DATABASE ANALYSIS

e-ISSN 1643-3750
© Med Sci Monit, 2026; 32: e954052
DOI: 10.12659/MSM.954052

Received: 2026.05.08
Accepted: 2026.06.23
Available online: 2026.07.01
Published: 2026.XX.XX

# Accuracy, Self-Reported Confidence, and Overconfidence of Large Language Models in Endodontics: An Evaluation Using National Specialty Examination Questions

Authors' Contribution:
Study Design A
Data Collection B
Statistical Analysis C
Data Interpretation D
Manuscript Preparation E
Literature Search F
Funds Collection G

ABCDEF **Mehmet Adıgüzel**

Department of Endodontics, Faculty of Dentistry, Hatay Mustafa Kemal University, Hatay, Türkiye



**Corresponding Author:** Mehmet Adıgüzel, Department of Endodontics, Faculty of Dentistry, Hatay Mustafa Kemal University, Hatay, Türkiye, Phone: +903262456060, e-mail: dt.mehmetadiguzel@gmail.com

**Financial support:** None declared
**Conflict of interest:** None declared

**Background:** This study aimed to evaluate the accuracy, confidence, and overconfidence behavior of large language models (LLMs) in endodontics using questions derived from a national specialty entrance examination.

**Material/Methods:** A total of 123 text-based endodontic questions from the Turkish Dental Specialty Examination (2017-2026) were included after excluding annulled and image-based questions. Three LLMs (ChatGPT, Claude, and Gemini) were assessed. Each model answered all questions using a standardized prompt and provided a confidence score (0%-100%). Accuracy was recorded as correct/incorrect. Overconfidence was defined as incorrect responses with ≥90% confidence. Statistical analyses were performed using Cochran's Q test, Friedman test, and post hoc pairwise comparisons with Bonferroni correction.

**Results:** Accuracy rates were 89.4% for ChatGPT, 76.4% for Claude, and 90.2% for Gemini, with significant differences among models ($\chi^2(2) = 21.00$, $P < 0.001$). Confidence scores differed significantly ($\chi^2(2) = 213.40$, $P < 0.001$), with Gemini demonstrating highest confidence (99.51 ± 1.49), followed by ChatGPT (90.88 ± 9.19) and Claude (80.22 ± 11.10). Overconfidence rates were 9.8% for Gemini and 5.7% for ChatGPT, while no overconfident responses were observed for Claude ($\chi^2(2) = 14.53$, $P = 0.001$). Despite similar accuracy between ChatGPT and Gemini, confidence patterns differed markedly, demonstrating that comparable accuracy does not necessarily reflect comparable reliability.

**Conclusions:** LLMs demonstrated high accuracy in answering text-based endodontic examination questions; however, significant differences were observed in confidence behavior and overconfidence patterns. The presence of high-confidence incorrect responses suggests that accuracy alone may be insufficient to fully evaluate model reliability. These findings highlight the importance of considering confidence-related behavior alongside accuracy when assessing LLM performance in examination-style endodontic tasks.

**Keywords:** Large Language Models • Endodontics • Artificial Intelligence

**Full-text PDF:** https://www.medscimonit.com/abstract/index/idArt/954052





Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

## Introduction

In recent years, large language models (LLMs) have emerged as powerful tools in healthcare, attracting increasing attention as potential aids in medical and dental education. Trained on extensive corpora of biomedical and general knowledge, these models have demonstrated considerable ability to process complex medical terminology, encode clinical knowledge, and generate coherent responses to clinical questions [1,2]. Several studies have also reported that LLMs can achieve high accuracy rates on standardized examinations, including the United States Medical Licensing Examination (USMLE), dental licensing examinations, and national specialty access examinations conducted in non-English languages, suggesting performance levels comparable to, or in some cases exceeding, those of human examinees in knowledge-based assessments [3-6].

Despite these promising findings, concerns remain regarding the extent to which such performance reflects true clinical competence. Emerging evidence indicates that although LLMs may perform well on factual recall and standardized examination tasks, their accuracy can decline when faced with problems requiring higher-order reasoning, contextual decision-making, and integration of multiple variables. This limitation has been attributed to inflexible pattern matching, hallucination, and overconfidence rather than genuine clinical inference. Accordingly, strong benchmark performance does not necessarily translate into real-world clinical competence, a discrepancy that has been described as the "knowledge-practice gap" [7,8].

Within dentistry, and particularly in endodontics, the evaluation of LLM performance remains limited. Recent studies have demonstrated moderate-to-high performance of LLMs on dental and endodontic question-answering tasks; however, substantial variability in accuracy, consistency, and clinical reliability has also been reported. Endodontics is a uniquely challenging domain due to its reliance on applied knowledge, spatial reasoning, radiographic interpretation, and the integration of multiple clinical variables during diagnostic and treatment-related decision-making. Nevertheless, most existing studies remain restricted to text-based multiple-choice questions or narrowly focused diagnostic scenarios, which may inadequately reflect the complexity of clinically contextualized decision-making and multimodal diagnostic processes [9-12].

Another notable gap in the literature is the scarcity of studies conducted using non-English, specialty-level examination datasets. The Turkish Dental Specialty Examination (Diş Hekimliği Uzmanlık Sınavı [DUS]), a national specialty entrance examination for dentistry in Turkey, provides a valuable, high-stakes, and underexplored framework for evaluating LLM performance in a real-world, high-stakes, and linguistically distinct context. However, there is a lack of comprehensive studies focusing specifically on endodontic questions within the DUS.

A particularly important concern is the tendency of LLMs to provide high-confidence responses even when incorrect. Such outputs may create a false perception of reliability, particularly in clinical settings where users may equate confidence with correctness. Recent studies have demonstrated that LLMs may exhibit limited differentiation in confidence between correct and incorrect responses, thereby reducing the reliability of uncertainty signaling and increasing the risk of misleading certainty in high-stakes decision-making [13-15].

Therefore, the present study aimed to provide an endodontics-specific evaluation of LLM performance using questions derived from the DUS, a non-English national specialty examination context that remains underrepresented in the literature, with particular emphasis on the relationship between answer accuracy and self-reported confidence. Unlike most previous studies that focused primarily on answer accuracy, the present study simultaneously evaluated answer accuracy, self-reported confidence, overconfidence behavior, and inter-model agreement, providing a more comprehensive assessment of LLM performance in endodontic examination settings. The primary inferential outcomes were answer accuracy, self-reported confidence, and overconfidence behavior. In addition, inter-model agreement was evaluated as an exploratory analysis to assess the consistency of response patterns across models.

Specifically, we investigated whether LLMs differed in terms of answer accuracy, whether self-reported confidence levels varied among models, and whether high-confidence incorrect responses (overconfidence) were observed. It was hypothesized that the evaluated LLMs would differ in answer accuracy and confidence-related behavior, and that high-confidence incorrect responses would be observed in at least some models.

Recent literature has emphasized the need for multidimensional evaluation frameworks that extend beyond accuracy alone [16]. The present findings contribute to this emerging perspective by providing a multidimensional assessment of LLM performance in endodontics.

## Material and Methods

### Study Design

This study was designed as a cross-sectional, comparative in silico analysis to evaluate the performance of LLMs on endodontic questions derived from the DUS, a national specialty examination. Ethical approval was not required because the study did not involve human participants, patient data, or animal subjects.





Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

### Data Source and Question Selection

The dataset consisted of all endodontic questions obtained from the DUS administered between 2017 and 2026. During this period, a total of 13 examinations were conducted, each including 10 endodontic questions.

All questions within this timeframe were included without sampling. Questions that had been officially annulled (eg, 1 question from the 2017 examination) were excluded. In addition, questions containing radiographic images, clinical photographs, or graphical elements were excluded due to the text-based nature of the evaluated models.

After applying these exclusion criteria, a total of 123 text-based questions with clearly established and verifiable correct answers were included in the final analysis.

### Large Language Models

Three LLMs were evaluated: ChatGPT (OpenAI; GPT-5.3), Claude (Anthropic; Claude Sonnet 4.6), and Gemini (Google; Gemini 3.1 Pro). ChatGPT, Claude, and Gemini were accessed through their respective publicly available web-based interfaces provided by OpenAI, Anthropic, and Google at the time of testing (April 2026). Model parameters were not modified, and no external tools were used during testing.

### Prompting Procedure

Each question was presented to the models using the following standardized Turkish prompt:

"Aşağıdaki çoktan seçmeli soruyu cevapla. Sadece bir seçeneği işaretle (A, B, C, D veya E). Açıklama yapma. Cevabına olan güven düzeyini yüzde olarak belirt (0%-100%). Soru: [question text]". The English translation is "Answer the following multiple-choice question. Select only one option (A, B, C, D, or E). Do not provide any explanation. Indicate your level of confidence in your answer as a percentage (0%-100%). Question: [question text]".

This prompt was used to ensure that each model provided a single selected answer and a numerical self-reported confidence score without generating explanatory text. To minimize order effects, questions were presented in a randomized sequence for each model. Each question was evaluated in a separate session to prevent context carryover. Each model generated only 1 response per question. A single-response design was intentionally adopted to better reflect real-world examination and user-interaction settings, where decisions are typically based on a single generated output rather than repeated sampling.

### Question Classification

For exploratory subgroup analyses, all questions were classified as either basic knowledge questions or clinical questions according to their primary cognitive requirement. Basic knowledge questions assessed factual knowledge, theoretical concepts, and foundational endodontic principles, whereas clinical questions required diagnostic reasoning, treatment planning, or clinical decision-making. Questions with potentially overlapping characteristics were reviewed and assigned based on the predominant competency being assessed.

### Outcome Measures

The primary inferential outcomes were answer accuracy, self-reported confidence, and overconfidence behavior. Inter-model agreement was additionally evaluated as an exploratory analytic component. For accuracy, model responses were compared with the official answer key and categorized as correct (1) or incorrect (0). For the confidence score, each model's self-reported confidence was recorded as a numerical percentage ranging from 0% to 100%. Overconfidence was evaluated as a secondary outcome and defined as an incorrect response accompanied by a confidence score of 90% or higher. The 90% threshold was selected to represent highly certain responses and to identify instances of potentially misleading certainty in incorrect outputs.

In this study, confidence-related behavior was operationalized using self-reported confidence scores and the frequency of high-confidence incorrect responses, thereby allowing comparison of confidence patterns and high-confidence error behavior across models.

Inter-model agreement was evaluated to explore the consistency of response patterns across models. Agreement was assessed based on correctness patterns for each question.

### Statistical Analysis

All statistical analyses were performed using SPSS software (IBM Corp, Armonk, NY, USA). Because each question was answered by all 3 models, the data were treated as paired (dependent) observations.

Differences in accuracy among the models were analyzed using Cochran's Q test. When a significant difference was detected, pairwise comparisons were conducted using McNemar tests with Bonferroni correction.

Confidence scores were summarized using mean ± standard deviation and median (interquartile range) to provide complementary measures of central tendency and dispersion for potentially non-normally distributed paired data. Because confidence







Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

scores represented paired observations and showed non-uniform distributions across models, comparisons were performed using nonparametric related-samples tests. Confidence scores were analyzed using the Friedman test for related samples. When significant differences were observed, post hoc pairwise comparisons were performed using the Wilcoxon signed-rank test with Bonferroni adjustment. Effect sizes were reported for confidence analyses. Kendall's W was calculated for the Friedman test, and effect size r was calculated for Wilcoxon signed-rank tests using the formula $r = Z/\sqrt{N}$. Effect sizes were interpreted according to conventional thresholds.

Inter-model agreement was additionally evaluated as an exploratory analysis using pairwise Cohen's kappa statistics and overall Fleiss' kappa. Agreement analysis was performed to determine whether models with similar overall accuracy consistently identified the same questions correctly, thereby providing insight into response consistency beyond accuracy alone. The unit of analysis was individual examination questions, and agreement analyses were based on correctness patterns rather than answer-option selections across the evaluated questions. Confidence intervals were calculated and reported for pairwise Cohen's kappa statistics; however, formal significance testing was not performed for agreement analyses.

The proportion of overconfident responses was also compared among models using Cochran's Q test, followed by pairwise McNemar tests when appropriate. Descriptive statistics were reported as frequencies and percentages for categorical variables and as mean ± standard deviation and median (interquartile range) for continuous variables.

A Bonferroni-adjusted significance threshold of $P < 0.017$ was applied for pairwise comparisons. Exploratory subgroup analyses were additionally performed according to question type (basic knowledge vs clinical questions) to evaluate potential differences in accuracy, confidence, and overconfidence patterns. Accuracy and overconfidence rates were compared using chi-square or Fisher's exact tests, whereas confidence scores were compared using independent-samples *t* tests. A *P* value of < 0.05 was considered statistically significant for overall tests.

For descriptive outcome estimates, 95% confidence intervals were calculated. Wilson confidence intervals were used for proportions, whereas t-based confidence intervals were used for mean confidence scores.

## Results

A total of 123 endodontic questions were evaluated by each model. Overall accuracy was 89.4% for ChatGPT, 76.4% for Claude, and 90.2% for Gemini. Accuracy differed significantly among the models (Cochran's Q, $\chi^2(2) = 21.00$, $P < 0.001$), with both ChatGPT and Gemini performing significantly better than Claude, whereas their accuracies were comparable (**Tables 1-3**). The comparative accuracy rates are illustrated in **Figure 1**.

Confidence scores also differed significantly among the models (Friedman test, $\chi^2(2)=213.40$, $P < 0.001$, Kendall's W=0.87), indicating a large effect size. Gemini showed the highest confidence scores (99.51 ± 1.49), followed by ChatGPT (90.88 ± 9.19), whereas Claude showed the lowest confidence scores (80.22 ± 11.10). Median (interquartile range) values for confidence scores are presented in **Table 1**, and score distributions are illustrated in **Figure 2**. Post hoc pairwise comparisons using the Wilcoxon signed-rank test demonstrated significant differences in confidence scores between all model pairs. Large effect sizes were observed for ChatGPT vs Claude (r = 0.77), Claude vs Gemini (r = 0.87), and ChatGPT vs Gemini (r = 0.82). Detailed pairwise comparison results are presented in **Table 3**.

The frequency of high-confidence incorrect responses also differed significantly among models (Cochran's Q, $\chi^2(2) = 14.53$, $P = 0.001$). Gemini had the highest overconfidence rate (9.8%), followed by ChatGPT (5.7%), whereas Claude had no overconfident responses. Pairwise comparisons showed that both Gemini and ChatGPT had significantly higher overconfidence rates than Claude, with no significant difference between Gemini and ChatGPT. The distribution of overconfidence rates across models is presented in **Figure 3**.

Overall, ChatGPT and Gemini achieved comparable accuracy; however, Gemini demonstrated substantially higher and less variable confidence, indicating a divergence between confidence and correctness. Agreement analysis showed moderate agreement between ChatGPT and Gemini ($\kappa = 0.51$, 95% CI 0.26-0.76), ChatGPT and Claude ($\kappa = 0.44$, 95% CI 0.25-0.63), and Claude and Gemini ($\kappa = 0.41$, 95% CI 0.21-0.60). Overall inter-model agreement was moderate (Fleiss' $\kappa = 0.44$), suggesting that the models did not consistently identify the same questions correctly.

Of the 123 included questions, 77 (62.6%) were classified as basic knowledge questions and 46 (37.4%) as clinical questions. Exploratory subgroup analyses showed no statistically significant differences in accuracy between basic and clinical questions for ChatGPT (89.6% vs 89.1%, $P = 0.933$), Claude (75.3% vs 78.3%, $P = 0.710$), or Gemini (92.2% vs 87.0%, $P = 0.342$). Similarly, confidence scores did not differ significantly between question types for ChatGPT (91.4 ± 9.5 vs 89.9 ± 8.7, $P = 0.381$), Claude (80.7 ± 11.7 vs 79.5 ± 10.0, $P = 0.558$), or Gemini (99.4 ± 1.6 vs 99.7 ± 1.3, $P = 0.354$). Overconfidence rates were also comparable between basic and clinical questions for ChatGPT (3.9% vs 8.7%, $P = 0.266$) and Gemini (7.8% vs 13.0%, $P = 0.342$), whereas no overconfident responses were observed for Claude in either question category (**Table 4**).





Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

**Table 1.** Overall performance metrics of the evaluated large language models.

| Model | Accuracy, n (%) (95% CI) | Confidence, mean ± SD (95% CI) | Median (IQR) | Overconfidence, n (%) (95% CI) |
|---|---|---|---|---|
| ChatGPT | 110 (89.4%) (82.8-93.7) | 90.88 ± 9.19 (89.24-92.52) | 95 (90-95) | 7 (5.7%) (2.8-11.3) |
| Claude | 94 (76.4%) (68.2-83.1) | 80.22 ± 11.10 (78.24-82.20) | 82 (72-90) | 0 (0.0%) (0.0-3.0) |
| Gemini | 111 (90.2%) (83.7-94.3) | 99.51 ± 1.49 (99.24-99.78) | 100 (100-100) | 12 (9.8%) (5.7-16.3) |

Notes: Comparison of accuracy, self-reported confidence scores, and overconfidence rates among ChatGPT, Claude, and Gemini in answering 123 endodontic specialty examination questions. Confidence scores are presented as mean ± standard deviation and median (interquartile range). Values in parentheses represent 95% confidence intervals. Overconfidence was defined as incorrect responses with confidence scores ≥ 90%.

**Table 2.** Overall statistical comparisons among large language models.

| Outcome | Test | Statistic | Effect size | *P* |
|---|---|---|---|---|
| Accuracy | Cochran's Q | $\chi^2(2)=21.00$ | — | < 0.001 |
| Confidence | Friedman | $\chi^2(2)=213.40$ | Kendall's W=0.87 | < 0.001 |
| Overconfidence | Cochran's Q | $\chi^2(2)=14.53$ | — | 0.001 |

Notes: Overall comparisons of accuracy, confidence scores, and overconfidence rates among the evaluated models. Cochran's Q test was used for binary outcomes (accuracy and overconfidence), and the Friedman test was used for paired confidence score comparisons.

**Table 3.** Pairwise comparisons of accuracy, confidence, and overconfidence between the evaluated large language models.

| Comparison | Accuracy *P* value | Accuracy OR (95% CI) | Confidence Z | Confidence effect size (r) | Confidence *P* value | Overconfidence *P* value |
|---|---|---|---|---|---|---|
| ChatGPT vs Claude | < 0.001 | 9.00 (1.95-41.6) | -8.538 | 0.77 | < 0.001 | 0.016 |
| Claude vs Gemini | < 0.001 | 9.50 (2.07-43.6) | -9.662 | 0.87 | < 0.001 | < 0.001 |
| ChatGPT vs Gemini | 1.000 | 0.83 (0.25-2.77) | -9.109 | 0.82 | < 0.001 | 0.227 |

Notes: Pairwise comparisons of accuracy, confidence scores, and overconfidence rates between ChatGPT, Claude, and Gemini. Accuracy and overconfidence were compared using McNemar tests, while confidence scores were compared using Wilcoxon signed-rank tests. Bonferroni-adjusted statistical significance was set at $P < 0.017$. Odds ratios (ORs) and 95% confidence intervals were calculated from discordant pairs in McNemar comparisons. Values greater than 1 indicate a higher likelihood of correct responses for the first model listed in each comparison.

## Discussion

The present study provides a comprehensive evaluation of LLM performance in endodontics using endodontic questions derived from a national specialty entrance examination.

A principal finding of this study is that although ChatGPT and Gemini demonstrated comparable accuracy levels, substantial differences emerged in confidence behavior. Gemini exhibited consistently high confidence with minimal variability, whereas Claude demonstrated lower and more dispersed confidence scores. These findings indicate that accuracy alone does not adequately reflect the reliability of LLM outputs. In particular, the combination of high confidence and incorrect responses observed in Gemini suggests a mismatch between self-reported confidence and correctness despite otherwise strong overall performance. Importantly, the near-ceiling confidence distribution observed in Gemini should be interpreted as a concentration of self-reported confidence scores rather than evidence of formal calibration or superior uncertainty estimation.







Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

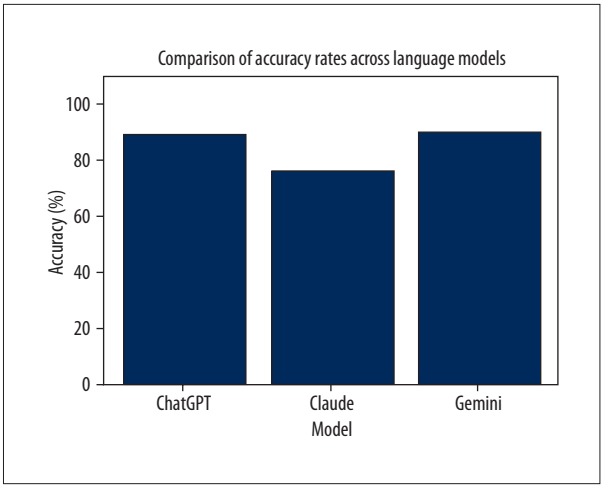


**Figure 1. Comparison of accuracy rates among large language models.** Bar chart showing the percentage of correctly answered endodontic examination questions by ChatGPT, Claude, and Gemini.

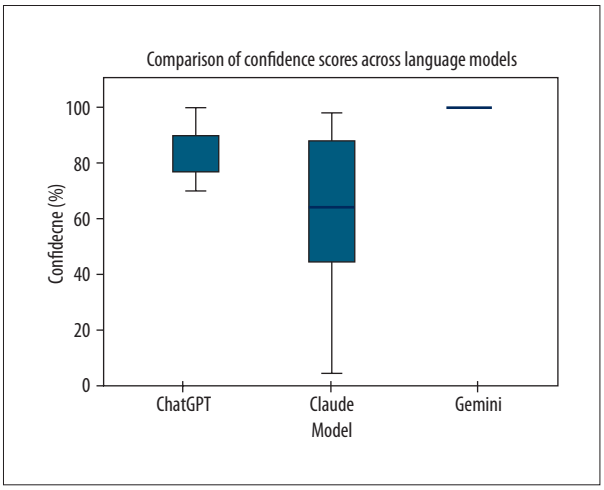


**Figure 2. Comparison of self-reported confidence scores among language models.** Boxplot showing the distribution of confidence scores (%) for ChatGPT, Claude, and Gemini across 123 endodontic questions. Boxes represent the interquartile range, horizontal lines indicate medians, and whiskers represent variability outside the upper and lower quartiles.

The presence of high-confidence incorrect responses represents a critical concern. From a clinical perspective, confidently delivered incorrect information may be more misleading than uncertain responses, potentially contributing to automation bias, cognitive offloading, and over-reliance on AI-generated outputs in clinical decision-making [14,17,18]. These findings emphasize that confidence-related behavior should be considered alongside accuracy in the evaluation of LLMs, particularly in high-stakes domains such as endodontics. Recent studies have similarly suggested that LLMs often exhibit inadequate

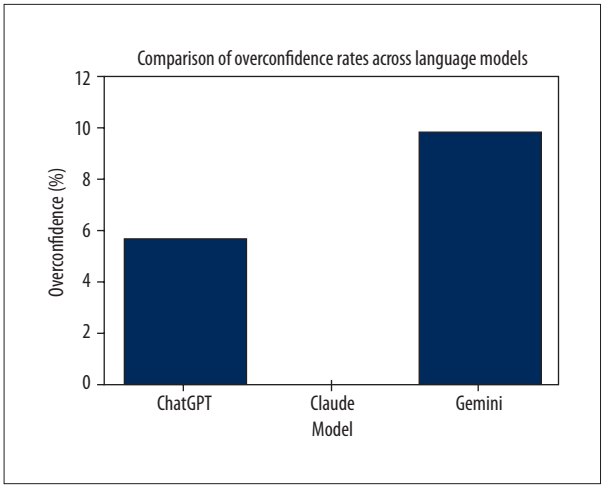


**Figure 3. Comparison of overconfidence rates among large language models.** Bar chart showing the proportion of incorrect responses accompanied by confidence scores ≥90% for each evaluated model.

differentiation in confidence between correct and incorrect responses, thereby limiting the reliability of uncertainty signaling [13]. The absence of overconfident responses in Claude, despite its lower overall accuracy, suggests that confidence calibration and answer correctness may represent distinct dimensions of model behavior.

Agreement analysis further revealed that, despite similar accuracy levels, models did not consistently converge on the same correct answers. Moderate agreement between models suggests that comparable performance may arise from different response patterns rather than shared reasoning processes. This variability indicates that LLMs may rely on distinct internal representations and decision pathways, reinforcing the need to assess model behavior beyond accuracy alone. Accordingly, high overall performance may not necessarily guarantee consistent performance patterns across evaluations [19]. Because agreement was evaluated according to correctness patterns rather than specific answer selections, the reported kappa values reflect consistency in identifying questions correctly rather than consistency in underlying reasoning processes or response content.

The inclusion of multiple LLMs may improve the external validity and generalizability of the present findings. Although ChatGPT and Gemini achieved comparable accuracy, substantial differences were observed in confidence-related behavior and response consistency. This suggests that conclusions based on a single model may not necessarily reflect the behavior of other contemporary LLMs. Therefore, comparative evaluations involving multiple models may provide a more comprehensive understanding of the strengths, limitations, and potential risks associated with the use of LLMs in dental education and assessment settings.





Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

**Table 4.** Exploratory subgroup analysis according to question type.

| Outcome | ChatGPT basic (n=77) | ChatGPT clinical (n=46) | *P* | Claude basic (n=77) | Claude clinical (n=46) | *P* | Gemini basic (n=77) | Gemini clinical (n=46) | *P* |
|---|---|---|---|---|---|---|---|---|---|
| Accuracy, n (%) | 69 (89.6) | 41 (89.1) | 0.933 | 58 (75.3) | 36 (78.3) | 0.710 | 71 (92.2) | 40 (87.0) | 0.342 |
| Confidence, mean ± SD | 91.4 ± 9.5 | 89.9 ± 8.7 | 0.381 | 80.7 ± 11.7 | 79.5 ± 10.0 | 0.558 | 99.4 ± 1.6 | 99.7 ± 1.3 | 0.354 |
| Overconfidence, n (%) | 3 (3.9) | 4 (8.7) | 0.266 | 0 (0.0) | 0 (0.0) | — | 6 (7.8) | 6 (13.0) | 0.342 |

Notes: Comparison of accuracy, self-reported confidence scores, and overconfidence rates between basic knowledge questions and clinical questions for ChatGPT, Claude, and Gemini. Basic knowledge questions assessed factual knowledge, theoretical concepts, and foundational endodontic principles, whereas clinical questions required diagnostic reasoning, treatment planning, or clinical decision-making. Confidence scores are presented as mean ± standard deviation. Overconfidence was defined as incorrect responses with confidence scores ≥ 90%. Basic knowledge questions assessed factual knowledge and theoretical concepts, whereas clinical questions required diagnostic reasoning, treatment planning, or clinical decision-making.

Previous studies evaluating LLM performance in dentistry and endodontics have primarily focused on accuracy, consistency, or diagnostic performance. For example, Suárez et al [12] evaluated the consistency and accuracy of ChatGPT-generated answers in endodontics, whereas Durmazpinar and Ekmekci [11] assessed the diagnostic performance of ChatGPT-4o in endodontic cases. Nguyen et al [10] compared the accuracy of multiple contemporary LLMs across dental disciplines. In contrast, the present study additionally evaluated self-reported confidence and overconfidence behavior, demonstrating that models with comparable accuracy may exhibit substantially different confidence-related risk profiles. These findings also contribute to the ongoing discussion of the "knowledge-practice gap." While LLMs appear capable of achieving high performance in structured, knowledge-based assessments, their reliability becomes less predictable when confidence and agreement patterns are considered. Models may therefore generate outputs that appear authoritative without reflecting true underlying understanding. Exploratory subgroup analyses further suggested that model accuracy, confidence, and overconfidence patterns were generally consistent across basic knowledge and clinical questions, indicating that the observed performance characteristics were not substantially influenced by question type.

Differences among the evaluated models further highlight the role of model architecture and training strategies in shaping performance characteristics. Although ChatGPT and Gemini achieved similar levels of accuracy, their confidence distributions and agreement patterns differed substantially, indicating that models with comparable correctness may still vary in their risk profiles. In contrast, Claude's lower confidence levels, despite reduced accuracy, may reflect a more conservative response strategy, potentially reducing the risk of misleading certainty. One possible explanation for this behavior may be reinforcement learning from human feedback, in which models are optimized to generate clear, helpful, and confident responses that align with human preferences, potentially at the expense of appropriate uncertainty expression. Recent studies have suggested that reinforcement learning from human feedback may intrinsically promote overconfidence by sharpening output probability distributions during reward optimization [20,21].

The findings of this study should be interpreted in light of several limitations. First, the use of text-based multiple-choice questions does not fully reflect the complexity of clinical endodontic practice, which frequently involves radiographic interpretation, visual information processing, and dynamic clinical decision-making. Consequently, the reported performance may overestimate model capabilities in real-world endodontic settings that require multimodal reasoning and integration of imaging data.

Second, each question was evaluated using a single model response. Although this approach was intended to reflect real-world user interactions, repeated sampling may produce different outputs because of the probabilistic nature of LLMs. Therefore, the observed accuracy, confidence, and overconfidence rates may not fully capture the variability of model behavior across repeated interactions. Although each question was treated as an independent observation, latent clustering according to topic content or question characteristics cannot be completely excluded and may have influenced model performance patterns.

Third, only 3 models were included, limiting generalizability. Additionally, the findings may be specific to endodontic knowledge domains and may not be directly generalizable to other dental or medical specialties with different cognitive and diagnostic demands. Furthermore, all questions were presented in Turkish, the original language of the DUS examination. Although this approach preserved the authentic examination context, the







Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

evaluated LLMs were primarily developed using predominantly English-language training data. Therefore, language-related differences in knowledge representation, comprehension, and response generation may have influenced model performance. Consequently, the findings may not be directly generalizable to equivalent examinations conducted in other languages. Finally, rapid model updates may affect reproducibility over time.

Despite these limitations, this study provides relevant insights into LLM performance in text-based, examination-style endodontic knowledge tasks by utilizing specialty examination data and incorporating a multidimensional evaluation framework. The findings demonstrate that while LLMs show promising performance in knowledge-based assessments, discrepancies in confidence behavior and inter-model agreement are important limitations. These results support recent calls for composite evaluation frameworks that integrate accuracy with self-reported confidence, overconfidence behavior, and response consistency. Future research should focus on multimodal evaluations incorporating imaging data, repeated-response designs, and the development of metrics that better capture applied reasoning and clinically contextualized performance.

In conclusion, although LLMs demonstrated high accuracy in text-based endodontic examination settings, differences in confidence behavior, overconfidence patterns, and agreement profiles indicate that these findings should be interpreted with caution. The present results do not establish LLM performance in multimodal, patient-specific, or real-time clinical decision-making contexts. Further studies using clinically contextualized and multimodal tasks are required before broader conclusions can be drawn regarding educational or clinical integration.

## Conclusions

Within the limitations of this study, LLMs demonstrated high accuracy in answering text-based endodontic examination questions; however, marked differences were observed in confidence behavior, overconfidence patterns, and agreement profiles. Although ChatGPT and Gemini showed comparable accuracy, Gemini exhibited substantially higher and less variable confidence, with more high-confidence incorrect responses. These findings indicate that accuracy alone is insufficient to assess LLM reliability and support the use of multidimensional evaluation frameworks that incorporate confidence-related behavior and response consistency alongside accuracy. Because the present study was limited to text-based examination questions, the findings should not be directly generalized to multimodal, patient-specific, or real-time clinical decision-making contexts. Future research should focus on multimodal assessments, repeated-response designs, and clinically contextualized tasks to better characterize the reliability and practical applicability of LLMs in endodontics.

### Acknowledgements

AI-assisted language editing tools were used during manuscript preparation. All outputs were critically reviewed and verified by the author.

### Department and Institution Where Work Was Done

Department of Endodontics, Faculty of Dentistry, Hatay Mustafa Kemal University, Hatay, Türkiye.

### Patient consent

Not applicable

### Declaration of Figures' Authenticity

All figures submitted have been created by the author who confirms that the images are original with no duplication and have not been previously published in whole or in part.

## References:

1. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80
2. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589-96
3. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLoS Digit Health. 2023;2(2):e0000198
4. Kasagga A, Sapkota A, Changaramkumarath G, et al. Performance of ChatGPT and large language models on medical licensing exams worldwide: A systematic review and network meta-analysis with meta-regression. Cureus. 2025;17(10):e94300
5. Liu M, Okuhara T, Huang W, et al. Large language models in dental licensing examinations: Systematic review and meta-analysis. Int Dent J. 2025;75(1):213-22
6. Benito P, Isla-Jover M, González-Castro P, et al. GPT-4o and OpenAI o1 performance on the 2024 Spanish competitive medical specialty access examination: Cross-sectional quantitative evaluation study. JMIR Med Educ. 2026;12(1):e75452
7. Kvist T, Hofmann B. Clinical decision making of post-treatment disease. Int Endod J. 2023;56(Suppl. 2):154-68
8. Gliga A, Imre M, Grandini S, et al. The limitations of periapical X-ray assessment in endodontic diagnosis: A systematic review. J Clin Med. 2023;12(14):4647
9. Yang Y, Jin Q, Zhu Q, et al. Beyond multiple-choice accuracy: Real-world challenges of implementing large language models in healthcare. Annu Rev Biomed Data Sci. 2025;8:305-16
10. Nguyen HC, Dang HP, Nguyen TL, et al. Accuracy of latest large language models in answering multiple choice questions in dentistry: A comparative study. PLoS One. 2025;20(1):e0317423





Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

11. Durmazpinar PM, Ekmekci E. Comparing diagnostic skills in endodontic cases: Dental students versus ChatGPT-4o. BMC Oral Health. 2025;25(1):457
12. Suárez A, Díaz-Flores García V, Algar J, et al. Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers. Int Endod J. 2024;57(1):108-13
13. Berkowitz JS, Patock JR, Nawaz A, et al. A crisis of overconfidence: Why confidence, not accuracy, is the real risk in clinical AI. BioData Min. 2026;19(1):10
14. Saadeh MI, Janhonen J, Beer E, et al. Automation complacency: Risks of abdicating medical decision making. AI Ethics. 2025;5(6):5783-93
15. Savage T, Wang J, Gallo R, et al. Large language model uncertainty proxies: Discrimination and calibration for medical diagnosis and treatment. J Am Med Inform Assoc. 2025;32(1):139-49
16. Kanithi PK, Christophe C, Pimentel MAF, et al. MEDIC: Towards a comprehensive framework for evaluating LLMs in clinical applications. arXiv. 2024;2409.07314 [Preprint posted online September 11, 2024]
17. Khera R, Simon MA, Ross JS. Automation bias and assistive AI: Risk of harm from AI-driven clinical decision support. JAMA. 2023;330(23):2255-57
18. Goddard K, Roudsari A, Wyatt JC. Automation bias: A systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012;19(1):121-27
19. Omar M, Agbareia R, Glicksberg BS, et al. Benchmarking the confidence of large language models in answering clinical questions: Cross-sectional evaluation study. JMIR Med Inform. 2025;13:e66917
20. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst. 2022;35:27730-44
21. Naderi N, Safavi-Naini SAA, Savage T, et al. Across generations, sizes, and types, large language models poorly report self-confidence in gastroenterology clinical reasoning tasks. NPJ Gut and Liver. 2026;3:6







Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]