JOURNAL of ENDODONTICS and RESTORATIVE DENTISTRY

E-ISSN: 2980-3748

OPEN ACCESS

creative commons

BY NC

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

Original Article

# Performance Comparison of Large Language Models in Dentistry: A Focus on Restorative Dentistry Using Turkish Specialty Exam Questions

Muhammed Baytar [a], Fatma Pertek Hatipoğlu [b]

[a] Department of Restorative Dentistry, Recep Tayyip Erdoğan University, Rize, Türkiye.
[b] Department of Endodontics, Recep Tayyip Erdoğan University, Rize, Türkiye.

ARTICLE INFO

Received: 17.09.2025
Completion of First Review: 25.12.2025
Accepted: 29.01.2026
Published: 01.03.2026

KEYWORDS

Artificial Intelligence
Dental Education
Large Language Models
Restorative Dentistry

CORRESPONDENCE

Muhammed Baytar
Department of Restorative Dentistry, Recep Tayyip Erdoğan University, Rize, Türkiye.
E-mail: muhammed.baytar@erdogan.edu.tr

CLINICAL SIGNIFICANCE

Large language models may support dental education and clinical knowledge reinforcement. Model-specific differences in accuracy, speed, and response depth highlight the importance of selecting appropriate systems based on educational or clinical objectives rather than relying on a single tool.

ABSTRACT

**Objectives**: This study aimed to comparatively evaluate the accuracy, response time, and content length of five large language models (LLMs), ChatGPT-4o, ChatGPT-o3-mini, Deepseek-v3, Google Gemini 2.0 Flash, and Microsoft Copilot, based on restorative dentistry questions from the Turkish Dental Specialty Exam (DUS).

**Materials and Methods:** A total of 100 multiple-choice questions from the restorative dentistry sections of the DUS (2016–2024) were presented to each LLM model in Turkish under standardized testing conditions. Model performance was assessed based on three primary metrics: answer accuracy, response generation time, and content length (word count). Temporal consistency was evaluated by re-submitting a 10% sample after two weeks. Statistical analyses were conducted to compare differences among the models.

**Results:** ChatGPT-o3-mini achieved the highest accuracy (96%), followed by ChatGPT-4o (90%), Deepseek-v3 (88%), and both Gemini and Copilot (85%). Microsoft Copilot was the fastest model (median: 3.19 s), while Deepseek-v3 was the slowest (median: 25.64 s). Google Gemini 2.0 Flash produced the longest responses (median: 218 words), whereas Microsoft Copilot generated the shortest (median: 34 words).

**Conclusion:** LLMs demonstrate promising potential for supporting dental education, particularly in restorative domains. Among the evaluated models, ChatGPT-o3-mini showed the highest overall accuracy, suggesting its relative suitability for knowledge-based tasks in dentistry. However, performance varied by model and topic, indicating that no single system is universally superior. Model selection should be guided by the intended application, whether speed, depth, or accuracy is prioritized. The use of standardized specialty exam questions offers a reliable framework for benchmarking LLM performance in domain-specific contexts.

## 1. Introduction

In recent years, the rapid advancements in Large Language Models (LLMs) have led to significant transformations in healthcare services, with widespread applications in clinical decision support systems, diagnostic processes, patient monitoring, and access to medical information.[1,2] These systems offer advantages such as accuracy, speed, and accessibility, providing substantial support to healthcare professionals and facilitating clinical workflows.[3,4] The opportunities presented by LLMS are increasingly being adopted in the field of dentistry as well. Owing to their capacity to respond to knowledge-based dental queries, LLMs are being recognized as effective tools in both educational settings and treatment planning.[5-7] These models assist in decision-making processes such as clinical guidance, material selection, and case evaluation, and hold considerable potential particularly in restorative dentistry, which is a practice-intensive domain.[8,9]

Among the advanced artificial intelligence models developed in recent times, ChatGPT-4o, ChatGPT-o3-mini, Deepseek-v3, Google Gemini 2.0 Flash, and Microsoft Copilot stand out by offering various advantages in knowledge-intensive fields such as dentistry.[6,7,10] ChatGPT-4o, with its text-input performance and high accuracy rates, offers broad application potential, while ChatGPT-o3-mini is notable for its compact structure and ability to generate fast and accurate responses. Deepseek-v3 tends to produce detailed and explanatory answers, particularly in the generation of scientific content and the communication of technical information. Google Gemini 2.0 Flash, on the other hand, distinguishes itself through its high response speed and user-friendly interface. Microsoft Copilot provides a distinct user experience by producing concise and focused answers. A systematic evaluation of these models' performance in responding to dental knowledge-based queries may contribute to a better understanding of their potential in both educational and clinical contexts.

Although some studies in the current literature have evaluated the performance of LLM models in the field of dentistry, most focus on a single model or are limited to superficial comparisons among a small number of systems.[6-8] In particular, there is a notable lack of comprehensive studies comparing advanced models in terms of their performance across specific subcategories of dental knowledge, including anatomical structures, microbiology, restorative materials, therapeutic treatments, and aesthetic Technologies.[10-12] Moreover, the absence of systematic analyses evaluating these models based on criteria such as response accuracy, generation time, and content scope hinders the reliable assessment of their potential for clinical and educational applications.[9,13] This limitation makes it difficult to objectively understand the strengths and weaknesses of each model and to confidently integrate them into clinical practice. In this context, a comparative evaluation of the responses provided by these LLMs to dental knowledge-based queries—in terms of accuracy, speed, and comprehensiveness—holds the potential to fill a critical gap in the literature.[14]

This study aims to comparatively evaluate the performance of the artificial intelligence models ChatGPT-4o, ChatGPT-o3-mini, Deepseek-v3, Google Gemini 2.0 Flash, and Microsoft Copilot in terms of their knowledge level, response time, and content scope in the field of dentistry. The models were analyzed based on objective criteria including accuracy rates, response generation times, and content lengths. Additionally, their performance was assessed across various subcategories of dental knowledge. To date, no study in the literature has examined such a wide range of LLMs while simultaneously providing an in-depth evaluation of dental knowledge domains. In this regard, the study is expected to reveal which LLMs are more reliable and efficient for use in dental education and clinical decision support systems.

The null hypothesis tested in this study is that there is no statistically significant difference among the models, ChatGPT-4o, ChatGPT-o3-mini, Deepseek-v3, Google Gemini 2.0 Flash, and Microsoft Copilot, in terms of the accuracy, response time, and content scope of their answers to dentistry-specific questions.

## 2. Materials and Methods

### 2.1. Study Design

This comparative observational study was designed to evaluate the performance of five advanced large language models,ChatGPT-4o, ChatGPT-o3-mini, Deepseek-v3, Google Gemini 2.0 Flash, and Microsoft Copilot, in answering dentistry-specific knowledge-based questions. The models' responses were assessed using objective performance criteria including accuracy, response time, and content length. Additionally, the consistency of responses over time was analyzed to determine the temporal reliability of each model.

### 2.2. A priori Power Analysis

A priori power analysis was conducted to justify the sample size for between-model accuracy comparisons. Because each multiple-choice question (MCQ) was evaluated by all five LLMs, accuracy comparisons were based on a paired design at the question level. Accordingly, statistical power was primarily determined by the proportion of discordant responses between model pairs rather than by marginal accuracy alone.

Assuming a two-sided significance level of 0.05, a total of 100 paired MCQs provides approximately 80% power to detect absolute differences in accuracy of approximately 10–12% between models under plausible discordance rates of 15–20%, using McNemar-type tests for paired proportions. When accounting for multiple pairwise comparisons, the minimum detectable difference increases to approximately 15%. These effect sizes were considered meaningful for benchmarking model performance in an educational and clinical context.

Category-level analyses were planned as secondary and exploratory due to limited numbers of MCQs in some domains, and no separate power calculations were performed for these subgroup analyses.

### 2.3. Model versioning and inference settings

All LLMs were evaluated using their publicly available web-based interfaces during a defined testing window (January–February 2025). Model versions and access tiers were recorded at the time of evaluation. ChatGPT-4o and ChatGPT-o3-mini were accessed via a paid subscription tier, whereas Google Gemini 2.0 Flash and Microsoft Copilot were accessed through their standard publicly available interfaces. Deepseek-v3 was accessed via its official web interface.

No custom system prompts were used. All queries were submitted using the default chat interface of each platform, and responses were generated using platform-default inference parameters. When model-specific generation settings were not user-adjustable via the interface, they were assumed to be fixed at their default values. All evaluations were conducted using the standard default response mode provided by each platform at the time of testing.

All models were prompted with identical user inputs. During testing, no safety-related refusals or content blocks were observed for the included multiple-choice questions; all models provided a substantive response to each prompt.

### 2.4. Data Source and Question Categories

The dataset consisted of 100 MCQs derived from Restorative Dentistry sections of the Dental Specialty Examination (DUS), administered in Türkiye by the Student Selection and Placement Center (OSYM) between 2016 and 2024. Each DUS exam comprises 120 questions, including 80 clinical science items and 40 basic science items. Restorative Dentistry is one of the eight clinical science disciplines represented in the clinical component, typically contributing 10 questions per exam.

The questions were selected to ensure broad coverage of key domains within restorative dental knowledge. In order to facilitate detailed analysis, each question was classified into one of six major subcategories of dental knowledge based on content:

1. Anatomical Structures and Oral Environment
2. Dental Caries and Other Lesions
3. Restorative Materials and Application Techniques
4. Therapeutic and Preventive Procedures
5. Aesthetic and Advanced Technologies
6. Microbiology and Oral Biofilm

Sample questions for each category along with their English equivalents are provided in Table 1, while the complete list of questions is included in the Supplemental File.

### 2.5. Testing Procedure of AI Models

All 100 questions were submitted to each AI model by the same operator (M.B) under standardized conditions to minimize operator bias. The questions were input in Turkish, reflecting their original language in the DUS exams. The models were not tested concurrently but sequentially in a controlled testing environment to avoid temporal or server-based variations that might influence performance.

Each model's response to each question was evaluated using the following three core metrics:

*Accuracy:* Binary scoring (correct/incorrect) based on the model's ability to select or generate the correct answer as per the official DUS key.

*Response Time:* Measured in milliseconds using a digital stopwatch initiated the moment a question was submitted and stopped once a complete answer was received.

*Response Length:* Calculated as the total word count of the generated response using Microsoft Word's word count tool.

All responses and performance metrics were recorded systematically, and comparative analyses were conducted.

### 2.6. Response Consistency and Reliability Testing

To evaluate the temporal consistency of the language models, a subset of 10 questions (10% of the total dataset) was re-submitted to each model exactly two weeks after the initial testing phase. The second-round responses were compared with the originals in terms of both accuracy (correctness of answer) and content stability (response time and word count).

### 2.7. Ethical Considerations

As this study did not involve human participants, patient data, or clinical intervention, ethical approval was not required. However, all data sources used (DUS questions) are publicly available through the official website of ÖSYM (https://www.osym.gov.tr), and were used solely for academic research purposes.

**Table 1.** The presentation of two different question examples and their answers from each category.

| Operative Dentistry Topics | Examples of questions | Choices |
|---|---|---|
| Anatomical Structures and Oral Environment | What is the name of the unmineralized dentin layer adjacent to the odontoblast cells in the pulp? | A) Mantle dentin<br>B) Circumpulpal dentin<br>C) Secondary dentin<br>D) Primary dentin<br>E) Predentin |
| Anatomical Structures and Oral Environment | What is the name of the optical image formed by the arrangement of prism groups in different directions to prevent enamel fracture? | A) Retzius line<br>B) Hunter-Schreger band<br>C) Enamel lamella<br>D) Perikymata<br>E) Enamel tuft |
| Dental Caries and Other Lesions | Which of the following is not one of the layers of an initial enamel caries lesion? | A) Dark zone<br>B) Body of the lesion<br>C) Translucent zone<br>D) Infected layer<br>E) Surface layer |
| Dental Caries and Other Lesions | Which of the following is not a risk factor for the development of root surface caries in elderly individuals? | A) Use of medications that cause dry mouth<br>B) Cariogenic diet<br>C) Increased salivary flow rate<br>D) Gingival recession<br>E) Use of partial dentures |
| Restorative Materials and Application Techniques | Which monomer is added to composite resins to reduce or control viscosity? | A) TEGDMA<br>B) Bis-GMA<br>C) UDMA<br>D) 10-MDP<br>E) 4-META |
| Restorative Materials and Application Techniques | Which component in composite resins binds the organic and inorganic phases together and ensures stress distribution? | A) Camphorquinone<br>B) Tertiary amine<br>C) Triethylene glycol dimethacrylate<br>D) Barium glass<br>E) Silane |
| Therapeutic and Preventive Treatments | Which of the following is used in the microabrasion procedure for the treatment of tooth discoloration? | A) Orthophosphoric acid<br>B) Maleic acid<br>C) Citric acid<br>D) Hydrochloric acid<br>E) Hydrofluoric acid |
| Therapeutic and Preventive Treatments | In which of the following conditions is treatment with the resin infiltration technique not recommended? | A) Fluorosis cases<br>B) Cavitated root surface caries<br>C) Hypomineralization cases<br>D) Initial interproximal caries<br>E) White spot lesions |
| Aesthetic and Advanced Technologies | Which of the following lasers is the least likely to be used in the treatment of dentin hypersensitivity? | A) $CO_2$ laser<br>B) Argon laser<br>C) Er:YAG laser<br>D) Er,Cr:YSGG laser<br>E) Nd:YAG laser |
| Aesthetic and Advanced Technologies | Which optical property is defined as reflecting short-wavelength light as blue and transmitting long-wavelength light as yellow/red? | A) Fluorescence<br>B) Translucency<br>C) Opalescence<br>D) Opacity<br>E) Transparency |
| Microbiology and Oral Biofilm | Which of the following bacteria is not a member of the Mitis group streptococci? | A) Streptococcus sanguinis<br>B) Streptococcus infantis<br>C) Streptococcus peroris<br>D) Streptococcus oralis<br>E) Streptococcus cristatus |
| Microbiology and Oral Biofilm | Which of the following microorganisms does not belong to the mutans group of oral streptococci? | A) Streptococcus salivarius<br>B) Streptococcus sobrinus<br>C) Streptococcus ferus<br>D) Streptococcus ratti<br>E) Streptococcus criceti |

The questions submitted to the chatbots were in Turkish without an English translation. However, the table has been translated into English for the readers' convenience.

### 2.7. Statistical Analysis

All statistical analyses were performed using Jamovi software (Version 2.3.28), with a significance threshold set at $p < 0.05$. To compare the overall accuracy rates among the five LLMs, the Cochran's Q test was applied, and McNemar's test was used for pairwise comparisons of correct versus incorrect responses. Differences in response time and word count across models were evaluated using the Kruskal–Wallis test, followed by Dwass–Steel–Critchlow–Fligner post hoc tests to determine pairwise significance. For domain-specific performance, accuracy across subcategories of restorative dentistry (e.g., anatomical structures, microbiology, materials) was assessed using the Chi-squared test. Additionally, to assess the temporal consistency and reliability of the models, 10% of the questions were re-submitted two weeks after the initial testing. In this phase, Cohen's Kappa was used to measure agreement in accuracy between time points, while

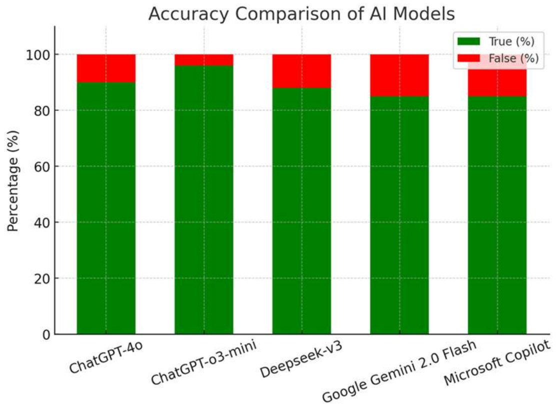

**Fig. 1.** Accuracy Comparison of Large Language Models in Dentistry

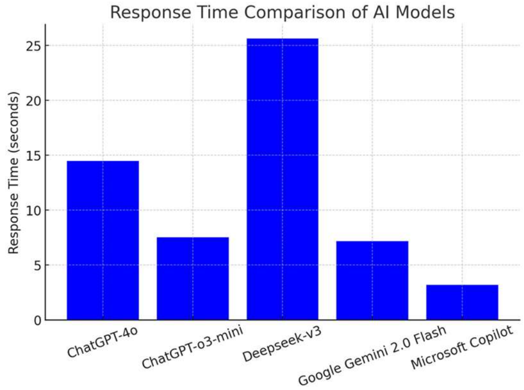

**Fig. 2.** Response Time Distribution Among AI Models

Concordance Correlation Coefficients (CCC) were calculated to assess consistency in response time and word count.

## 3. Results

The performance comparison of LLMs in terms of accuracy, response time, and word count revealed statistically significant differences across the evaluated systems (Table 2). Accuracy varied among the models, with ChatGPT-o3-mini achieving the highest proportion of correct responses (96%), which was significantly greater than Deepseek-v3 (88%), Google Gemini 2.0 Flash (85%), and Microsoft Copilot (85%) ($p = 0.011$). ChatGPT-4o demonstrated an intermediate accuracy level (90%), with no statistically significant difference from ChatGPT-o3-mini (Fig. 1).

Regarding response time, Microsoft Copilot responded the fastest, with a median time of 3.19 seconds, significantly lower than all other models ($p < 0.001$). Google Gemini 2.0 Flash and ChatGPT-o3-mini followed with similarly low response times (7.19 s and 7.54 s, respectively), whereas ChatGPT-4o (14.5 s) and particularly Deepseek-v3 (25.64 s) were significantly slower (Fig. 2).

In terms of word count, Google Gemini 2.0 Flash produced the most verbose responses (median: 218 words), significantly higher than all other models ($p < 0.001$). In contrast, Microsoft Copilot generated the shortest responses (median: 34 words), while ChatGPT-4o (144.5 words) and Deepseek-v3 (135.5 words) had comparable outputs. ChatGPT-o3-mini generated a significantly lower word count (92.5 words) than ChatGPT-4o and Deepseek-v3 but remained above Microsoft Copilot (Fig. 3).

Table 3 presents the accuracy rates of different LLMs across various dental knowledge categories. Overall, ChatGPT-o3-mini demonstrated the highest accuracy across most categories, achieving 100% accuracy in microbiology, restorative materials, therapeutic treatments, and aesthetic technologies. ChatGPT-4o and Deepseek-v3 also performed well, particularly in anatomical structures (88% and 92%, respectively) and restorative materials (94% and 91%). In contrast, Google Gemini 2.0 Flash and Microsoft Copilot showed lower accuracy in anatomical structures (79% and 75%) and restorative materials (94% and 85%), suggesting a relative weakness in these areas. Despite these differences, the p-values indicate no statistically significant differences among the models, implying comparable performance across most domains.

To evaluate temporal reliability, 10% of the questions were re-submitted to each model two weeks after initial testing. Accuracy agreement, assessed using Cohen's Kappa, was perfect (κ = 1.000) for ChatGPT-4o, ChatGPT-o3-mini, and Deepseek, indicating almost perfect reliability (κ ≥ 0.81).[15] Copilot showed substantial agreement (κ = 0.737), while Gemini demonstrated only moderate consistency (κ = 0.545). Response time and word count stability were evaluated using Concordance Correlation Coefficients (CCC). ChatGPT-4o (0.232, 0.769), o3-mini (–0.018, 0.643), and Deepseek (0.043, 0.567) exhibited moderate-to-strong concordance (CCC ≥ 0.40), while Gemini and Copilot showed poor agreement (CCC < 0.25).[16] (Table 4, Fig. 4).

## 4. Discussion

This study is among the first to systematically evaluate multiple LLMs using national specialty exam questions in restorative dentistry. The findings reveal that ChatGPT-o3-mini achieved the highest accuracy, while Copilot demonstrated the fastest response times and Gemini produced the most verbose outputs. These findings suggest that LLMs should not be evaluated through a singular definition of the "best system," but rather through a context-dependent approach that considers the suitability of different models for different tasks. Indeed, it was observed that models with higher content accuracy sometimes demonstrated slower response times, whereas models that generated faster responses tended to lag in content depth and coherence.[14,17] This underscores the need for purpose-driven model selection in application scenarios such as clinical decision support systems, patient education, preclinical learning, and the development of digital instructional materials.

The findings of this study align with the analysis conducted by Sallam, et al. [18], who reported that models like ChatGPT, Gemini, and Copilot exhibited inconsistent performance across different dental subfields. While these models showed high accuracy in certain areas, they sometimes produced content-deficient

**Table 2.** Comparative Performance of AI Models in Accuracy, Response Time, and Word Count

| | ChatGPT-4o | ChatGPT-o3-mini | Deepseek- v3 | Google Gemini 2.0 Flash | Microsoft Copilot | p-value |
|---|---|---|---|---|---|---|
| Accuracy | | | | | | |
| True | 90 (90%)$^{AB}$ | 96 (96%) $^{A}$ | 88 (88%)$^{B}$ | 85 (85%)$^{B}$ | 85 (85%)$^{B}$ | **0.011**$^{1}$ |
| False | 10 (10%)$^{AB}$ | 4 (4.0%)$^{A}$ | 12 (12%)$^{B}$ | 15 (15%)$^{B}$ | 15 (15%)$^{B}$ | |
| Response Time (second) | 14.5 (4.58-34.49)$^{C}$ | 7.535 (3.52-33.55)$^{A}$ | 25.64 (11.84-53.67)$^{D}$ | 7.19 (4.91-12.19)$^{A}$ | 3.185 (1.95-10.07)$^{B}$ | **<0.001**$^{2}$ |
| Word counts | 144.5 (28-315)$^{B}$ | 92.5 (29-315)$^{C}$ | 135.5 (34-286)$^{B}$ | 218 (87-341)$^{A}$ | 34 (11-103)$^{D}$ | **<0.001**$^{2}$ |

N (%), Median (Min-Max), $^{1}$ Cochran's Q test, McNemar Test for pairwise comparisons, $^{2}$ Kruskal Wallis test, Dwass-Steel-Critchlow-Fligner pairwise comparisons. Different uppercase superscript letters indicate significant difference (p<0.05).

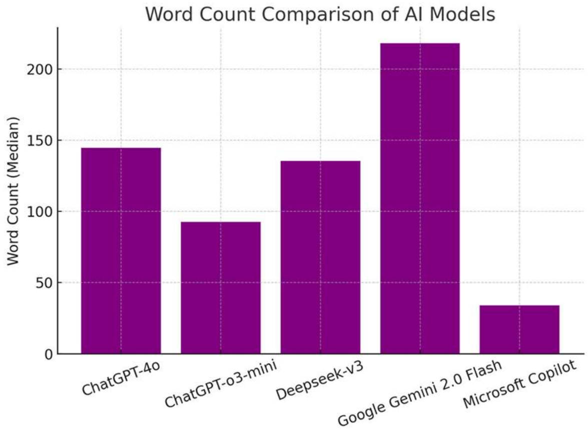

**Fig. 3.** Word Count Analysis of Model-Generated Answers

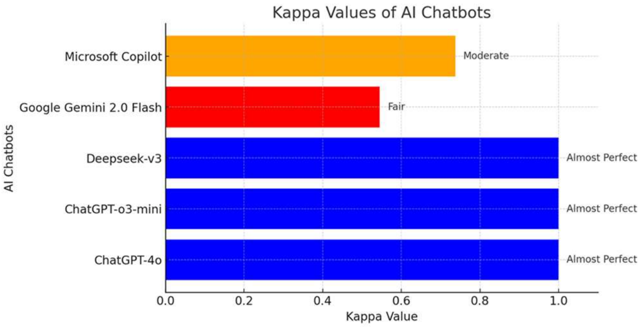

**Fig. 4.** Temporal Reliability of AI Model Responses Based on Cohen's Kappa

responses, particularly in microbiology and materials science domains.[18] Similarly, a comparative evaluation by Reyhan, et al. [19] emphasized intra-model response fluctuations and reproducibility issues, highlighting the need for careful consideration of validity and security, especially in exam-based assessments. Llorente de Pedro, et al. [20] further support these concerns, demonstrating notable inconsistencies in ChatGPT's responses to clinically structured endodontic questions, particularly under repeated prompt scenarios, thereby stressing the importance of test-retest reliability in educational and certification contexts.In conclusion, this study provides valuable empirical evidence toward understanding the suitability of advanced LLMs in clinical and pedagogical contexts within dentistry. Furthermore, identifying the strengths and limitations of different models lays a foundational framework for the development of future specialized dental LLMs.

When evaluated within this context, the low accuracy rates of LLMs in responding to basic science questions indicate that these systems still possess limited representational capacity in knowledge domains based on visual and biological data.[20] In particular, the decline in accuracy observed in disciplines such as anatomical structures, oral histology, and microbiology may stem from the underrepresentation of such content in the training datasets of LLMs. This outcome is consistent with the findings of Nguyen, et al. [14], who also highlighted the limited capacity of LLMs to generate accurate responses in dental scenarios requiring visual-spatial reasoning.

Moreover, the variability in model performance across different knowledge subcategories underscores the need to move beyond aggregated accuracy scores and adopt a topic-based performance mapping approach. For instance, the relatively high accuracy demonstrated by ChatGPT-o3-mini should not be viewed merely as superiority, but rather as a reflection of its broader and clinically optimized training corpus. Batool, et al. [21] also reported that such models tend to outperform others in medical content generation due to their enhanced compatibility with health-related terminology. Conversely, the lower accuracy scores observed in models like Google Gemini 2.0 Flash and Microsoft Copilot may be attributable not only to differences in training data but also to the primary design purposes of these systems. As these models are primarily developed for general-purpose productivity and assistant functions, their ability to deliver in-depth content in specialized technical domains (e.g., dental pharmacology, biomaterials classification) appears to be inherently limited. Overall, the findings reinforce the principle that "no single model fits all tasks." As the content, depth, and format of dental knowledge domains vary, it becomes clear that LLMs must be evaluated and selected according to a task-specific approach. This insight is not only a technical consideration but also suggests that the integration of LLM into dental education must be structured in alignment with pedagogical strategies.[19]

The differences observed in response times among LLMs are largely attributed to technical factors such as the architectural design of the systems, the volume of data used during training, and the underlying server infrastructures.[14,20,22] According to the study data, Microsoft Copilot achieved the shortest median response time (3.19 seconds), positioning it as the fastest model in this regard. Its tendency to produce relatively shorter and less context-rich responses may reduce processing load and thereby enhance speed. Similarly, Google Gemini 2.0 Flash and ChatGPT-o3-mini also demonstrated short response times, making them particularly advantageous for time-sensitive applications such as bedside decision support and urgent clinical guidance.[23]

In contrast, Deepseek-v3 exhibited the longest response time in the study (median: 25.64 seconds), making it the slowest model overall. This may be attributed to the model's architecture, which

**Table 3.** Accuracy Rates of AI Models Across Different Operative Dentistry Knowledge Categories

| Characteristics | ChatGPT-4o | ChatGPT-o3-mini | Deepseek- v3 | Google Gemini 2.0 Flash | Microsoft Copilot |
|---|---|---|---|---|---|
| Anatomical Structures and Oral Environment N = 24 (24%) | 21 (88%) | 23 (96%) | 22 (92%) | 19 (79%) | 18 (75%) |
| Dental Caries and Other Lesions N = 19 (19%) | 16 (84%) | 16 (84%) | 17 (89%) | 16 (84%) | 16 (84%) |
| Microbiology and Oral Biofilm N = 4 (4.0%) | 3 (75%) | 4 (100%) | 3 (75%) | 3 (75%) | 3 (75%) |
| Restorative Materials and Application Techniques N = 33 (33%) | 31 (94%) | 33 (100%) | 30 (91%) | 31 (94%) | 28 (85%) |
| Therapeutic and Preventive Treatments N = 9 (9.0%) | 9 (100%) | 9 (100%) | 7 (78%) | 7 (78%) | 9 (100%) |
| Aesthetic and Advanced Technologies N = 11 (11%) | 10 (91%) | 11 (100%) | 9 (82%) | 9 (82%) | 11 (100%) |
| p-value | 0.550[1] | 0.160[1] | 0.590[1] | 0.420[1] | 0.320[1] |

[1] Chi-squared test

**Table 4.** Test–Retest Reliability of Language Models Based on Repeated Submissions

| LLMs | Accuracy | Response Time (second) | Word counts |
|---|---|---|---|
| chat gpt 4o | 1.000 | 0.232 | 0.769 |
| chat gpt o3-mini | 1.000 | -0.018 | 0.643 |
| deepseek | 1.000 | 0.043 | 0.567 |
| gemini | 0.545 | 0.108 | -0.001 |
| co pilot | 0.737 | -0.003 | 0.217 |

Ten previously asked questions were re-submitted after two weeks. Accuracy was measured using Kappa statistics, while response time and word count stability were analyzed via concordance correlation coefficients.

appears to prioritize the generation of more detailed and comprehensive responses, thereby increasing processing time. However, the relationship between response time and content quality is not linear. Notably, Microsoft Copilot, the fastest model, demonstrated the lowest accuracy. Conversely, ChatGPT-o3-mini stood out with both a high accuracy rate and a balanced response time. These findings indicate that speed alone is not a reliable indicator of quality, and that content accuracy should be prioritized, especially in educational and clinical contexts.[14,20,24] In conclusion, it is evident that selecting LLMs requires establishing an optimal balance between speed and content quality.

When comparing the performance of artificial intelligence models in terms of response length, significant differences were observed among the systems. Google Gemini 2.0 Flash stood out by generating the longest and most comprehensive responses, with an average of 218 words, whereas Microsoft Copilot produced notably brief outputs, averaging only 34 words, thus offering minimal content generation.[19,25] Longer responses may provide advantages in contexts that require detailed clinical explanations, the conveyance of theoretical knowledge, or the presentation of alternative approaches in educational settings.[26] However, increased text length does not necessarily equate to higher quality.[27] Indeed, some studies have reported that excessively long responses may lead to redundant information or ambiguity, thereby complicating decision-making processes.[28] In this context, models such as ChatGPT-4o and Deepseek-v3, which produced moderately lengthy outputs, demonstrated a balanced performance by delivering content that was both informative and concise.[29,30] Although Google Gemini 2.0 Flash generated lengthy responses in terms of word count, its relatively lower accuracy rates suggest that verbosity does not always translate into meaningful information.[31] While comprehensive explanations may be preferred in educational environments, shorter, more direct, and high-accuracy content tends to be prioritized in clinical applications where time is limited.[32] Therefore, response length should not be regarded as an isolated indicator of quality; rather, it must be considered alongside context, accuracy, and the specific needs of the end user.[33]

When evaluating the performance of LLMs across dentistry-specific subdomains, significant differences in accuracy among categories become apparent.[19] In particular, models demonstrate higher overall accuracy in more clinically oriented and application-based domains such as restorative materials, therapeutic procedures, and esthetic technologies.[31] This can be attributed to the greater representation of such topics in training datasets and the clearer, more well-defined nature of the questions in these domains.[31] Similarly, the study by Lafourcade, et al. [34] in 2025, demonstrated that ChatGPT models exhibited higher accuracy and consistency in clinically oriented domains such as restorative dentistry and endodontics . In contrast, markedly lower accuracy rates were observed for models such as Google Gemini 2.0 Flash and Microsoft Copilot in foundational science categories like anatomical structures and microbiology.[28] These domains often involve complex terminology, require visual content, and demand interdisciplinary knowledge—factors that can challenge purely text-based models.[29] On the other hand, the strong performance of ChatGPT-o3-mini and ChatGPT-4o in these areas suggests that their training on more comprehensive and medically focused datasets provides a significant advantage.[32] These differences among LLMs clearly underscore the necessity of careful model selection based on the intended context of use.[33] As emphasized by Ong, et al. [32], selecting the most appropriate model for specific scenarios, such as education or clinical decision support, is crucial for achieving optimal outcomes.

One of the major strengths of this study lies in its systematic comparison of five contemporary artificial intelligence models across various dentistry-specific knowledge categories, using multidimensional criteria, namely accuracy, response time, and content comprehensiveness.[19] In the existing literature, most studies tend to focus on a single model or utilize a limited number of evaluation parameters; thus, comprehensive and multivariate analyses such as the present one remain relatively scarce.[31] Moreover, the assessment of each model not only in terms of overall performance but also across individual subcategories of knowledge provides a significant advantage in evaluating their suitability for both clinical and educational contexts.[28]

Nevertheless, this study has certain limitations. First, all questions were presented exclusively in Turkish, which may have adversely affected the performance of models with limited Turkish language support.[35] This highlights the variability in language-based response quality among LLMs.[32] In addition, the use of a limited number of multiple-choice questions per category may have restricted the depth of analysis in certain knowledge domains.[14,29,36] Moreover, although some of the evaluated models support multimodal inputs, the present assessment was restricted to text-only multiple-choice questions and did not incorporate image-based dental queries, such as radiographs, intraoral photographs, or clinical images.[37-39] As visual data play a central role in real-world dental diagnosis and treatment planning, this represents an important limitation and a recognized gap of text-only benchmarking approaches. Consequently, the findings should not be generalized to multimodal clinical scenarios involving image interpretation.

Given the continuously evolving nature of LLMs, the results obtained are specific to the particular model versions used at the time of data collection.[33] Furthermore, as the study is based on theoretical knowledge, the generalizability of its findings to real-world clinical applications may be limited.[37-39] Lastly, the ChatGPT-o3-mini model, which demonstrated the highest accuracy in our evaluation, is no longer available as a public-facing model. Consequently, reproducibility of these results using current versions (e.g., o4-mini or standard o3) may vary. Another limitation is that explicit user-selectable inference run modes (such as deep reasoning versus fast or low-latency modes) were not available via the public web interfaces of the evaluated models during the study period. As a result, all models were assessed under their default platform-defined execution settings, and potential performance differences attributable to alternative run modes could not be examined.[36,40] Despite these limitations, the present study objectively delineates the strengths and weaknesses of current LLMs in text-based dental knowledge assessment, thereby offering a meaningful contribution to the literature, particularly in the context of dental education and clinical decision-support systems.

This study highlights the performance differences of LLMs across dentistry-specific knowledge domains and provides valuable insights into their potential for clinical and educational use. In particular, the superior balance of accuracy and content demonstrated by ChatGPT-based models suggests their suitability for educational applications and clinical decision support. Future studies incorporating broader datasets, multilingual input, and case-based scenarios will be essential to further validate and deepen these findings, ultimately supporting the effective integration of LLM into dental practice.

## 5. Conclusion

This study compared the performance of five distinct LLMs across dentistry-specific knowledge domains, revealing that each model exhibits unique strengths. While ChatGPT-o3-mini stood out in terms of accuracy, Microsoft Copilot excelled in response speed, and Google Gemini 2.0 Flash delivered the most comprehensive content. The findings clearly indicate that the selection of an LLMs should be based on the intended purpose of use. In an era of rapidly advancing AI technologies, such multidimensional and comparative analyses serve as valuable guidance for both clinical and educational applications.

## References

1. Arjumand B. The application of artificial intelligence in restorative dentistry: A narrative review of current research. *Saudi Dent J*. 2024.
2. de Magalhães AA, Santos AT. Advancements in diagnostic methods and imaging technologies in dentistry: A literature review of emerging approaches. *J Clin Med*. 2025;14(4):1277.
3. Batra AM, Reche A. A new era of dental care: Harnessing artificial intelligence for better diagnosis and treatment. *Cureus*. 2023;15(11).
4. Abdulakhatov J, Oxunov B. A review of advancements of artificial intelligence in dentistry. *Web Med J Med Pract Nurs*. 2025;3(3):118–142.
5. Dhopte A, Bagde H. Smart smile: Revolutionizing dentistry with artificial intelligence. *Cureus*. 2023;15(6).
6. Ghods K, Azizi A, Jafari A, Ghods K. Application of artificial intelligence in clinical dentistry: A comprehensive review of literature. *J Dent*. 2023;24(4):356.
7. Carrillo-Perez F, Pecho OE, Morales JC, Paravina RD, Della Bona A, Ghinea R, et al. Applications of artificial intelligence in dentistry: A comprehensive review. *J Esthet Restor Dent*. 2022;34(1):259–280.
8. Shan T, Tay FR, Gu L. Application of artificial intelligence in dentistry. *J Dent Res*. 2021;100(3):232–244.
9. Revilla-León M, Gómez-Polo M, Vyas S, Barmak AB, Özcan M, Att W, et al. Artificial intelligence applications in restorative dentistry: A systematic review. *J Prosthet Dent*. 2022;128(5):867–875.
10. Shafi I, Fatima A, Afzal H, Díez I del T, Lipari V, Breñosa J, et al. A comprehensive review of recent advances in artificial intelligence for dentistry e-health. *Diagnostics*. 2023;13(13):2196.
11. Semerci ZM, Yardımcı S. Empowering modern dentistry: The impact of artificial intelligence on patient care and clinical decision making. *Diagnostics*. 2024;14(12):1260.
12. Dutta A, Pasricha N, Singh RK, Revanna R, Ramanna PK. Harnessing artificial intelligence in dentistry: Enhancing patient care and diagnostic precision. *J Dent Sci*. 2024.
13. Moeini A, Torabi S. The role of artificial intelligence in dental diagnosis and treatment planning. *J Oral Dent Health Nexus*. 2025;2(1):14–26.
14. Nguyen HC, Dang HP, Nguyen TL, Hoang V, Nguyen VA. Accuracy of latest large language models in answering multiple-choice questions in dentistry: A comparative study. *PLoS One*. 2025;20(1):e0317423.
15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
16. Akoglu H. User's guide to correlation coefficients. *Turk J Emerg Med*. 2018;18(3):91–93.
17. 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(1):573.
18. Sallam M, Al-Adwan AS, Mijwil MM, Abdelaziz DH, Al-Qaisi A, Ibrahim OM, et al. Technology readiness, social influence, and anxiety as predictors of university educators' perceptions of generative AI usefulness and effectiveness. *Front Artif Intell*. 2025;8:1571527.
19. Reyhan AH, Mutaf Ç, Uzun İ, Yüksekyayla F. A performance evaluation of large language models in keratoconus. *J Clin Med*. 2024;13(21):6512.
20. Llorente de Pedro M, Suárez A, Algar J, Díaz-Flores García V, Andreu-Vázquez C, Freire Y. Assessing ChatGPT's reliability in endodontics. *Appl Sci*. 2025;15(10):5231.
21. Batool H, Mukhtar A, Khawaja SG, et al. Knowledge distillation and transformer-based framework for automatic spine CT report generation. IEEE Access. 2025; ahead of print.
22. Xu J, Wang Y. Enhancing healthcare recommendation systems with multimodal LLMs-based MOE architecture. Paper presented at: 5th International Conference on Signal Processing and Machine Learning; 2025.
23. Naqvi WM, Ganjoo R, Rowe M, et al. Critical thinking in the age of generative AI. *Front Artif Intell*. 2024;8:1571527.
24. Hatia A, Doldo T, Parrini S, et al. Accuracy and completeness of ChatGPT-generated information on interceptive orthodontics. *J Clin Med*. 2024;13(3):735.
25. Sallam M, Alasfoor IM, Khalid SW, et al. Chinese generative AI models rival ChatGPT-4 in ophthalmology queries. *Narra J*. 2025;5(1):e2371.
26. Maes SH. The circle of life for LLMs. 2025; ahead of print.
27. Neha F, Bhati D. A survey of DeepSeek models. Authorea Preprints. 2025.
28. Gao T, Jin J, Ke ZT, Moryoussef G. A comparison of DeepSeek and other LLMs. arXiv preprint. 2025;arXiv:250203688.
29. Joshi S. A comprehensive review of DeepSeek. 2025; ahead of print.
30. Jelodar H, Meymani M, Razavi-Far R. Large language models for source code analysis. arXiv preprint. 2025;arXiv:250317502.
31. Rahman A, Mahir SH, Tashrif MTA, et al. Comparative analysis of DeepSeek, ChatGPT, and Gemini. arXiv preprint. 2025;arXiv:250304783.
32. Ong QC, Ang CS, Chee DZY, et al. Advancing health coaching with LLMs. *Artif Intell Med*. 2024;157:103004.
33. Agha AS. Evaluating AI efficiency in backend software development. 2025; ahead of print.
34. Lafourcade C, Kérourédan O, Ballester B, Richert R. Accuracy and contextual understanding of LLMs in dentistry. *J Dent*. 2025;157:105764.
35. Moreno-Molina M, Suresh A, Colman RE, Rodwell TC. Facilitating user interaction with the tuberculosis mutation catalogue using AI tools. bioRxiv. 2025.
36. Nguyen VA, Ha TBN, Tran MN, et al. Speed–accuracy trade-off of LLMs on oral surgery MCQs. *Sci Rep*. 2025;15(1):40657.
37. Carboni L. Code generation on a diet. 2024; ahead of print.
38. Nguyen VA, Vuong TQT, Nguyen VH. Benchmarking LLM vision in oral anatomy. *PLoS One*. 2025;20(10):e0335775.
39. Nguyen VA, Nguyen VH, Vuong TQT, Truong QT, Nguyen TT. Comparative reasoning in oral pathology diagnosis. *PLoS One*. 2025;20(12):e0340220.
40. Nguyen VA, Vuong TQT, Nguyen V. Comparative performance of deep-reasoning and lightweight LLMs on oral implantology MCQs. *Int J Prosthodont*. 2025;38:1–20.

**AI Declaration**

Chatgpt 4o was used for language editing to improve grammar and clarity. All data collection, screening, analysis, visualization, and interpretation were performed manually and independently verified by the author.

**Conflict of Interest**

All authors declare no conflicts of interest.

**CRediT Author Statement**

M. B. : Methodology, Investigation, Data curation
F.P.H. : Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing, Supervision, Project administration

**Data Availability Statement**

Data available from the corresponding author on reasonable request.

**Ethics Approval**

Not applicable.

**Funding**

No external funding.

**How to cite this article:**

Baytar M, Pertek Hatipoğlu F. Performance Comparison of Large Language Models in Dentistry: A Focus on Restorative Dentistry Using Turkish Specialty Exam Questions. *J Endod Restor Dent*. 2026; 4(1):7-14.
doi: 10.5281/zenodo.18585386