

Quality and Readability of Large Language Models in Dental Radiology: A Comparative Analysis Using FRE, FKGL, and DISCERN

Dental Radyolojide Büyük Dil Modellerinin Kalitesi ve Okunabilirliği: FRE, FKGL ve DISCERN ile Karşılaştırmalı Bir Analiz

Emre Sözen¹, Furkan Özbey².

¹ Emre Sözen, DDS, PhD

Department of Dentomaxillofacial Radiology, Faculty of Dentistry, Afyonkarahisar Health Sciences University, Afyonkarahisar, Türkiye.

E-mail: dtemresozen@gmail.com

ORCID: <https://orcid.org/0000-0001-9767-7162>

² Furkan Özbey, DDS

Department of Dentomaxillofacial Radiology, Faculty of Dentistry, Afyonkarahisar Health Sciences University, Afyonkarahisar, Türkiye.

E-mail: furkanozbey3@gmail.com

ORCID: <https://orcid.org/0000-0002-3123-5190>

*Corresponder author: Emre Sozen, DDS, PhD

Address: Afyonkarahisar Health Sciences University, Faculty of Dentistry, Department of Dentomaxillofacial Radiology, 03030 Afyonkarahisar – Türkiye

Telephone: +90 542 293 14 96

Fax: +90 (272) 216 70 80.

E-mail: dtemresozen@gmail.com

ORCID: <https://orcid.org/0000-0001-9767-7162>

Abstract

Objective: This study aimed to compare the readability, content quality, and reliability of responses generated by five widely used large language models (LLMs), namely ChatGPT, Claude, Gemini, DeepSeek, and Copilot, to original questions in dental radiology.

Methods: A total of 20 expert-developed and reviewed questions covering four thematic categories, namely the biological effects of radiation, protective measures, imaging techniques, and clinical protocols and standards, were submitted to each LLM under standardized conditions without prior conversational context. The generated responses were evaluated using the Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL) indices for readability and an adapted version of the DISCERN instrument for content quality and reliability. Statistical comparisons were performed using the Friedman test with Bonferroni-adjusted pairwise comparisons.

Results: Overall differences among the models were statistically significant for FRE ($p = .024$), FKGL ($p < .001$), and DISCERN ($p < .001$) scores. DeepSeek had the highest mean FRE score, ChatGPT had the highest mean FKGL score, and Claude had the highest mean DISCERN score. Topic-specific differences were not significant for FRE, were limited to the Protective Measures category for FKGL ($p = .030$), and were significant across all four domains for DISCERN ($p \leq .003$). All models generated responses that exceeded the recommended readability levels for patient-oriented materials.

Conclusion: LLM-generated responses in dental radiology demonstrated high linguistic complexity, while their content quality and reliability varied considerably across models. None of the models consistently produced responses that were both sufficiently readable and of high quality across all evaluated domains. These findings highlight the importance of carefully reviewing LLM-generated dental radiology content before its use for educational or communication purposes.

Keywords: Artificial intelligence, dental radiology, large language models, quality, readability.

Özet

Amaç: Bu çalışmada, yaygın olarak kullanılan beş büyük dil modelinin (LLM), ChatGPT, Claude, Gemini, DeepSeek ve Copilot'un, dental radyolojiye ilişkin özgün sorulara verdikleri yanıtların okunabilirlik, içerik kalitesi ve güvenilirlik açısından karşılaştırmalı olarak değerlendirilmesi amaçlanmıştır.

Yöntem: Radyasyonun biyolojik etkileri, koruyucu önlemler, görüntüleme teknikleri ile klinik protokoller ve standartlar olmak üzere dört tematik kategoriyi kapsayan, uzmanlar tarafından geliştirilen ve değerlendirilen toplam 20 soru, önceden herhangi bir konuşma bağlamı olmaksızın standartlaştırılmış koşullar altında her bir LLM'ye yöneltilmiştir. Oluşturulan yanıtlar, okunabilirliğin değerlendirilmesi amacıyla Flesch Okuma Kolaylığı (FRE) ve Flesch-Kincaid Sınıf Düzeyi (FKGL) indeksleri, içerik kalitesi ve güvenilirliğin değerlendirilmesi amacıyla ise uyarlanmış bir DISCERN aracı kullanılarak incelenmiştir. İstatistiksel karşılaştırmalar, Bonferroni düzeltmeli ikili karşılaştırmalarla birlikte Friedman testi kullanılarak gerçekleştirilmiştir.

Bulgular: Modeller arasındaki genel farklılıklar FRE ($p = 0,024$), FKGL ($p < 0,001$) ve DISCERN ($p < 0,001$) skorları açısından istatistiksel olarak anlamlı bulunmuştur. DeepSeek en yüksek ortalama FRE skoruna, ChatGPT en yüksek ortalama FKGL skoruna ve Claude en yüksek ortalama DISCERN skoruna sahip olmuştur. Konuya özgü farklılıklar FRE açısından anlamlı bulunmazken, FKGL açısından yalnızca Koruyucu Önlemler kategorisiyle sınırlı kalmış ($p = 0,030$) ve DISCERN açısından dört alanın tamamında anlamlı bulunmuştur ($p \leq 0,003$). Tüm modeller, hasta odaklı materyaller için önerilen okunabilirlik düzeylerini aşan yanıtlar üretmiştir.

Sonuç: Dental radyolojide LLM'ler tarafından oluşturulan yanıtlar yüksek düzeyde dilsel karmaşıklık gösterirken, içerik kalitesi ve güvenilirlik modeller arasında belirgin farklılık göstermiştir. Modellerin hiçbirisi, değerlendirilen tüm alanlarda hem yeterli okunabilirliğe hem de yüksek içerik kalitesine sahip yanıtları tutarlı biçimde üretememiştir. Bu bulgular, LLM'ler tarafından oluşturulan dental radyoloji içeriklerinin eğitim veya iletişim amacıyla kullanılmadan önce dikkatle gözden geçirilmesinin önemini göstermektedir.

Anahtar Kelimeler: Büyük dil modelleri, dental radyoloji, kalite, okunabilirlik, yapay zeka.

Received: 17/05/2026

Accepted: 10/08/2026

Published Online: 30/08/2026

Cite this article: Sözen E, Özbey F. Quality and Readability of Large Language Models in Dental Radiology: A Comparative Analysis Using FRE, FKGL, and DISCERN. Turk J Health S. 2026;7:2:79-85. doi: 10.29228/tjhealthsport.95453



INTRODUCTION

Radiation is the emission of energy in the form of electromagnetic waves or particles [1]. Ionizing radiation, a high-energy form frequently used in both medicine and dentistry, has the capacity to ionize atoms. This characteristic makes it suitable for both diagnostic and therapeutic applications [2]. In dentistry, radiographic imaging plays an essential role in diagnosis, treatment planning, and disease monitoring. Techniques such as intraoral radiographs, panoramic imaging, and cone-beam computed tomography (CBCT) provide detailed anatomical information about the oral and maxillofacial region, supporting clinical decision-making processes [3].

Although radiation doses used in dental imaging are generally low and are not expected to cause deterministic effects, repeated or unnecessary exposures increase the cumulative radiation dose and may theoretically increase the risk of stochastic effects. Radiation effects are broadly classified as tissue reactions, formerly termed deterministic effects, which occur above a threshold dose, and stochastic effects, such as cancer, for which the probability of occurrence is assumed to increase with dose without a threshold [4]. Therefore, understanding the biological effects of radiation and applying appropriate protective measures is crucial for the safety of both patients and healthcare providers. Principles such as ALARA (As Low As Reasonably Achievable) and ALADA (As Low As Diagnostically Acceptable) aim to minimize radiation exposure while maintaining diagnostic quality, and are fundamental responsibilities for dental practitioners [5].

In recent years, advancements in artificial intelligence (AI) technologies, particularly in natural language processing (NLP), have led to the emergence of large language models (LLMs) as important tools for information access and content generation. These models are developed using deep learning architectures, contain billions of parameters, and are trained on large-scale datasets. Examples include ChatGPT, Gemini, Claude, DeepSeek, and Copilot [6-8]. LLMs hold notable potential in healthcare applications, particularly in responding to information-based queries and generating personalized learning materials [9, 10].

In this context, the quality of health information depends not only on its content but also on how clearly and readably it is presented. Readability reflects the linguistic complexity of written content and may influence how easily health-related information can be processed. This consideration is particularly important in fields such as dentistry, where specialized terminology is frequently used. Therefore, the readability of AI-generated health information has become an important evaluation criterion [11]. As a general benchmark for accessible health communication, organizations such as

the American Medical Association (AMA) and the National Institutes of Health (NIH) recommend that patient education materials should not exceed a sixth-grade reading level [12].

Moreover, the quality and reliability of health information are also of paramount importance. The original DISCERN instrument, a validated tool for assessing the quality of written health information, enables the systematic evaluation of reliability, balance, transparency, and the overall quality of the information provided [13]. This dual perspective, encompassing readability and content quality, forms the core of the present study's analytical framework.

While several studies have assessed the readability of AI-generated responses in general medical domains, systematic comparisons of the readability and content quality of multiple LLMs in dental radiology remain limited [14-16]. This represents an important knowledge gap, given the increasing use of LLMs for accessing and generating health-related information.

This study aims to compare the readability and content quality of responses generated by five widely used LLMs, namely ChatGPT, Claude, Gemini, DeepSeek, and Copilot, to 20 original questions in dental radiology. Readability was evaluated using the Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL) scores, while content quality and reliability were assessed using an adapted version of the DISCERN instrument.

MATERIALS AND METHODS

Ethical Approval

This study does not require approval from an institutional ethics committee, as it does not involve human participants, biological samples, personal data, or any form of clinical intervention. All LLMs evaluated in this study are publicly accessible, and data were obtained exclusively through digital content analysis.

Study Design and Question Development

This study was designed as a comparative observational content analysis to evaluate the readability and content quality of responses generated by different LLMs. Four key topics related to radiation safety and radiology education were determined based on curriculum relevance and expert consensus:

- Biological effects of radiation
- Protective measures
- Imaging techniques
- Clinical protocols and standards

Five open-ended questions were developed for each subject area, resulting in a total of 20 questions. The questions were

prepared and reviewed by two academic oral and maxillofacial radiologists with 13 and 5 years of academic and clinical experience, respectively.

The experts independently evaluated the questions in terms of relevance, clarity, scientific accuracy, and coverage of the selected subject areas. Disagreements were resolved through discussion and consensus. The question set was revised until both experts agreed that each question was clear, evidence-based, and representative of commonly encountered informational needs in dental radiology.

Selection and Technical Characteristics of the LLMs

Five widely used LLM platforms were evaluated between May 15 and 17, 2025, using their official web interfaces. The models included ChatGPT with GPT-4o (OpenAI), Gemini with Gemini 2.5 Pro (Google), Claude with Claude 3.7 Sonnet (Anthropic), Microsoft Copilot in standard chat mode (Microsoft), and DeepSeek-V3-0324 in non-thinking mode (DeepSeek). All platforms were accessed through their publicly available free web interfaces without a paid subscription. Web search, deep research, extended-thinking modes, file analysis, plugins, memory, personalization, and other auxiliary tools were disabled or not used during data collection. Because all models were accessed through their web interfaces, generation parameters such as temperature, top-p, and maximum token length were not user-adjustable; therefore, the default settings of each platform were retained.

Response Generation and Data Collection

Data collection was conducted between May 15 and 17, 2025. All questions were submitted in English. Each question was entered exactly as written, without an additional system prompt, explanatory instruction, readability request, or follow-up question.

Each question was submitted in a new and independent conversation session to prevent the preceding questions and responses from influencing subsequent outputs. No prior conversation context was provided to any model.

The models were evaluated in the following order: ChatGPT, Gemini, Claude, Copilot, and DeepSeek. The 20 questions were presented to all models in the same predetermined order to ensure consistency and comparability across platforms. Data collection was conducted during the morning hours under comparable access conditions. Only the first complete response generated by each model was included. Responses were not regenerated, edited, shortened, expanded, or clarified through follow-up prompts. The use of the first response was intended to represent the output that a typical user would receive after an initial query.

A single response was obtained for each question-model combination. Therefore, the study represents a single-run

evaluation of first-response performance rather than an assessment of within-model response variability. All responses were copied and stored verbatim in plain-text format. The date, time, model name, model version, and question code associated with each response were recorded in a standardized data collection form.

Readability Assessment

The readability of each response was evaluated using the FRE and FKGL indices.

The FRE score was calculated using the following formula [17]:

$$\text{FRE} = 206.835 - (1.015 \times \text{average sentence length}) - (84.6 \times \text{average number of syllables per word})$$

FRE scores range from 0 to 100, with higher scores indicating greater readability and lower linguistic complexity.

The FKGL score was calculated using the following formula [18]:

$$\text{FKGL} = (0.39 \times \text{average sentence length}) + (11.8 \times \text{average number of syllables per word}) - 15.59$$

FKGL scores correspond approximately to United States school grade levels, with higher scores indicating that a more advanced educational level is required to read the text.

Readability analyses were conducted using Microsoft Word, Microsoft 365. The proofing and readability language was set to English. The same software version and language settings were used for all responses.

Readability scores were initially calculated by one researcher and independently verified by a second researcher. Any discrepancies were rechecked against the original response and resolved through verification of the text preparation and software settings. Because FRE and FKGL are automatically calculated using predefined formulas, inter-rater reliability was not calculated for these measurements.

Content Quality and Reliability Assessment Using DISCERN

Beyond syntactic readability, the content quality and reliability of the responses were evaluated using an adapted version of the DISCERN instrument. The original DISCERN is a validated 16-item instrument, with each item rated on a 5-point scale, developed to assess the quality of written health information intended to support informed treatment decisions [13]. For the purposes of the present study, the instrument was adapted by two oral and maxillofacial radiologists to improve its applicability to LLM-generated responses in dental radiology. Each item was reviewed and, where necessary, reworded to reflect the informational and domain-specific nature of the evaluated responses. The modifications were discussed and finalized by consensus between the two radiologists, and the adapted items are presented in Table 1.

Table 1. Adapted DISCERN Questions for Evaluating LLM Responses in Dental Radiology

No	Adapted DISCERN Questions for Evaluating LLM Responses in Dental Radiology
1	Is the purpose of the response clearly stated?
2	Does the response achieve its intended purpose?
3	Is the response relevant to the domain of dental radiology?
4	Are sources or references mentioned or implied?
5	Is there any indication of the information's currency or validity?
6	Is the response balanced and unbiased?
7	Are additional sources of information or consultation suggested?
8	Are uncertainties, limitations, or alternative interpretations addressed?
9	Are clinical mechanisms or justifications adequately explained?
10	Are the potential benefits of the suggested practices explained?
11	Are risks or limitations of the procedure mentioned?
12	Is it explained what may happen if no procedure is performed?
13	Does the response address patient safety or quality-of-life factors?
14	Is it indicated that multiple approaches or alternatives exist?
15	Is shared decision-making or collaboration implied?
16	Overall, is the response reliable, comprehensive, and satisfactory?

All 16 items were applied to every response using the same scoring framework. A score of 1 indicated that the criterion was not addressed or was not evident in the response, whereas a score of 5 indicated that the criterion was fully and explicitly satisfied; scores of 2 to 4 represented intermediate levels of fulfillment. No items were excluded from scoring, resulting in total scores ranging from 16 to 80.

Both oral and maxillofacial radiologists independently evaluated all 100 model-question responses. Inter-rater reliability for the total DISCERN scores was assessed using the intraclass correlation coefficient based on a two-way mixed-effects model with absolute agreement. The ICC for single measures was 0.912 (95% CI, 0.875–0.939), and the ICC for average measures was 0.954 (95% CI, 0.933–0.969), indicating excellent inter-rater reliability. Following the reliability analysis, scoring discrepancies were resolved through discussion and consensus. The resulting consensus scores were used as quantitative indicators of content quality and reliability in the comparative analyses.

Statistical Analysis

All data were organized in Microsoft Excel and analyzed using IBM SPSS Statistics for Windows, version 26.0 (IBM Corp., Armonk, NY, USA). The distributions of the FRE, FKGL, and DISCERN scores were assessed using the Shapiro-Wilk test. Descriptive data were presented as mean and standard deviation.

Because the same 20 questions were submitted to all five LLMs, the scores obtained from the models were treated as related measurements at the question level. Differences among the five models in overall FRE, FKGL, and DISCERN scores were evaluated using the Friedman test. When a statistically significant overall difference was identified, Bonferroni-adjusted pairwise comparisons were performed to determine which models differed from one another.

The same analytical approach was applied separately to each of the four question domains. Because each domain consisted of five questions, domain-specific comparisons were considered exploratory. All statistical tests were two-sided. The significance level for the overall Friedman tests was set

at $p < 0.05$, and Bonferroni-adjusted p values below 0.05 were considered statistically significant in pairwise comparisons. P values below 0.001 were reported as $p < 0.001$.

RESULTS

Table 2. Overall comparison of LLMs based on FRE, FKGL, and DISCERN scores

LLM	FRE Score Mean $\pm$ SD	FKGL Score Mean $\pm$ SD	DISCERN Score Mean $\pm$ SD
ChatGPT	16.53 $\pm$ 19.99 ^{ab}	14.95 $\pm$ 3.07 ^a	34.60 $\pm$ 5.72 ^a
Gemini	25.57 $\pm$ 22.18 ^{ab}	13.81 $\pm$ 3.77 ^{ab}	31.65 $\pm$ 5.59 ^{ac}
Claude	19.20 $\pm$ 10.40 ^{ab}	12.57 $\pm$ 1.54 ^{ab}	57.40 $\pm$ 13.01 ^b
Copilot	14.44 $\pm$ 14.63 ^a	10.71 $\pm$ 1.97 ^b	33.30 $\pm$ 3.77 ^a
DeepSeek	29.32 $\pm$ 21.31 ^b	13.18 $\pm$ 3.39 ^a	27.45 $\pm$ 4.06 ^c
P*	.024	< .001	< .001

* Friedman test; significant results were followed by Bonferroni-adjusted pairwise comparisons. Different superscript letters within a column indicate adjusted $p < .05$.

Table 2 summarizes the overall FRE, FKGL, and DISCERN scores of the five LLMs. Statistically significant differences were observed among the models for FRE ($p = .024$), FKGL ($p < .001$), and DISCERN ($p < .001$) scores. DeepSeek had the highest mean FRE score (29.32 $\pm$ 21.31), whereas Copilot had the lowest (14.44 $\pm$ 14.63). Bonferroni-adjusted pairwise comparisons showed that the FRE score of DeepSeek was significantly higher than that of Copilot; no other pairwise differences were significant.

For FKGL, ChatGPT had the highest mean score (14.95 $\pm$ 3.07), whereas Copilot had the lowest (10.71 $\pm$ 1.97). Copilot had significantly lower FKGL scores than ChatGPT and DeepSeek. No statistically significant differences were observed among the remaining model pairs.

Claude had the highest mean DISCERN score (57.40 $\pm$ 13.01), whereas DeepSeek had the lowest (27.45 $\pm$ 4.06). Claude had significantly higher DISCERN scores than all other models. DeepSeek had significantly lower scores than ChatGPT, Claude, and Copilot, but did not differ significantly from Gemini.

Table 3. FRE scores of LLMs according to topic

Topics	ChatGPT Mean $\pm$ SD	Gemini Mean $\pm$ SD	Claude Mean $\pm$ SD	Copilot Mean $\pm$ SD	DeepSeek Mean $\pm$ SD	p^*
Biological Effects of Radiation	13.90 $\pm$ 24.46	34.88 $\pm$ 11.29	27.09 $\pm$ 4.07	30.49 $\pm$ 10.96	21.83 $\pm$ 20.10	.265
Protective Measures	10.94 $\pm$ 14.70	18.86 $\pm$ 26.97	18.51 $\pm$ 8.02	16.60 $\pm$ 10.67	36.16 $\pm$ 15.90	.217
Imaging Techniques	21.01 $\pm$ 25.02	28.64 $\pm$ 22.33	22.17 $\pm$ 9.52	11.38 $\pm$ 8.82	20.02 $\pm$ 28.47	.600
Clinical Protocols and Standards	20.23 $\pm$ 19.17	19.86 $\pm$ 27.66	9.05 $\pm$ 11.15	5.71 $\pm$ 8.79	39.26 $\pm$ 18.15	.055

* P values were obtained using the Friedman test.

Table 3 presents the FRE scores of the LLMs according to topic. Although the highest mean FRE scores were observed for Gemini in the Biological Effects of Radiation and Imaging Techniques domains and for DeepSeek in the Protective Measures and Clinical Protocols and Standards domains, no statistically significant differences were observed among the models in any of the four domains ($p = .055-.600$).

Table 4. FKGL scores of LLMs according to topic

Topics	ChatGPT	Gemini	Claude	Copilot	DeepSeek	p*
	Mean ±SD	Mean ±SD	Mean ±SD	Mean ±SD	Mean ±SD	
Biological Effects of Radiation	14.89±3.09	11.89±0.92	11.41±0.34	12.07±1.55	11.91±2.46	.119
Protective Measures	16.44±2.48 ^a	15.82±3.14 ^{ab}	12.60±1.12 ^{ab}	14.18±2.63 ^{ab}	9.68±2.47 ^b	.030
Imaging Techniques	14.54±4.23	13.24±3.86	12.20±1.13	13.98±1.08	12.31±4.06	.569
Clinical Protocols and Standards	13.91±2.62	14.28±3.88	14.07±2.00	15.17±1.29	8.92±3.96	.060

* Friedman test; significant results were followed by Bonferroni-adjusted pairwise comparisons. Different superscript letters within a row indicate adjusted p < .05.

Table 4 presents the FKGL scores of the LLMs according to topic. A statistically significant difference among the models was observed only for Protective Measures (p = .030). No statistically significant differences were found for Biological Effects of Radiation (p = .119), Imaging Techniques (p = .569), or Clinical Protocols and Standards (p = .060).

For Biological Effects of Radiation, ChatGPT had the highest mean FKGL score (14.89 ± 3.09), whereas Claude had the lowest (11.41 ± 0.34). For Imaging Techniques, ChatGPT had the highest mean score (14.54 ± 4.23), whereas Claude had the lowest (12.20 ± 1.13).

For Protective Measures, ChatGPT had the highest mean FKGL score (16.44 ± 2.48), whereas DeepSeek had the lowest (9.68 ± 2.47). Bonferroni-adjusted pairwise comparisons showed that the FKGL score of DeepSeek was significantly lower than that of ChatGPT; no other pairwise differences were significant.

For Clinical Protocols and Standards, Copilot had the highest mean FKGL score (15.17 ± 1.29), whereas DeepSeek had the lowest (8.92 ± 3.96); however, the overall difference among the models was not statistically significant (p = .060).

Table 5. DISCERN scores of LLMs according to topic

Topics	ChatGPT	Gemini	Claude	Copilot	DeepSeek	p*
	Mean ±SD	Mean ±SD	Mean ±SD	Mean ±SD	Mean ±SD	
Biological Effects of Radiation	30.20±3.34 ^{ab}	26.20±1.30 ^a	36.40±2.60 ^b	34.00±2.54 ^{ab}	25.60±5.50 ^a	.002
Protective Measures	37.40±6.58 ^{ab}	38.80±4.91 ^{ab}	66.40±1.81 ^a	35.80±2.94 ^{ab}	31.20±3.49 ^b	.003
Imaging Techniques	33.20±4.32 ^{ab}	29.80±1.92 ^{ab}	65.80±2.77 ^a	33.20±3.56 ^{ab}	25.80±2.28 ^b	.001
Clinical Protocols and Standards	37.60±5.85 ^a	31.80±3.70 ^a	61.00±5.43 ^{ab}	30.20±4.38 ^b	27.20±2.28 ^b	.001

* Friedman test; significant results were followed by Bonferroni-adjusted pairwise comparisons. Different superscript letters within a row indicate adjusted p < .05.

Table 5 presents the DISCERN scores of the LLMs according to topic. Statistically significant differences among the models were observed for Biological Effects of Radiation (p = .002), Protective Measures (p = .003), Imaging Techniques (p = .001), and Clinical Protocols and Standards (p = .001).

For Biological Effects of Radiation, Claude had the highest mean DISCERN score (36.40 ± 2.60), whereas DeepSeek had the lowest (25.60 ± 5.50). Claude had significantly higher scores than Gemini and DeepSeek. For Protective Measures, Claude had the highest mean score (66.40 ± 1.81), whereas DeepSeek

had the lowest (31.20 ± 3.49); the pairwise difference between Claude and DeepSeek was statistically significant.

For Imaging Techniques, Claude had the highest mean DISCERN score (65.80 ± 2.77), whereas DeepSeek had the lowest (25.80 ± 2.28); the pairwise difference between Claude and DeepSeek was statistically significant. For Clinical Protocols and Standards, Claude had the highest mean score (61.00 ± 5.43), whereas DeepSeek had the lowest (27.20 ± 2.28). ChatGPT and Gemini had significantly higher scores than Copilot and DeepSeek, while Claude did not differ significantly from the other models in the pairwise comparisons.

DISCUSSION

This study compared the readability and content quality of responses generated by five LLMs to questions in dental radiology. The expanding use of LLMs across medical specialties has increased interest in their potential applications in healthcare [19]. Overall, the evaluated responses demonstrated high linguistic complexity, while content quality and reliability varied significantly among the models. The radiology-focused responses evaluated in this study demonstrated low readability according to both FRE and FKGL measures. Mean FRE scores ranged from 14.44 ± 14.63 for Copilot to 29.32 ± 21.31 for DeepSeek, indicating very low readability. Similarly, mean FKGL scores ranged from 10.71 ± 1.97 for Copilot to 14.95 ± 3.07 for ChatGPT and were considerably above the recommended sixth-grade or lower reading level for patient education materials [12]. Although the questions in this study were not limited to patient-oriented communication, these thresholds provide a useful reference for evaluating the linguistic accessibility of health-related information. The findings indicate that the dental radiology responses generated by the models generally contained a high level of linguistic complexity.

In contrast, previous studies conducted on more general or patient-oriented dental topics have reported more readable outputs. Esmailpour et al. [20] reported FRE scores of 53.05 for ChatGPT and 64.94 for Gemini in responses to frequently asked questions about dental prostheses. In the same study, FKGL scores were 10.45 for ChatGPT, 7.82 for Gemini, and 8.38 for Copilot. Studies conducted by Güven et al. [14] and Bayraktar et al. [13] also demonstrated that responses related to traumatic dental injuries and pediatric dentistry generally had FRE scores ranging from 41 to 52 and lower FKGL scores than those observed in the present study.

These differences among studies may partly reflect the terminological and conceptual complexity of dental radiology. Radiology-related questions include multisyllabic and specialized terms involving the biological effects of ionizing radiation, ALARA and ALADA principles, imaging modality

selection, and technical imaging parameters. This may contribute to longer sentences, lower FRE scores, and higher FKGL values. However, direct comparisons should be interpreted cautiously because question content, model versions, prompting procedures, and assessment methods differ among studies.

Some topic-dependent differences among the models were observed in the topic-specific FRE and FKGL analyses (Tables 3 and 4). Gemini generated the highest mean FRE score in the “Biological Effects of Radiation” domain, whereas DeepSeek generated the lowest mean FKGL score in the “Protective Measures” domain. In this domain, the FKGL score of DeepSeek was significantly lower than that of ChatGPT, indicating lower linguistic complexity. ChatGPT had the highest mean FKGL scores in the “Biological Effects of Radiation” and “Imaging Techniques” domains, whereas Copilot had the highest mean FKGL score in the “Clinical Protocols and Standards” domain. However, most topic-specific comparisons were not statistically significant. This suggests that the models tended to produce similar levels of linguistic complexity across different subject areas and that their overall performance remained largely consistent.

These findings indicate that although LLMs can be used to generate health information and educational materials, the readability of the resulting texts should be evaluated before their direct use. Requesting responses at a specified educational level and subsequently simplifying complex terminology may improve the accessibility of the generated content. These approaches may be particularly useful in fields such as dental radiology, where technical language is used extensively.

In this study, mean DISCERN scores ranged from 27.45 ± 4.06 for DeepSeek to 57.40 ± 13.01 for Claude. Claude scored significantly higher than all other models, while DeepSeek scored significantly lower than ChatGPT, Claude, and Copilot, but did not differ significantly from Gemini. These findings indicate statistically significant differences among the evaluated models in terms of content quality and reliability.

Güven et al. [14] evaluated responses related to traumatic dental injuries using a modified eight-item DISCERN scale with a maximum score of 40. In that study, ChatGPT-3.5 received a score of approximately 29, whereas ChatGPT-4.0 and Gemini received scores of approximately 31 and 32, respectively. Kılınç and Mansız [21] evaluated ChatGPT responses to orthodontic questions using a five-point adaptation of DISCERN and reported moderate reliability scores of approximately 3.00 for general questions and 2.38 for treatment-related questions. These findings indicate that the quality of dental information generated by LLMs may vary according to the model used, the question domain, and the assessment method [22]. However, direct comparisons among studies are limited because DISCERN adaptations

and scoring methods differ. These findings suggest that LLM-generated dental radiology content may benefit from expert review and readability-focused refinement before being used for educational or clinical communication purposes.

The FRE and FKGL results showed that although some models generated detailed responses, the language used frequently exceeded the recommended readability levels. Uranbey et al. [23] reported that ChatGPT could generate diagnostic and treatment-related recommendations for oral pathologies, but discrepancies between artificial intelligence outputs and expert decisions remained in complex scenarios. Güven et al. [14] also demonstrated that readability was an important area of evaluation in responses related to traumatic dental injuries. These results support the importance of professional oversight, particularly for LLM-generated content intended for patients or designed to support clinical communication.

From an educational perspective, the linguistic complexity of the generated responses may limit the use of these texts as independent educational resources. Bayraktar et al. [13] reported that LLM responses to parental questions in pediatric dentistry demonstrated higher readability. Helvacioğlu-Yigit et al. [15] showed that most educational content generated by chatbots in oral and maxillofacial radiology exceeded the recommended seventh-grade reading level [12]. These findings demonstrate that the ability to generate detailed information is not always accompanied by the ability to produce accessible and easily readable content. Future studies may benefit from evaluating the actual comprehension performance of users at different educational levels, in addition to readability and content quality.

One of the strengths of this study was its standardized data collection procedure. The same 20 questions were presented to all models in the same predetermined order, and each question was submitted in a new and independent session without prior conversational context. Thus, contextual carryover between responses was minimized. In addition, readability was evaluated using two established indices, and the automatically generated scores were independently verified by a second researcher. The evaluation of multiple LLMs across four different dental radiology subject areas also enabled comparisons among the models.

This study has several limitations. Only one response was obtained for each question-model combination, and the relatively small question set, particularly the five questions per domain, limits the generalizability of the findings and makes topic-specific analyses exploratory. The models were accessed through free web interfaces, and generation parameters, including response length, could not be controlled. FRE and FKGL assess linguistic complexity rather than factual accuracy or actual comprehension. The adapted DISCERN instrument was developed through expert consensus and

was not independently validated. Finally, because LLMs are continuously updated, the findings are specific to the English-language questions, model versions, platform conditions, and data collection period evaluated in this study.

CONCLUSION

This study provides a comparative evaluation of LLM-generated responses in terms of readability and content quality within the specialized context of dental radiology. The outputs of all evaluated models exceeded the recommended readability levels for accessible health communication, while their content quality and reliability varied considerably. Claude achieved the highest DISCERN scores overall and across all four domains, and its overall DISCERN score was significantly higher than those of all other models. However, none of the models consistently demonstrated favorable performance in both readability and DISCERN scores across all evaluated domains. These findings indicate that LLM-generated dental radiology content may require readability-focused prompting, expert review, and post-editing before being used for educational or communication purposes. Future studies should evaluate larger question sets, repeated model outputs, and actual user comprehension to determine how effectively LLM-generated content can support dental education and clinical communication.

Contribution to the Literature: This study provides a comparative evaluation of five widely used large language models in dental radiology by considering both readability and content quality. Unlike evaluations based on a single performance dimension, the study combines FRE and FKGL readability measures with an adapted DISCERN assessment of content quality and reliability across four dental radiology domains. The findings demonstrate that higher content quality does not necessarily coincide with greater readability and highlight the need to consider both dimensions when evaluating LLM-generated dental radiology content.

Ethics Committee Approval: Ethics committee approval was not required because the study involved no human participants, identifiable personal data, biological materials, or clinical interventions and was based solely on the analysis of text generated by publicly accessible LLM interfaces.

Acknowledgements: None.

Author Contributions: Conceptualization: ES. Study design and protocol: ES, FO. Data collection: ES, FO. Data analysis and interpretation: ES, FO. Writing - original draft: ES. Writing - review and editing: ES, FO. All authors read and approved the final manuscript.

Declaration of Interests: The authors have no conflicts of interest to declare.

Funding: This study received no funding.

Data Availability: The datasets used and/or analyzed during the

current study are available from the corresponding author on reasonable request.

REFERENCES

1. Apte K, Bhide S. Basics of radiation. In: Verma S, Srivastava AK, editors. *Advanced radiation shielding materials: radiation and radiological protection*. Amsterdam: Elsevier; 2024. p. 1-23. <https://doi.org/10.1016/B978-0-323-95387-0.00013-3>.
2. White SC, Pharoah MJ. *Oral radiology: principles and interpretation*. 6th ed. St. Louis: Mosby Elsevier; 2009.
3. Scarfe WC, Farman AG. What is cone-beam CT and how does it work? *Dent Clin North Am*. 2008;52(4):707-730. <https://doi.org/10.1016/j.cden.2008.05.005>.
4. Cousins C, Miller DL, Bernardi G, Rehani MM, Schofield P, Vanió E, et al. Radiological protection in cardiology. *ICRP Publication 120*. *Ann ICRP*. 2013;42(1):1-125. <https://doi.org/10.1016/j.icrp.2012.09.001>.
5. Jajui PP, Jajui SP. Cone-beam computed tomography: time to move from ALARA to ALADA. *Imaging Sci Dent*. 2015;45(4):263-265. <https://doi.org/10.5624/isd.2015.45.4.263>.
6. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. *arXiv [Preprint]*. 2023. [arXiv:2303.13375](https://arxiv.org/abs/2303.13375). <https://doi.org/10.48550/arXiv.2303.13375>.
7. Yilmaz BE, Gokkurt Yilmaz BN, Ozbey F. Artificial intelligence performance in answering multiple-choice oral pathology questions: a comparative analysis. *BMC Oral Health*. 2025;25:573. <https://doi.org/10.1186/s12903-025-05926-2>.
8. Akpınar H. Comparison of responses from different artificial intelligence-powered chatbots regarding the All-on-Four dental implant concept. *BMC Oral Health*. 2025;25:922. <https://doi.org/10.1186/s12903-025-06294-7>.
9. Wang F, Preininger A. AI in health: state of the art, challenges, and future directions. *Yearb Med Inform*. 2019;28(1):16-26. <https://doi.org/10.1055/s-0039-1677908>.
10. Sözen E. Large language model accuracy in dental radiology: effects of cognitive complexity and content domain. *Sci Rep*. 2026 Jul 20. <https://doi.org/10.1038/s41598-026-63160-4>.
11. Alshehri A, Alghofaili N, Alshunaiber R, Alkadi L. Quality and readability assessment of internet-based information on common prosthodontic treatments. *Int J Prosthodont*. 2022;35(1):62-67. <https://doi.org/10.11607/ijp.7063>.
12. Rouhi AD, Ghanem YK, Yolchieva L, Saleh Z, Joshi H, Moccia MC, et al. Can artificial intelligence improve the readability of patient education materials on aortic stenosis? A pilot study. *Cardiol Ther*. 2024;13(1):137-147. <https://doi.org/10.1007/s40119-023-00347-0>.
13. 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(2):105-111. <https://doi.org/10.1136/jech.53.2.105>.
14. Bayraktar Nahir C. Can ChatGPT be guide in pediatric dentistry? *BMC Oral Health*. 2025;25:9. <https://doi.org/10.1186/s12903-024-05393-1>.
15. Guven Y, Ozdemir OT, Kavan MY. Performance of artificial intelligence chatbots in responding to patient queries related to traumatic dental injuries: a comparative study. *Dent Traumatol*. 2025;41(3):338-347. <https://doi.org/10.1111/edt.13020>.
16. Helvacdoglu-Yigit D, Demirturk H, Ali K, Tamimi D, Koenig L, Almashraqi A. Evaluating artificial intelligence chatbots for patient education in oral and maxillofacial radiology. *Oral Surg Oral Med Oral Pathol Oral Radiol*. 2025;139(6):750-759. <https://doi.org/10.1016/j.oooo.2025.01.001>.
17. Flesch R. A new readability yardstick. *J Appl Psychol*. 1948;32(3):221-233. <https://doi.org/10.1037/h0057532>.
18. Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS. Derivation of new readability formulas: automated readability index, Fog Count and Flesch Reading Ease formula for Navy enlisted personnel. Millington (TN): Naval Technical Training Command; 1975. Research Branch Report 8-75.
19. Lechien JR, Rameau A. Applications of ChatGPT in otolaryngology-head neck surgery: a state-of-the-art review. *Otolaryngol Head Neck Surg*. 2024;171(3):667-677. <https://doi.org/10.1002/ohn.807>.
20. Esmailpour H, Rassaie V, Babaei Hemmati Y, Falahchahi M. Performance of artificial intelligence chatbots in responding to the frequently asked questions of patients regarding dental prostheses. *BMC Oral Health*. 2025;25:574. <https://doi.org/10.1186/s12903-025-05965-9>.
21. Kılınç DD, Mansız D. Examination of the reliability and readability of Chatbot Generative Pretrained Transformer's responses to questions about orthodontics and the evolution of these responses in an updated version. *Am J Orthod Dentofacial Orthop*. 2024;165(5):546-555. <https://doi.org/10.1016/j.jado.2023.11.012>.
22. Sözen E, Akpınar H, Yaman S. Comparative performance of large language models for patient-oriented support in dental trauma emergencies. *BMC Oral Health*. 2026;26:665. <https://doi.org/10.1186/s12903-026-08037-8>.
23. Uranbey Ö, Özbey F, Kaygısız Ö, Ayrancı F. Assessing ChatGPT's diagnostic accuracy and therapeutic strategies in oral pathologies: a cross-sectional study. *Cureus*. 2024;16(4):e58607. <https://doi.org/10.7759/cureus.58607>.