. Med. Internet Res.* **2017**, *19*, e9. [CrossRef]
11. Aldhafeeri, L.; Aljumah, F.; Thabyan, F.; Alabbad, M.; AlShahrani, S.; Alanazi, F.; Al-Nafjan, A. Generative AI Chatbots Across Domains: A Systematic Review. *Appl. Sci.* **2025**, *15*, 11220. [CrossRef]
12. Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* **2023**, *11*, 887. [CrossRef]
13. Lee, P.; Bubeck, S.; Petro, J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N. Engl. J. Med.* **2023**, *388*, 1233–1239. [CrossRef]
14. Verma, S.; Sharma, R.; Deb, S.; Maitra, D. Artificial intelligence in marketing: Systematic review and future research direction. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100002. [CrossRef]
15. Chen, L.; Zaharia, M.; Zou, J. How is ChatGPT's behavior changing over time? *arXiv* **2023**, arXiv:2307.09009. [CrossRef]
16. Gupta, M.; Virostko, J.; Kaufmann, C. Large language models in radiology: Fluctuating performance and decreasing discordance over time. *Eur. J. Radiol.* **2025**, *182*, 111842. [CrossRef] [PubMed]
17. Akan, B.; Dindaroğlu, F.Ç. Content and Quality Analysis of Websites as a Patient Resource for Temporomandibular Disorders. *Turk. J. Orthod.* **2020**, *33*, 203–209.
18. Bronda, S.; Ostrovsky, M.G.; Jain, S.; Malacarne, A. The role of social media for patients with temporomandibular disorders: A content analysis of Reddit. *J. Oral Rehabil.* **2022**, *49*, 1–9. [CrossRef]
19. Cannatà, D.; Galdi, M.; Russo, A.; Scelza, C.; Michelotti, A.; Martina, S. Reliability and Educational Suitability of TikTok Videos as a Source of Information on Sleep and Awake Bruxism: A Cross-Sectional Analysis. *J. Oral Rehabil.* **2025**, *52*, 434–442. [CrossRef]
20. Camargo, E.S.; Quadras, I.C.C.; Garanhani, R.R.; de Araujo, C.M.; Stuginski-Barbosa, J. A Comparative Analysis of Three Large Language Models on Bruxism Knowledge. *J. Oral Rehabil.* **2025**, *52*, 896–903. [CrossRef]
21. Hassan, M.G.; Abdelaziz, A.A.; Abdelrahman, H.H.; Mohamed, M.M.Y.; Ellabban, M.T. Performance of AI-Chatbots to Common Temporomandibular Joint Disorders (TMDs) Patient Queries: Accuracy, Completeness, Reliability and Readability. *Orthod. Craniofac. Res.* 2025; *ahead of print*. [CrossRef]
22. Kula, B.; Kula, A.; Bagcier, F.; Alyanak, B. Artificial intelligence solutions for temporomandibular joint disorders: Contributions and future potential of ChatGPT. *Korean J. Orthod.* **2025**, *55*, 131–141. [CrossRef]
23. Sallam, M.; Barakat, M.; Sallam, M. A Preliminary Checklist (METRICS) to Standardize the Design and Reporting of Studies on Generative Artificial Intelligence–Based Models in Health Care Education and Practice: Development Study Involving a Literature Review. *Interact. J. Med. Res.* **2024**, *13*, e54704. [CrossRef]
24. Bernard, A.; Langille, M.; Hughes, S.; Rose, C.; Leddin, D.; Veldhuyzen Van Zanten, S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. *Am. J. Gastroenterol.* **2007**, *102*, 2070–2077. [CrossRef]
25. Shoemaker, S.J.; Wolf, M.S.; Brach, C. Development of the Patient Education Materials Assessment Tool (PEMAT): A new measure of understandability and actionability for print and audiovisual patient information. *Patient Educ. Couns.* **2014**, *96*, 395–403. [CrossRef]
26. Charnock, D.; Shepperd, S.; Needham, G.; Gann, R. DISCERN: An instrument for judging the quality of written consumer health information on treatment choices. *J. Epidemiol. Community Health* **1999**, *53*, 105–111. [CrossRef] [PubMed]

27. Sallam, M.; Barakat, M.; Sallam, M. Pilot Testing of a Tool to Standardize the Assessment of the Quality of Health Information Generated by Artificial Intelligence-Based Models. *Cureus* **2023**, *15*, e49373. [CrossRef]
28. Flesch, R. A new readability yardstick. *J. Appl. Psychol.* **1948**, *32*, 221–233. [CrossRef] [PubMed]
29. Kincaid, J.; Fishburne, R.; Rogers, R.; Chissom, B. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel, Institute for Simulation and Training. Available online: https://stars.library.ucf.edu/istlibrary/56 (accessed on 1 April 2025).
30. Bellinger, J.R.; De La Chapa, J.S.; Kwak, M.W.; Ramos, G.A.; Morrison, D.; Kesser, B.W. BPPV Information on Google Versus AI (ChatGPT). *Otolaryngol. Head Neck Surg.* **2024**, *170*, 1504–1511. [CrossRef] [PubMed]
31. Incerti Parenti, S.; Bartolucci, M.L.; Biondi, E.; Maglioni, A.; Corazza, G.; Gracco, A.; Alessandri-Bonetti, G. Online Patient Education in Obstructive Sleep Apnea: ChatGPT versus Google Search. *Healthcare* **2024**, *12*, 1781. [CrossRef]
32. Makrygiannakis, M.A.; Giannakopoulos, K.; Kaklamanos, E.G. Evidence-based potential of generative artificial intelligence large language models in orthodontics: A comparative study of ChatGPT, Google Bard, and Microsoft Bing. *Eur. J. Orthod.* **2024**, *46*, cjae017. [CrossRef]
33. Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; et al. WebGPT: Browser-assisted question-answering with human feedback. *arXiv* **2021**, arXiv:2112.09332. [CrossRef]
34. Santamato, V.; Tricase, C.; Faccilongo, N.; Iacoviello, M.; Marengo, A. Exploring the Impact of Artificial Intelligence on Healthcare Management: A Combined Systematic Review and Machine-Learning Approach. *Appl. Sci.* **2024**, *14*, 10144. [CrossRef]
35. Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. Constitutional AI: Harmlessness from AI Feedback. *arXiv* **2022**, arXiv:2212.08073. [CrossRef]
36. Rodhen, R.M.; de Holanda, T.A.; Barbon, F.J.; de Oliveira da Rosa, W.L.; Boscato, N. Invasive surgical procedures for the management of internal derangement of the temporomandibular joint: A systematic review and meta-analysis regarding the effects on pain and jaw mobility. *Clin. Oral Investig.* **2022**, *26*, 3429–3446. [CrossRef]
37. Durham, J.; Wassell, R.W. Recent advancements in temporomandibular disorders (TMDs), Rev. Recent advancements in temporomandibular disorders (TMDs). *Rev. Pain* **2011**, *5*, 18–25. [CrossRef]
38. Ghurye, S.; McMillan, R. Orofacial pain—An update on diagnosis and management, Br. Orofacial pain—An update on diagnosis and management. *Br. Dent. J.* **2017**, *223*, 639–647. [CrossRef]
39. Manfredini, D.; Lombardo, L.; Siciliani, G. Temporomandibular disorders and dental occlusion. Temporomandibular disorders and dental occlusion. A systematic review of association studies: End of an era? *J. Oral Rehabil.* **2017**, *44*, 86–87. [CrossRef]
40. Roganović, J.; Radenković, M.; Miličić, B. Responsible Use of Artificial Intelligence in Dentistry: Survey on Dentists' and Final-Year Undergraduates' Perspectives. *Healthcare* **2023**, *11*, 1480. [CrossRef] [PubMed]
41. Van Grootel, R.J.; Buchner, R.; Wismeijer, D.; van der Glas, H.W. Towards an optimal therapy strategy for myogenous TMD, physiotherapy compared with occlusal splint therapy in an RCT with therapy-and-patient-specific treatment duration. *BMC Musculoskelet. Disord.* **2017**, *18*, 76. [CrossRef] [PubMed]
42. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Chen, D.; Dai, W.; et al. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 3571730. [CrossRef]
43. Borges do Nascimento, I.J.; Pizarro, A.B.; Almeida, J.M.; Azzopardi-Muscat, N.; Gonçalves, M.A.; Björklund, M.; Novillo-Ortiz, D. Infodemics and health misinformation: A systematic review of reviews. Bull. *World Health Organ.* **2022**, *100*, 544–561. [CrossRef]
44. World Health Organization. Ethics and governance of artificial intelligence for health: Large multi-modal models. In *WHO Guidance*; World Health Organization: Geneva, Switzerland, 2025; ISBN 978-92-4-008475-9.
45. Williamson, S.M.; Prybutok, V. Balancing Privacy and Progress: A Review of Privacy Challenges, Systemic Oversight, and Patient Perceptions in AI-Driven Healthcare. *Appl. Sci.* **2024**, *14*, 675. [CrossRef]
46. Hager, P.; Jungmann, F.; Holland, R.; Bhagat, K.; Hubrecht, I.; Knauer, M.; Vielhauer, J.; Makowski, M.; Braren, R.; Kaissis, G.; et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **2024**, *30*, 2613–2622. [CrossRef]
47. Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Papernot, N.; Anderson, R.; Gal, Y. AI models collapse when trained on recursively generated data. *Nature* **2024**, *631*, 755–759. [CrossRef] [PubMed]
48. Rao, K.D.; Peters, D.H.; Bandeen-Roche, K. Towards patient-centered health services in India—A scale to measure patient perceptions of quality. *Int. J. Qual. Health Care* **2006**, *18*, 414–421. [CrossRef] [PubMed]
49. Snyder, C.F.; Wu, A.W.; Miller, R.S.; Jensen, R.E.; Bantug, E.T.; Wolff, A.C. The role of informatics in promoting patient-centered care. *Cancer J.* **2011**, *17*, 211. [CrossRef]
50. Howell, M.D. Generative artificial intelligence, patient safety and healthcare quality: A review. *BMJ Qual. Saf.* **2024**, *33*, 748–759. [CrossRef]

51. Abd-Alrazaq, A.; AlSaad, R.; Alhuwail, D.; Ahmed, A.; Healy, P.M.; Latifi, S.; Aziz, S.; Damseh, R.; Alrazak, S.A.; Sheikh, J. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. *JMIR Med. Educ.* **2023**, *9*, e48291. [CrossRef]
52. Bernal, J.; Mazo, C. Transparency of Artificial Intelligence in Healthcare: Insights from Professionals in Computing and Healthcare Worldwide. *Appl. Sci.* **2022**, *12*, 10228. [CrossRef]